Coins 1964

1

A MODEL FOR EXPERIENTIAL STORAGE IN NEURAL NETWORKS*

F. Rosenblatt

Cornell University

INTRODUCTION

This is a preliminary exposition of a model for long-term sequential memory in the brain. The model was first formulated about a year ago, and has since been subjected to a series of numerical studies which have revealed a capacity considerably beyond our original expectations. A more comprehensive treatment of the theory, with emphasis on its biological and biochemical aspects, is currently in preparation.[46] The objective of this paper is to summarize the main quantitative results (particularly those which may be of engineering interest) and to suggest some of the possible applications of this model to the problems of memory, recall, and the learning of heuristic programs and algorithms. This seems, in fact, to be the first instance of a model of the perceptron variety which is sophisticated enough to learn computerlike programs employing stored data and stored instructions.

The reader will find that an introductory knowledge of perceptron theory is essential for an adequate understanding of the following presentation. A general orientation can be obtained from Block,[4] or from the first eight chapters of Rosenblatt.[44] To bring the perceptual aspects of the theory up to date, the summary of recent work in Ref. 45 might prove helpful. The present paper assumes some knowledge of the pattern recognition and discrimination capabilities of perceptrons, and concentrates entirely on the question of how a record of such sensory experience can be stored and

* The work reported in this paper has been supported jointly by the Information Systems Branch of the Office of Naval Research, and by the National Science Foundation.

recalled, for periods comparable in duration to a human lifetime, by a biologically plausible network of neurons.

An adequate theory of memory must satisfy at least two criteria:

1. It must employ a recording mechanism (or "trace mechanism") which is physically, mathematically, and biologically plausible.
2. It must provide a demonstration that this trace mechanism, when incorporated into a biologically plausible neural network, can in fact account for the basic psychological phenomena of memory and recall.

In order to motivate the particular trace mechanism which is proposed, we shall undertake a demonstration of the second point first, introducing a detailed biochemical mechanism only after its required characteristics have been clarified.

The problem of memory as a physiological phenomenon has, of course, been dealt with in a vast volume of literature. Recent reviews by Gerard[21,22] Morell,[40,41] John,[31] and Eccles,[13] and some of the recent symposia[17,50] are helpful in indicating the directions of current research. The variety of specific theories which have been proposed in recent years are exemplified by those of Hebb,[24,26] Lashley,[35] Culbertson,[9] Hydén,[29] Wechsler,[58] Briggs and Kitto,[6] Gaito,[20] Smith,[55] Milner,[39] and Roy.[47] These theoretical treatments vary widely in mathematical rigor and in the range of memory phenomena which they attempt to encompass. Unfortunately, those which attempt to treat the psychological phenomena most comprehensively, such as Hebb's theory, are apt to be lacking in rigor, while the most rigorous of them (such as Roy's ingenious model) are apt to be superficial and unconvincing psychologically. Many of these "theories" are in fact, no more than suggestions of a possible approach, while others, such as the imaginative and currently fashionable notion of a "tape recorder molecule" in the cellular RNA, seem to do such violence to our basic conceptions of both physics and physiology (without, in fact, satisfactorily explaining any phenomena of memory) that they can only be regarded as a desperate attempt to fill a theoretical vacuum.

The experimental literature on memory physiology, which used to be concerned chiefly with the effects of lesions and surgical ablation of brain tissue, has recently begun to yield a number of interesting reports on electrophysiological and neurochemical influences in conditioning and learning experiments. Some of this literature is covered by the reviews mentioned in the last paragraph. Representative of recent contributions are those of Gerard, Chamberlain, and Rothschild,[23] Hebb,[25] Scoville and Milner,[51] and Milner and Penfield.[38] The psychological literature, much of which is clearly relevant to the problem, is too extensive to summarize here. Instead, we shall present a brief list of empirical phenomena which an adequate theory should be able to account for. This list, while far from exhaustive, gives us something to aim for in evaluating a proposed model. Since our primary interest is in human memory, the phenomena listed are all to be found in man; it seems likely that if we can account for these, then a simplified or modified version of the model is likely to prove applicable to memory phenomena in simpler nervous systems as well.

The primary phenomena which we would like to explain include the following:

1. Ability to recapitulate past experience in proper temporal order.
2. Selective recall; effects of "cognitive set," attention, and suggestion.
3. "Free association" (dreams; transitions and jumps between remembered events).
4. Retention and subsequent recall of originally "unnoticed" events.
5. Poor memory for sequences with low diversity (e.g., strings of digits), in contrast to sequences of nonrepetitive events.
6. Modification of stored information (cf. Bartlett[2]).
7. Effect of practice on accuracy of sequential recall.
8. Effects of "reinforcement" (pleasure, pain, reward, punishment) on memory.
9. Forgetting (transient and permanent).
10. "Repression," psychogenic amnesias, and subsequent recovery of memory.
11. Heightened accessibility of memory under hypnosis.
12. Posthypnotic suggestion.
13. Extra-high stability of early memory.
14. Low stability of recent memory in senility.
15. Lapse of memory during sleep or unconsciousness.
16. Retrograde amnesia (due to shock, cold, concussion, epileptic convulsion).
17. Recovery from retrograde amnesia in original temporal order.
18. Consolidation time (brief period following an event during which shock or trauma leads to irrecoverable loss of memory).
19. Hallucinatory recall of sequences under temporal lobe stimulation (Penfield[42]).
20. Effects of localized lesions and electrical stimulation in aphasia, agnosia, and related disorders (cf. Penfield and Roberts[43]).
21. Distributed memory and functional equivalence of cortical regions (cf. Lashley[35]).
22. Incapacity for retention of new experience, without interference with recall of old experience, temporary memory, or motor learning, following hippocampal lesions (cf. Milner and Penfield,[38] and Scoville and Milner[51]).

In examining the above list, it can be seen that the phenomena have been arranged in a rough sequence from purely psychological to primarily physiological ones. The list emphasizes qualitative, rather than quantitative, phenomena. On some of the most important quantitative questions, such as the total amount of information stored in human memory, investigators are hopelessly at odds, ranging from estimates of 1.5×10^6 bits (Miller,[37]) to 10^{21} bits for a model which assumes storage in protein molecules (von Foerster[57]). An interesting review of these estimates has been presented by Schaefer,[49] while some of the theoretical considerations are discussed by Brown.[7] Until recently, it seemed to this writer unlikely that such a thing as continuous and complete recording of sensory experience could be made physically plausible. The model proposed by Culbertson, for example,[9] would require about 3×10^8 modifiable connections to record one second of visual experience, or about 10^{18} connections to record continuously for the duration of a human lifetime (taking this to be about 100 years). Nonetheless, the results of the present theory force us to reconsider the possibility of almost complete recording, as will be seen from the numerical results in the following sections.

Despite the many unanswered questions which may be raised in connection with the above items, few investigators would be likely to deny that these are empirically well-established phenomena; consequently, a brain model which seems intrinsically incapable of dealing with them must be judged inadequate as a theory. It will clearly be impossible to deal with all of this evidence satisfactorily in the short space of this paper; a more comprehensive discussion will be forthcoming in Ref. 46. Nonetheless, even in this brief space we hope to show that a considerable number of the phenomena on our checklist can find plausible explanations in terms of the proposed model.

Before concluding these introductory remarks, a word is in order concerning different types of memory. The word "memory" has been used to cover a wide range of empirical observations, including conditioned reflexes, perceptual learning, the learning of goal-directed behavior, the temporary storage of information (such as telephone numbers), and the retention of experience from the remote past. It has become increasingly clear to those working in the field that we are probably dealing not with a single mechanism but with a variety of different mechanisms. We should not necessarily expect that the mechanism which enables a flatworm to modify a tropic reaction to light[30] is the same mechanism which enables an actor to recite *Hamlet*. In particular, the work of Milner, Penfield, and others on the effects of hippocampal lesions (mentioned in Item 22 of the preceding list) suggests that a clear distinction should be made between the mechanisms of experiential recording, experiential recall, short-term memory, and the learning of motor skills in man, since it is possible to completely obliterate one of these capabilities without, in any way, interfering with the others. The distinction between "short-term" and "long-term" memory has now been rather widely accepted among psychologists (cf. Hebb,[25] Milner,[39] and Konorski[34]). There is, in fact, a popular belief that some form of temporary "dynamic storage" (e.g., a reverberating trace system) is a necessary precondition for the establishment of a permanent recording. This belief has been fostered by the idea that a permanent change (such as synaptic growth) must take a considerable time to establish, and by the consolidation period (Item 18 on the above list) which is revealed by studies of amnesia. Actually, the evidence for such a dependent relationship between temporary and permanent storage mechanisms seems much too scanty to be assumed without question—particularly in view of the persistence of temporary memory after the long-term recording mechanism has been incapacitated by hippocampal ablation. In the present theory we shall not require an "active" short-term memory mechanism to precede permanent recording; the consolidation time, in this system, comes about as a result of slow chemical reactions which do not require an "active trace" to support them.

This paper, then, will concentrate on the problem of the long-term storage of experience, and the mechanism which enables us to retrieve information about past events. Some of the more "primitive" types of memory (such as the association of responses to stimuli, and certain types of perceptual learning) have been demonstrated in earlier work on perceptron theory.[5,44,45] Up to this time, however, none of these networks have met the challenge of being able to recapitulate a sequence of experienced events, no matter how well the events may have been "recognized" at the time they occurred. The model which is proposed for this will now be considered in detail in the following section.

DESCRIPTION OF THE MODEL

A neural network, capable of learning to give recognition responses to sensory patterns with some degree of generalization to "similar" stimuli, may take the form shown in Fig. 1. For the sake of explicitness, we shall take this model (representing a fairly general perceptron) as the starting point for the development of a sequential memory. It should be borne in mind, however, that the basic principles of the memory model could work equally well with a number of other "neural networks" (such as Widrow's Adaline or Madaline[59]) as the perceptual part of the system. Networks of the type shown here have successfully learned such tasks as alphabet character recognition and speech recognition. In its simplest form (the "simple perceptron") the model shown in Fig. 1 is reduced to a three-layer network, with a single response unit, and a "retina" of sensory points connected directly to the set of internuncial neurons, or *A*-units (association units). The *A*-units are threshold elements, which respond to any combination of input signals whose sum exceeds the *A*-unit threshold. In more sophisticated models, the intervening network (stimulus-transformation network) acts as a recoding system, which may detect such features as straight lines or edges in the stimulus pattern, transmitting only information about these important features to the *A*-units.[45] Short-time sequences, rather than momentary stimuli, may form the input patterns; these may be encoded in the association system as a nontemporal (spatial) pattern by means of a distribution of transmission delays in the *S* to *A* network, or by means of a closed-loop cross-coupled network (either in the *A*-system itself or prior to it), or else by means of a combination of "On" and "Off" neurons in the early layers which signal the onset and termination of the activity induced by a moving or changing stimulus.[44] There is increasing evidence that the last of these three mechanisms may be largely responsible for motion detection in the cat's visual system (Hubel and Wiesel[27]).

Whatever the preliminary transformations may be, we shall be chiefly concerned with the succession of states induced in the association system by the sequence of stimuli from the environment. Each activity state of the *A*-units (represented by the set of units which are "on" or active at the time) bears information about the stimulus, or recent succession of stimuli, which has just occurred. It is now well established that this information is sufficient to permit an arbitrary identifying response to be associated either to a particular stimulus or to a class of similar stimuli, if the network is properly designed.[44] When an *R*-unit receives a superthreshold signal from the set of active *A*-units, it emits a signal which serves to identify the current stimulus.

The assignment of discriminating responses to stimulus classes is generally carried out by means of an "error-correction procedure" in which the weights of connections from the *A*-units to the *R*-unit (or *R*-units) are modified, whenever an "error" occurs in the response of the perceptron. The "motivational system" which is responsible for recognizing such errors and applying the necessary reinforcement to the connections may consist of an outside experimenter or trainer, or it may be built into the network itself, as a "reinforcement-control system".[44] In any case, it is not shown in Fig. 1 or in any of the subsequent figures; the memory processes with which we will be chiefly concerned are automatic, and do not depend on the "reinforcement-control system" in any necessary way. An alternative training procedure (the *S*-controlled procedure) assumes that reinforcement will be supplied continuously, modifying the weights of the *A*- to *R*-unit connections for each stimulus in a direction which will tend to give the correct response for that stimulus. For this *S*-controlled procedure, a detailed mathematical analysis of the resulting distribution of *R*-unit input signals is now available (Joseph,[32] and Rosenblatt[44]). Since the signal distributions resulting from the error-correction procedure have been less well analyzed than the signal distributions in the *S*-controlled procedure, Joseph's analysis of the *S*-controlled procedure will be used in this paper to describe the signal distribution to the *R*-units which might be expected to exist after a period of training.

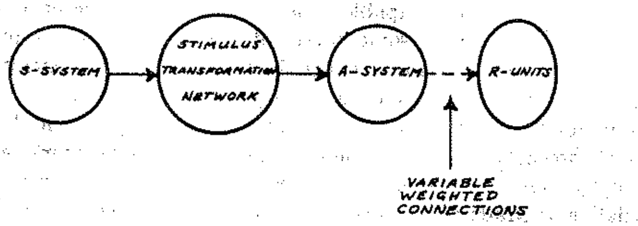
S-SYSTEM
STIMULUS TRANSFORMATION NETWORK
A-SYSTEM
R-UNITS
VARIABLE WEIGHTED CONNECTIONS

Figure 1

The sequential memory model will operate, basically, by reconstituting the succession of *A*-unit activity states which occurred when the original experience took place. This reconstitution is, generally, far from perfect, but it will be shown that it can be made close enough to the original activity states to permit the previously learned responses to occur, or, alternatively, to learn new responses to stimuli in retrospect, which will then generalize satisfactorily to stimuli appearing in the environment. In order to do this, an auxiliary network is necessary, as shown in Fig. 2.

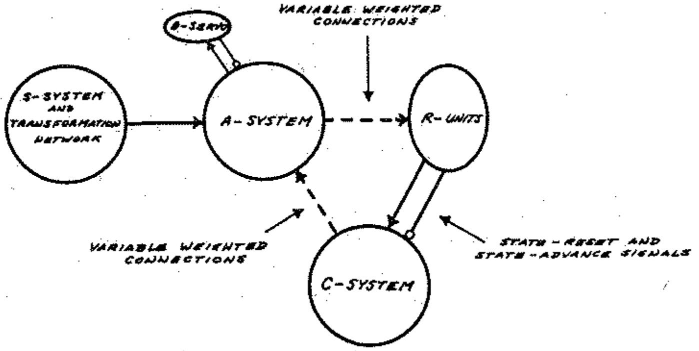
B-SERVO
VARIABLE WEIGHTED CONNECTIONS
S-SYSTEM AND TRANSFORMATION NETWORK
A-SYSTEM
R-UNITS
VARIABLE WEIGHTED CONNECTIONS
STATE-RESET AND STATE-ADVANCE SIGNALS
C-SYSTEM

Figure 2

In this figure, as in the subsequent ones, broken arrows are used to represent adaptive connections (with variable weights), while solid arrows represent fixed connections. The circles represent sets of "neurons" or functionally analogous units. A normal arrowhead generally represents excitatory connections (or mixtures of excitatory and inhibitory connections), while a small circle in place of the arrowhead represents inhibitory connections.

The two main additions to the system shown in the previous figure are the threshold servomechanism for the association system, and the C-system, or clock network, which has variable connections to the A-units, and input connections from the R-units. The Θ-servo is simply a negative feedback system which tends to maintain a constant level of activity in the association system. It might consist, physiologically, of a set of cells whose input connections are drawn from the whole of the association network, and whose output connections deliver an inhibitory signal to all A-units, which increases with the magnitude of the input signal. With such a control mechanism the association system will tend to find and maintain a constant level of activity despite changes in the distribution or intensity of input signals. Such mechanisms have been proposed previously by Beurle[3] and Rosenblatt.[44]

The A-units may receive signals from two sources, apart from the servo system itself. Normally, their chief input source would be the sensory network, which is assumed to send strong signals to the A-units whenever sensory events occur. The second source is the set of adaptive connections from the C-network (the functioning of which will be elaborated shortly). These connections, however, are assumed to be limited to weights which are considerably smaller in magnitude than the weights of the S- to A-connections. Consequently, as long as sensory signals are arriving at the A-units, the state of the association system will be "S-determined," the signals which might be coming in simultaneously from the C-system constituting only a negligible perturbation in the total input signals. Under the action of the Θ-servo, the A-units will act essentially like high-threshold units in a simple perceptron, and the C-system will have little or no influence on the operation of the primary information channel, from S to A to R. In this state (as long as sensory inputs continue) the perceptron can be trained or interrogated in the usual fashion, and all previous analyses of such performances remain applicable. On the other hand, when sensory signals cease (either due to lack of environmental stimulation or due to an active cutoff mechanism in the perceptron itself, which might be controlled by one of the R-units) the Θ-servo will immediately act to lower the thresholds of the A-units until previous activity levels are restored. Under these conditions, the relatively weak signal component coming from the C-system becomes the primary determinant of the state of the A-system, and the A-units will respond to the C-network as if it were an alternate sensory field.

We must now consider the C-system itself in greater detail. Several alternative organizations are illustrated in Fig. 3. The C-network, as its name suggests, operates as a "clock" for the memory of sequences. This clock may either be synchronous (progressing through a sequence of states at a rate which is independent of external events) or asynchronous, in which case it advances from one state to the next only when a suitable trigger-event occurs to make it do so. A synchronous clock is exemplified by a simple cross-coupled network (Fig. 3*a*) which will advance through a succession of states, each determined by the preceding state, with a speed which depends only on the transmission time of the connections and synaptic delays. The Θ-servo acts to prevent "blowups" or extinction of activity. Random networks of this type have been analyzed in Chap. 18 of Ref. 44, and elsewhere. While it would be quite possible for our model to operate with such a simple mechanism, the asynchronous clock, which permits the events constituting the recorded sequence to occur at one rate and to be recalled later at a different rate, is inherently of much greater interest.

Two variations of the asynchronous clock are shown in Figs. 3*b* and 3*c*. In each case, the C-network is subdivided into two sets (or layers) of neurons. One layer consists of "On" neurons, which deliver a sustained burst of impulses in response to an excitatory input signal; the second layer

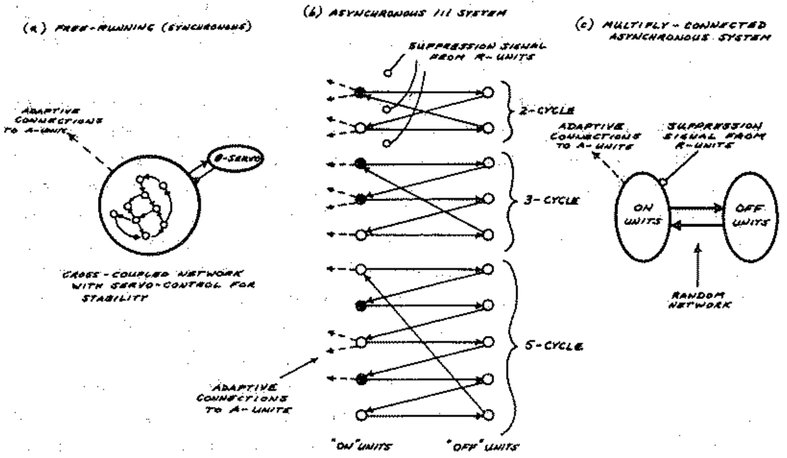
(a) FREE-RUNNING (SYNCHRONOUS)
ADAPTIVE CONNECTIONS TO A-UNITS
Θ-SERVO
CROSS-COUPLED NETWORK WITH SERVO-CONTROL FOR STABILITY
(b) ASYNCHRONOUS 1/1 SYSTEM
SUPPRESSION SIGNAL FROM R-UNITS
2-CYCLE
3-CYCLE
5-CYCLE
ADAPTIVE CONNECTIONS TO A-UNITS
"ON" UNITS
"OFF" UNITS
(c) MULTIPLY-CONNECTED ASYNCHRONOUS SYSTEM
ADAPTIVE CONNECTIONS TO A-UNITS
SUPPRESSION SIGNAL FROM R-UNITS
ON UNITS
OFF UNITS
RANDOM NETWORK

Figure 3

consists of "Off" neurons, which are effectively inhibited during an input signal, but deliver a brief burst of high-frequency impulses when the input ceases. Such "On" and "Off" neurons are known to exist in the cerebral cortex, and extensive recordings of their activity are available (cf. Jung,[33] Florey,[19] Hubel and Wiesel,[27] Sandel and Kiang[48]). The physiological mechanism underlying the Off responses is not well understood. In some cases it is possible that the cell is part of a more complex network which, in fact, delivers excitatory signals to it upon the termination of a stimulus. A more plausible explanation in most cases, however, is that there is an intracellular servomechanism tending to keep the membrane potential at its normal, resting level, despite the effect of transmitter substances which either hyperpolarize or depolarize the membrane. Such a cell would tend to deliver a brief burst after an excitatory stimulus began, which would terminate as the servomechanism began to operate, and would similarly deliver a burst after the cessation of an inhibitory input signal, since its membrane would suddenly be left under the depolarizing influence of the servo system, in the absence of the hyperpolarizing effect which it was combating. The observations of Sandel and Kiang[48] are particularly suggestive of such a mechanism. A particular type of long-lasting after-discharge following stimulation has been studied by B. D. Burns,[8] who attributes it to a network of "Type B" neurons. It seems likely that these cells, once stimulated, will continue to discharge indefinitely until some inhibitory signal occurs to cut them off. The mechanism is likely to depend on differential rates of repolarization in different parts of the cell, which Burns has demonstrated could lead to a continuing volley of impulses. It is quite tempting, although not essential, to identify the "On" units in Fig. 3 with Burns' Type B neurons.

The manner of operation of the asynchronous clock network can best be understood from Fig. 3*b*. Assume that those "On" units which are filled in solidly in the diagram are active at the present time. They will continue to emit impulses (if they are of the Burns Type B variety) until some inhibitory signal arrives to cut them off. This inhibitory signal is provided by an "On" burst or "Off" burst of short duration, from any of the R-units, signaling some change in the response of the perceptron, and thus the beginning or end of a distinguishable event. If we assume that the coupling from the R-units to the C-network is dense enough and powerful enough, then any change in response will momentarily quench the activity of the On units in the C-system. During all of the time that these On units have been firing, however, they have not only been transmitting signals back to the A-units (by way of the variable connections, which will soon be discussed in detail); they have also been sending "priming signals" to the Off units, which thus begin to fire as soon as the On units are cut off. This Off burst occurs only in the subset of cells which were connected to the active On units. These cells will immediately transmit excitatory signals back to the On layer, activating a new subset of On units, which will then continue to fire until it is finally quenched by the next change in the R-units. Thus the C-system will advance through a deterministic succession of states, changing abruptly to a new state whenever the response of the perceptron is altered in a significant fashion.

Ultimately, since the number of C-units must be finite, the network must return to its initial state, and the cycle will repeat. It can readily be seen that the activity configuration shown in the network of Fig. 3*b* will repeat every thirty steps, since it factors into prime-numbered cycles of durations 2, 3, and 5, respectively. Such a network organization can easily yield cycles of immense duration with only a small number of neurons. Its cycle-time will be equal to the product of all of its prime subloops which are not "silent" (no units active) or "saturated" (all units active). Moreover, a large number of different initial conditions are possible which will lead to totally different "life histories" for the state sequence of the network. For example, an initial state with only one active neuron in each loop will yield a cycle which must be distinct, in all its states, from that which results from an initial condition with two active neurons in one or more of the loops.

While a network of this sort would be quite useful for engineering purposes, and quite satisfactory so far as the operation of the memory system is concerned, it is clearly unbiological in its requirement of successive prime numbers for the order of its subcycles, and in its 1-to-1 connectivity. A slightly more plausible network results if we retain the 1 : 1 constraint, but allow the connections between On and Off units to be made at random. A number of computer studies have been made by Trevor Barker and the writer to determine the expected cycle times for initially random activity states of such 1 : 1 networks. Based on samples of 1,000 networks of each size, with 50 percent of the units active, the following cycle times were obtained:*

Number of "On" units in network (N_c)	Mean cycle time	Variance of cycle times	Minimum cycle time	Maximum cycle time
10	8.91	4.96	1	30
100	30,677	164,370	~ 40	~4 × 10^6
200	1,518,299	10,057,894	~150	~3 × 10^8
400	114,963,471	93,358,746	~250	>2 × 10^9

* The numbers in the first line are exact theoretical values; the subsequent lines are estimated from samples of 1,000 cases.

Thus it is clear that with a network of biological size, numbering at least a few thousand C-units, the probability that a state would repeat during a long period is very slight indeed, despite the random choice of connections and starting state.

A more plausible biological model is shown in Fig. 3c. Here each On-unit may be connected to many Off-units, and vice-versa, the only important constraint being that On-units should not be connected to one another (or at least not strongly) in order to prevent the activity pattern from spreading to all units in the network. A Θ-servo could again be employed to govern the level of activity. Unlike the 1 : 1 network there is the possibility with the multiply-connected model that two different starting states might lead into the same activity state of the network. This and other complexities make the cycle times exceedingly hard to estimate, but from previous observations of cross-coupled systems in computer simulation programs (e.g. Farley and Clark[16]) as well as from theoretical considerations, it seems likely that cycles will tend to be at least as long as in the 1 : 1 network, and possibly much longer. In all that follows, it will be assumed that the C-system is of sufficient size that the likelihood of a state repeating itself, without the network having been deliberately reset, is entirely negligible.†

Although the successive states of the C-system form a deterministic sequence, each state being a predictable consequence of the preceding one, their interrelationships (particularly the measure of the intersections of active sets at different times) are, generally, indistinguishable from those that would pertain to a collection of randomly chosen states. Thus, if the initial state in a 1 : 1 network, for example, is selected at random (with a probability P of any unit being on or off) the measure of the intersection between the initially active set and any following active set will have an expected value of P^2. Such a sequence, then, can be considered a "quasi-random-state sequence," since the measures of all active sets and their intersections will have a distribution which is indistinguishable (except in specially contrived cases) from a random-state sequence. This property will be seen to be of great importance in the subsequent analysis of the memory system.

It now remains to see how the states of the C-system can be made to induce a succession of states in the A-system corresponding to a recorded sequence of stimuli. For this we must specify more precisely the modification mechanism of the C-unit to A-unit connections.

† Despite the arguments given above for its plausibility, the writer is convinced that the C-system, in its present form, is the least plausible part of the model. At the present time a more "realistic" form of C-system is being investigated, which makes use of a statistically homogeneous network containing only one kind of neuron, with states represented by frequency modes rather than On/Off activity states. This will be reported in Ref. 46.

We assume that each A-unit receives connections from a fraction M of the "On" units in the C-network. The connection system from C to A is a many-to-many system, the only important constraint being that the choice of connections to particular A-units should be statistically independent of the particular sets of C-units which are likely to be active in different "clock states." To be explicit, it will be assumed that the connections to each A-unit originate from a set of MN_c points chosen at random with a uniform probability distribution (where N_c is the total number of units in the "On" layer of the C-system). Thus, if a fraction Q_c of the C-units are active in any given state, it is expected that a fraction MQ_c are actually transmitting signals to any particular A-unit.

The modification of connections takes place according to a rule which has been called the gamma-system in perceptron terminology. Chiefly because it permits us to obtain more rigorous equations, the gamma-system will be employed for the A- to R-unit connections as well as the C- to A-unit connections. For the A-R connections, however, the simpler alpha-system, which changes only the weights of connections from active units, would undoubtedly work equally well.[44]

The γ-system reinforcement procedure is conservative in the weights of connections to a given unit; that is, the sum of the weights of all input connections to any A-unit (or R-unit) must remain constant. Therefore, if the active connections should gain in weight at time t, the inactive connections must lose a compensating amount. γ-systems have been defined and analyzed in detail for simple perceptrons in Ref. 44. In the case of the C- to A-unit connection weights, the "reinforcement procedure" operates as follows:

Let w_{ij} = weight of connection from the ith C-unit (c_i) to the jth A-unit (a_j).

$c_i^*(t)$ = activity state of c_i at time t. $c_i^* = 1$ if c_i is active, 0 otherwise.

$a_i^*(t)$ = activity state of a_i at time t. $a_i^* = 1$ if a_i is active, 0 otherwise.

η = unit of reinforcement, generally taken as 1.

Q_c = fraction of C-units active.

Two variations of the reinforcement rule will be considered: the *asymmetric model*, which modifies only connections to active association units, and the *symmetric model*, which modifies connections to inactive A-units as well as to active ones.

For the asymmetric model, the change in the weight w_{ij} at time t takes the form

$$\Delta w_{ij}(t) = w_{ij}(t + \Delta t) - w_{ij}(t)$$

$$= \eta \cdot a_j^* (t) \left[c_i^* (t) - \frac{1}{MN_c} \sum_{i=1}^{MN_c} c_i^* (t) \right] \qquad (1a)$$

$$E\Delta w_{ij}(t) \doteq \eta \cdot a_j^* (t)[c_i^* (t) - Q_c] \qquad (2a)$$

The index i in this equation ranges over the set of $C-$units connected to a_j.

For the symmetric model, the corresponding equations are

$$\Delta w_{ij}(t) = \eta \cdot (-1)^{a_j^* (t)+1} \left[c_j^* (t) - \frac{1}{MN_c} \sum_{i=1}^{MN_c} c_i^* (t) \right] \qquad (1b)$$

$$E\Delta w_{ij}(t) = \eta \cdot (-1)^{a_j^* (t)+1}[c_i^* (t) - Q_c] \qquad (2b)$$

That is, for an active A-unit, the change in weights is the same in both models, but for an inactive A-unit (in which case there is no change in the asymmetric model) Δw_{ij} is the negative of what it would be for an active unit.

In the recording of a memory sequence, the following succession of events occurs. It is assumed that the C-system is set to some initial activity state, by any one of several mechanisms which will be discussed in more detail later. This could be achieved, in the simplest case, by activating one of the R-units which forces the on-units of the C-network to the desired starting condition. This R-unit, in turn, could be trained to respond to a starting command, such as the name of the recorded sequence. A starting mechanism of this sort is suggested in Fig. 2. The initial weights of C-to-A connections are assumed to be zero.

With the C-system in its initial state, the first stimulus pattern of the sequence appears in the sensory system, and induces a corresponding activity state in the association units. Say, for example, the first stimulus is a triangle. The set of A-units responding to this triangle will then have their connections from the active C-units augmented in value, according to equation (1a) or (1b). As long as the triangle remains on the "retina," signals transmitted from the C-units to the A-units will tend to be ignored, due to the action of the Θ-servo, and the relatively high weights of connections from the sensory system. On the other hand, if the same C-state should recur, without the presence of a retinal input, the Θ-servo will lower the effective thresholds of the A-units, and the augmented connection weights to the previously active A-units will tend to reactivate the same set of units which responded to the triangle. In the case of the asymmetric model, there will be no systematic attempt to turn off the "improper units," which did not respond to the triangle, but the Θ-servo will tend to find a level at which only the units receiving the strongest input signals will be reactivated, which has essentially the same effect. In the case of the symmetric model, there is an additional tendency to turn off the improper units, due to the negative weights which have been acquired by the connections from active C-units to inactive A-units, Eq. (1b).

As soon as the triangle is replaced by the next stimulus (say a square) which is sufficiently different so that the response of the perceptron changes, the C-system will advance to its second state, which we have seen to be statistically independent of the first, although it is a deterministic consequence of it. Due to this statistical independence and the use of the γ-system, it will be shown in the following section that the expected value of the signal now received by any A-unit from the C-system is equal to zero. Consequently, the modifications of the connections which now take place to the set of A-units responding to the square will have the same effect (except for a slight noise effect) as if no previous memory had been recorded.

If the square, in turn, is replaced by another triangle the change in response (whether correct or not is immaterial) will cause the C-system to advance to its third state, from which the expected signal to the A-units will again be zero. A new change in weights then occurs as before. This process continues indefinitely until the C-system either recycles (an unlikely possibility) or is deliberately reset.

To see how the system acts in recall, suppose the response which resets the C-system is evoked, followed by a "silent period," during which no sensory inputs occur. The Θ-servo, striving to normalize the activity level in the A-system, now lowers the thresholds to the point where the A-units begin to respond to the C-unit signals. As we have seen, the first state of the C-system will tend to reactivate the set of A-units responding to the first triangle (without any interference, other than random-noise effects, from any subsequently recorded memory). As soon as this state is, in fact, reconstituted in the association system, however, the triangle response should occur, and this response will advance the C-system to its next state. This state induces the A-unit activity pattern corresponding to the square, and as soon as this is responded to, the C-system is advanced again. Thus the association states corresponding to the entire sequence of sensory events tends to be reconstituted, in proper temporal order. If the states are reconstituted accurately enough they can be used for teaching the perceptron new discriminations, in retrospect, or for applying subsequently learned discriminations to events which were improperly recognized at the time they occurred. None of this interferes with the sequential memory system, which is independent of changes in the A- to R-unit network.

Due to the fact that the expected interaction between recorded events is zero, the sequences which can be stored may be extremely long. Ultimately, noise effects, which show up as a gradually increasing variance in the transmitted signals, grow to such a degree that they effectively mask the residual traces of previous memory, and the system saturates. Before this happens, the accuracy with which the association states are reconstituted gradually diminishes, and consequently the discriminatory responses which occur to remembered stimuli become less and less accurate. In evaluating the performance of this model, the most important question is the probability that a discriminatory response to a remembered stimulus is correct, after a long history of experience has been recorded. The estimation of this probability is the task which is undertaken in the following section.

ANALYSIS OF PERFORMANCE PROBABILITIES

The measurement of memory performance will be based upon the following experiment:

Assume that the perceptron sees a long sequence of stimulus patterns, $S_1, S_2, S_3 \ldots, S_t$. Each stimulus persists for an equal period of time, Δt, which for convenience is set equal to unity. Among these n stimuli, there is at least one occurrence of a stimulus S_x which the perceptron has been taught to recognize by emitting the response R_x. It is assumed that the perceptron has been taught to suppress the response R_x for all stimuli other than S_x. The level of performance of the perceptron in discriminating S_x from other stimuli is to be treated as a parameter of the problem. At the start of the sequence of n stimuli, the C-system is initialized, and recording goes on throughout the sequence. The C-system is then set back to its initial state, with no external stimuli present, so that the perceptron begins to recapitulate the remembered sequence. Suppose the stimulus S_x originally occurred at time t_x (measured from the start of the sequence). Then when the C-system has advanced to the t_x-th state, the response R_x should occur. We wish to calculate the probability $P(R_x)$ that this response does, in fact, occur at the appropriate point in the recapitulated sequence.

The perceptron is characterized by the choice of the symmetric or asymmetric reinforcement rule [Eqs. (1a) or (1b)] and by the following parameters:

N_a = number of A-units
N_c = number of C-units in the "on" layer
M = fraction of C-units connected to each A-unit ($0 < M \leq 1$)

Q_a = proportion of A-units activated by a stimulus (i.e., the measure of the A-unit activity level maintained by the Θ-servo).
Q_c = proportion of C-units active in any given clock state.

For a plausible biological system, we would require N_a and $N_c \leq 10^9$, MN_c (the number of connections to an A-unit) $\leq 1{,}000$, and Q_a and Q_c probably no greater than 0.1 or 0.2.

In the analysis of S-controlled reinforcement procedures,[44] it was shown that the probability of a correct response could be very closely approximated by assuming a normal distribution of input signals to the R-unit. The probability of a correct response to stimulus S_x (when S_x is actually present on the retina) is then given by

$$P(R_x|S_x) = \Phi\,[E(u_x)/\sigma(u_x)] \tag{3}$$

where $E(u_x)$ = expected value of signal to R-unit when S_x is present
$\sigma(u_x)$ = standard deviation of the signal u_x
$\Phi(z)$ = cumulative normal distribution function, from $-\infty$ to z, i.e.,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\,dx$$

The above distribution of signals may be taken over a collection of different perceptrons, or over a set of training sequences, or over a set of possible choices of the test stimulus, S_x. For any given perceptron and training sequence, of course, the actual signal is deterministic, and, assuming the response to be correctly learned, we could set $E(u_x)/\sigma(u_x) = \infty$, in applying the above equation. Since we are interested in studying the effects of imperfect recognition, or low levels of discrimination performance, upon recall, however, we will permit an arbitrary choice of the ratio $E(u_x)/\sigma(u_x)$ to characterize the initial performance level of the perceptron. The computation of this ratio for a number of different kinds of discrimination experiments with γ-system reinforcement is presented in Chap. 8 of Ref. 44.

A special case which is likely to be of interest is that of a perceptron which has been trained on only two stimuli, S_x and S_y, where the perceptron is taught to give the response R_x for S_x, but to suppress this response for S_y. In this case, if S_x and S_y are not identical or nearly identical stimuli, correct performance is virtually certain, and it is safe to take $E(u_x)/\sigma(u_x) = \infty$.

Suppose, then, that the perceptron has been trained to some level of performance for which $E(u_x)$ and $\sigma(u_x)$ are known. When the memory sequence has been recorded and the state of the C-system is restored to the conditions which existed when S_x appeared on the retina, some set of A-units, more or less similar to those activated by S_x, will be activated by the signals from the C-units. In particular, suppose the stimulus originally activated a set of $N^+ = Q_a N_a$ "proper units," the remaining $N^- = N_a - N^+$ units being designated "improper units." Of the N^+ proper units, the C-system now activates n^+ proper units; it also activates n^- improper units. If $n^+ = N^+$, and $n^- = 0$, this means that the original A-unit state has been reconstituted exactly, and the probability of a correct response would be just what it was if the stimulus actually appeared on the retina. In general, however, this condition is not likely to occur. For a given set of n^+ proper units reactivated, and n^- improper units activated, we must determine the new probability of obtaining a correct response.

To estimate this probability, we assume, first of all, that the set of proper A-units reactivated are not systematically related, in any way, to the weights of their A-R connections. In this case, the n^+ proper units would be expected to receive a fraction n^+/N^+ of the total signal from the proper set, $E(u_x)$. Similarly, the n^- improper units activated are assumed to be unrelated to the weights of their A-R connections, so that they will be expected to receive a fraction n^-/N^- of the total weight of the N^- improper A-units. But, if we assume that a γ-system has been employed for training the A-R network, the expected sum of the weights of the improper set must be equal to $-E(u_x)$, so that the expected value of the regenerated signal to the R-unit will be

$$\left[\frac{n^+}{N^+} - \frac{n^-}{N^-}\right] E(u_x) \tag{4}$$

By the same reasoning which was just applied to the expected value of the regenerated signal, the variance will likewise be expected to redistribute in a uniform fashion over the active units. (This will be rigorously true if the sets of A-units activated by different stimuli are statistically independent of one another, as occurs in a "binomial model" simple perceptron in an environment of random stimuli. In that case, the variance of the weight of any given A-unit a_i will be the same as the variance for any other A-unit, a_j.) For the n^+ units of the proper set, the new variance will be $(n^+/N^+)\,\sigma^2(u_x)$. The N^- improper units are inferred to have a total variance of $(N^-/N^+)\,\sigma^2(u_x)$, and consequently the variance of the n^- units which are actually active will be $(n^-/N^+)\,\sigma^2(u_x)$. Combining these components, the variance of the reconstituted signal becomes

$$\frac{n^+ + n^-}{N^+}\,\sigma^2(u_x) \tag{5}$$

and, taking the ratio of Eq. (4) and the square root of Eq. (5) we obtain the expression

$$Z(n^+, n^-) = \frac{\dfrac{n^+}{N^+} - \dfrac{n^-}{N^-}}{\sqrt{\dfrac{n^+ + n^-}{N^+}}} \cdot \frac{E(u_x)}{\sigma(u_x)}$$
$$= \frac{\dfrac{n^+}{Q_a N_a} - \dfrac{n^-}{(1 - Q_a)\,N_a}}{\sqrt{\dfrac{n^+ + n^-}{Q_a N_a}}} \cdot \frac{E(u_x)}{\sigma(u_x)}$$

and consequently (given n^+ and n^-), the probability of a correct response is

$$P(R_x) = \Phi\,[Z(n^+, n^-)] \tag{6}$$

Actually, however, any combination of $n^+ \leqq N^+$ and $n^- \leqq N^-$ might possibly occur. Let $P(n^+, n^-)$ be the probability that some particular set of n^+ proper units and n^- improper units is activated. This probability should be the same for any choice of the n^+ proper units and the n^- improper units, due to symmetry considerations. There are $\binom{N^+}{n^+}$ ways of choosing the proper set, and $\binom{N^-}{n^-}$ ways of choosing the improper set. Thus the general equation for $P(R_x)$ takes the form:

$$P(R_x) = \sum_{n^+=0}^{N^+} \sum_{n^-=0}^{N^-} \binom{N^+}{n^+}\binom{N^-}{n^-} P(n^+, n^-)\,\Phi[Z(n^+, n^-)] \tag{7}$$

The only unknown quantity in this equation is $P(n^+, n^-)$, which must now be analyzed.

Let us consider two extreme possibilities. First, the signals to different A-units from the C-system may be totally uncorrelated with one another, in which case $P(n^+, n^-)$ will be the product of the probabilities of activating each of the n^+ proper units and the n^- improper units individually. At the other extreme, the signals might all be perfectly correlated with one another, in which case either every A-unit will be turned on, or every A-unit will be turned off jointly, the probability of activating the entire set being the same as the probability of activating any one individually. In the first of these cases, due to a sort of "majority decision" effect, performance will be greatly improved by having a large number of A-units, while in the second case the probability of a correct response from a great number of units will be no better than the probability of a correct response with only one or two units in the system. Thus the correlation between the signals to different A-units, transmitted from the C-system, is seen to be a consideration of prime importance. The correlation between the signals to units a_i and a_j, measured over the set of C-system states, will be called ρ_{ij}. This correlation is, in general, nonzero, and will be different for the symmetric and asymmetric models as will be seen shortly.

Assuming that the correlations can be obtained, it will still be necessary to know the probability of turning on a given proper unit or improper unit, a_j, in order to develop an equation for $P(n^+, n^-)$. If we again assume a Gaussian distribution for the input signals to the A-units from the C-system (which will, in fact, be an extremely close approximation, since the signals consist of sums of a great number of increments and decrements which are added or subtracted at random (or pseudo-random) times during the recorded sequence) we have an analogous expression to Eq. (3),

h :

$$P\{a_j^* = 1\} = \Phi\left[\frac{E(u_j) - \theta_j}{\sigma(u_j)}\right] = \Phi(h) \tag{8}$$

where u_j = signal to a_j from the C-system
Θ_j = threshold of a_j

The expected value $E(u_j)$ can be considered to consist of two contributions, the first due to the reinforcement which occurred at the time of stimulus S_x, and the second due to all other stimuli in the recorded sequence. Each connection to a_j which is active at the time of recapitulation was also active at the time S_x originally occurred, since the state of the C-system is presumed to be identical. There are a total of Q_cMN_c such connections. Consequently, the expected signal contribution from S_x is

$$Q_cMN_c \cdot E(\Delta w_{ij}|c_i^* = 1). \tag{9}$$

From any other state of the C-system, however, only a fraction Q_c of the active units will be common with those which are active at the present time, since the quasi-independence of different C-states guarantees that the measure of any intersection will be Q_c. Thus from any such state other than the one previous occurrence of the present state, the expected contribution to the signal to a_j will be

$$Q_c^2MN_c \cdot E(\Delta w_{ij}|c_i^* = 1) + (Q_c - Q_c^2)\, MN_c \cdot E(\Delta w_{ij}|c_i^* = 0). \tag{10}$$

Taking the magnitude of the reinforcement increment η to be unity, and substituting in (10) for the asymmetric model, from Eq. (2a), we obtain

$$Q_c^2MN_c\, a_j^*(1 - Q_c) - (Q_c - Q_c^2)\, MN_c\, a_j^*(Q_c) = 0.$$

Similarly, substituting in Eq. (10) for the symmetric model, from Eq. (2b), yields

$$Q_c^2MN_c(-1)^{\overset{*}{a}_j+1}\, (1 - Q_c) - (Q_c - Q_c^2)\, MN_c(-1)^{\overset{*}{a}_j+1}\, (Q_c) = 0.$$

Thus the only contribution to $E(u_j)$ which survives is that given in Eq. (9). Substituting for the expectations, and taking $\eta = 1$, we obtain

$$E(u_j) = Q_cMN_ca_j^*(S_x)(1 - Q_c) \qquad \text{(Asymmetric case)} \tag{11a}$$

$$E(u_j) = Q_cMN_c(-1)^{\overset{*}{a}_j(S_x)+1}\, (1 - Q_c) \qquad \text{(Symmetric case)} \tag{11b}$$

where $a_j^*(S_x)$ = activity state of a_j in response to stimulus S_x.

The variance $\sigma^2(u_j)$ of the signal to a_j will receive an increment for every stimulus for which the connections to a_j are reinforced. In the asymmetric model, such reinforcements occur only when a_j is active; in the symmetric model, they occur for every stimulus, being equal in their expected magnitude but opposite in sign, depending upon whether $a_j^* = 1$ or 0. These differences in sign will not affect the variance, however, which will receive a fixed positive increment for every stimulus. An exact expression for the variance which results from a series of γ-system reinforcements has been obtained by Joseph,[32] and can also be found in Chap. 8 of Ref. 44. We assume here that Q_c is constant for all C-states, that the intersection between the active sets in two C-states has measure Q_c^2, and the triple intersection between three active C-sets has measure Q_c^3. For these conditions (which would apply with randomly chosen C-sets with large values of N_c) Joseph's formula for the γ'-system is applicable [Eq. (8.8) in Ref. 44]. Specifically, with appropriate changes in symbols, and assuming each C-state to occur only once, this equation becomes

$$\sigma^2(u_j) = MN_c \sum_{i=1}^{r} \sum_{k=1}^{r} r_ir_k[(Q_{ikx} - Q_c^3) - 2Q_c(Q_{ix} - Q_c^2) - (Q_{ix} - Q_c^2)(Q_{kx} - Q_c^2)] \tag{12}$$

where the indices i and k range over the set of C-states for which the connections to a_j were reinforced, r = the number of such states, and r_i (the sign of the reinforcement) is always +1 for the asymmetric model and $(-1)^{a_j^*(S_j)+1}$ for the symmetric model. Q_{ij} = the measure of the intersection between the ith and jth C-states ($= Q_c^2$ if $i \neq j$ and Q_c if $i = j$). Q_{ijk} = measure of the intersection between the ith, jth, and kth C-states ($= Q_c^3$ if $i \neq j \neq k$, Q_c^2 if exactly two indices are identical, and Q_c if $i = j = k$). Counting the number of terms with similar and dissimilar indices, assuming $x \neq i$ or k, and taking account of the signs r_i, it can be seen that all terms for $i \neq k$ vanish, and we are left with the expression

$$\sigma^2(u_j) = MN_c \sum_{i=k} [Q_c^2 - Q_c^3] = MN_c \tau (Q_c^2 - Q_c^3). \qquad (13)$$

If x = some i or k, then we must add the increment

$$\Delta\sigma^2(u_j) = MN_c[Q_c - 3Q_c^2 + 3Q_c^3 - Q_c^4]$$

to Eq. (13). For large values of τ, this increment will obviously be negligible, and Eq. (13) will be taken as the estimate of the variance throughout the following.

For the symmetric model, since reinforcement occurs regardless of the activity states of the A-units, $\tau = t$ (the total number of stimuli recorded). For the asymmetric model, τ depends on the number of stimuli by which the unit a_j was activated; specifically, $\tau = \varrho_a t$, where ϱ_a = the fraction of stimuli activating a_j. In a nonrepetitive random environment, we can assume that $\varrho_a = Q_a$, but in a systematic environment (for example, one with only two stimuli constituting a long sequence) this will not generally be true. Thus, from Eq. (11) and (13), for the two cases of interest, we have

$\hat{Q}_a$

$$\frac{E(u_j) - \theta}{\sigma(u_j)} = \begin{cases} \sqrt{\dfrac{MN_c(1 - Q_c)}{\hat{Q}_a t}}\,[a_j^*(S_x)] - \dfrac{\theta}{\sigma(u_j)} & \text{(Asymmetric case)} \\[2ex] \sqrt{\dfrac{MN_c(1 - Q_c)}{t}}\,[(-1)^{a_j^*(S_x)+1}] - \dfrac{\theta}{\sigma(u_j)} & \text{(Symmetric case)} \end{cases} \qquad (14)$$

Note that for the symmetric case, the ratio $E(u_j)/\sigma(u_j)$ for a "proper" unit is equal in magnitude, but opposite in sign to the ratio for an "improper" unit. We can assume, then, that the Θ-servo will set the threshold at a level close to zero, which would provide the best cutting-point between the units receiving high input signals and those receiving low input signals. The threshold will therefore be assumed to be exactly zero, in this model, for the following analysis. For the asymmetric case, on the other hand, the ratio $E(u_j)/\sigma(u_j)$ will be 0 for an improper unit, while for a proper unit it will have the value of the left-hand term in Eq. (14). In this case, the Θ-servo will tend to find a level halfway between the 0 signal which is expected as the input to improper units, and the signal expected by proper units. But this yields a distribution for the probability of activating proper

w.r.t. S_x

or improper units which is entirely equivalent to assuming a *zero* threshold, and a Gaussian distribution with $E(u_j)/\sigma(u_j)$ of exactly half of the magnitude shown in Eq. (14), with $E(u_j)$ having opposite signs for proper and improper units. Since this treatment permits both the symmetric and asymmetric models to be handled in an identical fashion, it will be adopted in the following analysis. Thus, the probability of activating the unit a_j is given by

$$P\{a_j^* = 1\} = \Phi(h) \qquad (15)$$

where

$$h = \begin{cases} \pm\sqrt{\dfrac{MN_c(1 - Q_c)}{4\hat{Q}_a t}} & \text{(Asymmetric case)} \\[2ex] \pm\sqrt{\dfrac{MN_c(1 - Q_c)}{t}} & \text{(Symmetric case)} \end{cases} \qquad (16)$$

The sign of h is positive for proper units, and negative for improper units.

For the case in which the correlation between A-unit input signals, P_{ij}, can be assumed to be zero for all $i \neq j$, we would immediately be able to obtain the probability

$$P(n^+, n^-) = \Phi(+h)^{n^+} \cdot \Phi(-h)^{n^-}.$$

not really independent

For the correlated case, however, we will have to face the problem of calculating the probability that each of $(n^+ + n^-)$ correlated Gaussian variables is positive. In general, this is known to be an insoluble problem. For the particular case in hand, fortunately, a solution is possible. In order to deal with this, however, we must first estimate the values of the correlation coefficients, P_{ij}.

It can be seen that there are two possible sources of correlation effects. One is the set of C-units which is connected both to a_i and to a_j. If the origins of the connections to the A-units are selected at random, this set will have an expected measure equal to M^2. (By selecting disjoint C-sets for the connections to each A-unit, the correlation can, in fact, be reduced to zero, but with only a limited number of C-units available, this leads to a reduction of h which more than compensates for any advantages which might acrue.) The other possible source of correlation comes from the joint activity of the A-units a_i and a_j themselves. The probability that $a_i^* = a_j^*$ will be designated q_{ij}. Thus if a_i and a_j are always either both on or both off (as might occur if they both respond to some stimulus S_x and to no other stimuli) q_{ij} will be equal to 1. If they are always in opposite states, $q_{ij} = 0$. If they are each activated independently, with a probability $Q_a = 0.5$, then $q_{ij} = 0.5$. For any other value of Q_a, however, q_{ij} must be either greater than or less than 0.5. For notational convenience in the following discussion, let $\boxed{q_{ij} = q.}$ — $\rho(a_i^* = a_j^*)$

We will first compute ρ_{ij} for the symmetric case. By definition,

$$\rho_{ij} = \frac{\text{cov}\ (u_i, u_j)}{\sigma(u_i)\ \sigma(u_j)}. \tag{17}$$

The value of $\sigma(u_i)$ has already been obtained, but we must still compute the covariance of the signals. For this we have

$$\begin{aligned}\text{cov}(u_i, u_j) &= \frac{1}{t}\sum_{k=1}^{t} (u_i(k) - \underset{k}{E}(u_i), (u_j(k) - \underset{k}{E}(u_j)) \\ &= \frac{1}{t}\sum_{k=1}^{t} [\tilde{u}_i(k) + u_{ij}(k) - E(\tilde{u}_i) - E(u_{ij})][\tilde{u}_i(k) \\ &\qquad + u_{ji}(k) - E(\tilde{u}_j) - E(u_{ji})]\end{aligned}$$

where $u_i(k)$ = signal to a_i from the kth C-state

$\tilde{u}_i(k)$ = signal to a_i from the set of C-units connected to a_i and not to a_j

$\tilde{u}_j(k)$ = signal to a_j from the set of C-units connected to a_j and not to a_i

$u_{ij}(k)$ = signal to a_i from the set of C-units connected to both a_i and a_j

$u_{ji}(k)$ = signal to a_j from the set of C-units connected to both a_i and a_j.

It can easily be shown that the terms coming from the "unique" connections to a_i and to a_j will cancel out of the above expression, only the "common" set of connections contributing to the covariance. Thus we obtain (if $E(\tilde{u}_i \cdot \tilde{u}_j) = E(\tilde{u}_i)E(\tilde{u}_j)$, $E(\tilde{u}_i \cdot u_{ji}) = E(\tilde{u}_i)E(u_{ji})$.)

$$\text{cov}\ (u_i, u_j) = \frac{1}{t}\sum_{k=1}^{t} [u_{ij}(k) - E(u_{ij})][u_{ji}(k) - E(u_{ji})] \tag{18}$$

Now let $\Delta_i^\nu(k)$ = contribution to $[u_{ij}(k) - E(u_{ij})]$ from the νth stimulus

$\Delta_j^\nu(k)$ = contribution to $[u_{ji}(k) - E(u_{ji})]$ from the νth stimulus.

Note that $|\Delta_i^\nu| = |\Delta_j^\nu|$ since the same set of C-units is involved for both a_i and a_j. Also, note that the sign of Δ_i^ν always agrees with the sign of the total signal increment for the νth stimulus.* For convenience, we can let

* Actually, it is easy to show that $E(u_{ij}) = 0$, so that Δ_i^ν is actually equal to the signal increment.

sgn Δ_i^ν = sgn Δ_j^ν for the first qt stimuli, and let the signs be opposite for the remaining stimuli in the sequence. Thus,

$$u_{ij}(k) - E(u_{ij}) = \sum_{\nu=1}^{qt} \Delta_i^\nu(k) + \sum_{\nu=qt+1}^{t} \Delta_i^\nu(k)$$

$$u_{ji}(k) - E(u_{ji}) = \sum_{\nu=1}^{qt} \Delta_i^\nu(k) - \sum_{\nu=qt+1}^{t} \Delta_i^\nu(k)$$

and

$$\begin{aligned}\text{cov}\ (u_i, u_j) &= \frac{1}{t}\sum_{k=1}^{t}\left[\left(\sum_{\nu=1}^{qt} \Delta_i^\nu(k)\right)^2 - \left(\sum_{\nu=qt+1}^{t} \Delta_i^\nu(k)\right)^2\right] \\ &= \frac{1}{t}\sum_{k=1}^{t}\left(\sum_{\nu=1}^{qt} \Delta_i^\nu(k)\right)^2 - \frac{1}{t}\sum_{k=1}^{t}\left(\sum_{\nu=qt+1}^{t} \Delta_i^\nu(k)\right)^2\end{aligned}$$

which can be seen to be a difference between two variances. The first term represents the contribution to the variance $\sigma^2(u_{ij})$ due to the first qt stimuli, and the second term represents the contribution to the variance due to the remaining $(1 - q)t$ stimuli. Since each stimulus in the sequence contributes an equal increment to the variance of the signal, this becomes

$$\begin{aligned}\text{cov}\ (u_i, u_j) &= q\sigma^2(u_{ij}) - (1 - q)\ \sigma^2(u_{ij}) \\ &= (2q - 1)\ \sigma^2(u_{ij}).\end{aligned}$$

But since the intersection of the sets of C-units connected to a_i and a_j is a fraction of the set connected to either A-unit alone (for $i \neq j$), it can readily be seen that $\sigma^2(u_{ij}) = M\sigma^2(u_i)$. Thus, substituting in Eq. (17), we obtain (for $i \neq j$)

$$\rho_{ij} = \frac{(2q - 1)\ M\sigma^2(u_i)}{\sigma(u_i)\ \sigma(u_j)} = (2q_{ij} - 1)\ M \quad \text{(Symmetric case)} \tag{19a}$$

$\sigma(u_i) = \sigma($

For the asymmetric case, Eq. (17) and (18) still apply. $\Delta_i^\nu(k)$ and $\Delta_j^\nu(k)$ are defined as before. Note, however, that in this case the sign of these increments is always positive, of magnitude Δu whenever a reinforcement occurs, and of magnitude 0 when the A-unit in question is inactive. Suppose both a_i and a_j are active for ${}_{Q_a}t$ stimuli. Then we have

$$u_{ij}(k) - E(u_{ij}) = \hat{Q}_a t\ \Delta u = u_{ji}(k) - E(u_{ji}).$$

Consequently, Eq. (18) becomes

$$\text{cov}\ (u_i, u_j) = \frac{1}{t}\sum_{k=1}^{t} [u_{ij}(k) - E(u_{ij})]^2 = \sigma^2(u_{ij})$$

and we obtain (for $i \neq j$)

$$\rho_{ij} = \frac{\sigma^2(u_{ij})}{\sigma(u_i)\,\sigma(u_j)} = M \qquad \text{(Asymmetric case)} \tag{19b}$$

Note that in the symmetric model, it is theoretically possible to guarantee a zero correlation by guaranteeing that $q_{ij} = 0.5$ for $i \neq j$. This would be true if the environment consists of random stimuli, and Q_a is kept at 0.5 by the Θ-servo. These conditions, however, are quite implausible biologically, and the high value of ϱ_a would be far from optimum in most discrimination experiments. In the asymmetric model, on the other hand, ρ_{ij} is entirely independent of q and consequently of Q_a, depending only on M. This corresponds to the worst possible case of the symmetric model, in which $2\bar{q} - 1 = 1$. These relatively high correlations are compensated' however, by the appearance of $\hat{Q}_a$ in the expression for h [Eq. (16)]. Here it is clear that by keeping Q_a small, the probability of activating any given A-unit correctly becomes correspondingly large, which tends to offset the effects of the increased correlation between signals, as we shall see.

At this point, we have established all of the necessary prerequisites for the analysis of $P(n^+, n^-)$. This probability (that a particular set of n^+ proper units and n^- improper units and no others are activated) can be rephrased as follows: Given N_a normally distributed random variables with unit variance and mean 0, and a matrix R of correlation coefficients ρ_{ij}, we require the probability that the first n^+ variables are $< h$, the next n^- variables are $< -h$, the next $N^+ - n^+$ are $< -h$, and the remaining $N^- - n^-$ variables are all $< h$, where the quantity h is defined in Eq. (16) and ρ_{ij} is defined by Eq. (19), for $i \neq j$. For $i = j$, ρ_{ij} is obviously equal to 1. The method which was finally obtained for achieving a tractable solution to this problem was suggested by Milton Sobel, and has been described by Curnow and Dunnet in Ref. 10. The method is applicable to any case in which the correlation matrix, R, has the structure $\rho_{ij} = \alpha_i\alpha_j$ for $i \neq j$, where $-1 \leq \alpha_i \leq +1$. This condition is clearly satisfied for the asymmetric model, where $\alpha_i = \sqrt{M}$. It is also satisfied for the symmetric case, if we assume q_{ij} to be equal for all pairs i and j $(i \neq j)$.

For the above conditions the following analysis is applicable.* Let $Z_1, Z_2, \ldots, Z_n$ be n standardized normal variables with correlation coefficients ρ_{ij} satisfying the above constraint. Then the variables Z_i can be generated from $n + 1$ independent normal variables $(X_1, X_2, \ldots, X_n, Y)$ by substituting

$$\overset{X_i}{\cancel{Z}_i} = \sqrt{1 - \alpha_i^2}\,\overset{Z_i}{\cancel{X}_i} + \alpha_i Y. \tag{20}$$

* This treatment follows that of Curnow and Dunnet. It is also possible to reduce the multivariate normal distribution, in this case, to a sum of products of Hermite polynomials, but the resulting equation becomes quite unmanageable for large n.

Go to p46

original training of the perceptron permits) due to the growth of h. As the length of the recorded sequence, t, becomes large, on the other hand, h diminishes and the probability of a correct response approaches 0.5.

Of greater interest is the asymptotic behavior of the amount of information stored in the C-system as the recorded sequence grows in length. To estimate this, assume that the stimulus sequence activates random A-states, with $Q_a = .5$, and M small, so that $\rho_{ij} \approx 0$ for both the symmetric and asymmetric models. Then we have the following:

Information content of original stimulus $= N_a$ bits.
Information in S-sequence of t stimuli $= tN_a$ bits.
Information content of reconstituted A-state $= x$ bits.

$$x = H(S) - H_n(S)$$

where $H(S)$ = entropy of stimulus representation $= N_a$ bits
$H_n(S)$ = equivocation (measure of uncertainty in the stimulus, given the reconstituted A-state)

Information stored in C-network $= tx$ bits.

Minimum information stored per C-A connection $= H_c = tx$/total number of connections $= tx/N_aMN_c$.

Lower bound for number of distinguishable weight levels required per connection (with optimal coding) to achieve the obtained storage capacity $= 2^{H_c}$.

The only quantity in the above equations for which we still lack an explicit expression is $H_n(S)$. But this is given by the expression

$$H_n(S) = -\sum_i P(S_i|S)\log P(S_i|S) \equiv -\sum_i P_i(S)\log P_i(S) \tag{29}$$

where $P(S_i|S)$ = probability of obtaining the A-state for stimulus S_i when the correct state is S. This is symbolized by $P_i(S)$.

Therefore the information content of the reconstituted (remembered) A-state is

$$x = N_a + \sum_i P_i(S)\log P_i(S).$$

After Shannon,[54] for the above case, if $P = \Phi(h)$ = probability that the activity state of a given A-unit is correctly reconstituted, the equivocation is given by

$$H_n(S) = -N_a[P\log_2 P + (1 - P)\log_2(1 - P)] \tag{30}$$

and x is given by

$$x = N_a[1 + P\log_2 P + (1 - P)\log_2(1 - P)] \tag{31}$$

Thus, substituting back in the previous expressions, we obtain the following formula for the minimum density of information stored per C-connection (for random stimuli):

$$H_c = \frac{t[1 + P\log_2 P + (1-P)\log_2(1-P)]}{MN_c}$$

$$= \frac{t}{MN_c}\left[1 + \frac{1}{\ln(2)}(P\ln P + (1-P)\ln(1-P))\right] \qquad (32)$$

We are interested in determining the limit of this quantity as $t \to \infty$. Call this limit H_∞. Note that as $t \to \infty$, $P = \Phi(h) \to (0.5 + \phi'(0)\cdot h) = \left(0.5 + \frac{h}{\sqrt{2\pi}}\right)$. Let $C = MN_c/t$. Then $h = k\sqrt{C}$, where, for the symmetric model, $k = \sqrt{1 - Q_c}$ and for the asymmetric model, $k = \sqrt{\frac{1 - Q_c}{4Q_a}}$. As $t \to \infty$, C goes to zero. Let $g = k/\sqrt{2\pi}$. Thus, substituting in Eq. (32), we obtain:

$$H_\infty = \operatorname*{Lim}_{C\to 0}\frac{1}{C}\left\{\frac{1}{\ln 2}[(0.5 + g\sqrt{C})\ln(0.5 + g\sqrt{C})\right.$$

$$\left. + (0.5 - g\sqrt{C})\ln(0.5 - g\sqrt{C})]\right\}$$

$$= \frac{1}{C} + \frac{1}{2C\ln 2}\ln(0.5 + g\sqrt{C}) + \frac{g}{\sqrt{C}\ln 2}\ln(0.5 + g\sqrt{C})$$

$$+ \frac{1}{2C\ln 2}\ln(0.5 - g\sqrt{C}) - \frac{g}{\sqrt{C}\ln 2}\ln(0.5 - g\sqrt{C})$$

$$= \frac{1}{C} + \frac{1}{2C\ln 2}\left\{\ln(0.5) + 2\left[\frac{g\sqrt{C}}{1+g\sqrt{C}} + \frac{1}{3}\left(\frac{g\sqrt{C}}{1+g\sqrt{C}}\right)^3 + \cdots\right]\right\}$$

$$+ \frac{g}{\sqrt{C}\ln 2}\left\{\ln(0.5) + 2\left[\frac{g\sqrt{C}}{1+g\sqrt{C}} + \frac{1}{3}\left(\frac{g\sqrt{C}}{1+g\sqrt{C}}\right)^3 + \cdots\right]\right\}$$

$$+ \frac{1}{2C\ln 2}\left\{\ln(0.5) - 2\left[\frac{g\sqrt{C}}{1-g\sqrt{C}} + \frac{1}{3}\left(\frac{g\sqrt{C}}{1-g\sqrt{C}}\right)^3 + \cdots\right]\right\}$$

$$- \frac{g}{\sqrt{C}\ln 2}\left\{\ln(0.5) - 2\left[\frac{g\sqrt{C}}{1-g\sqrt{C}} + \frac{1}{3}\left(\frac{g\sqrt{C}}{1-g\sqrt{C}}\right)^3 + \cdots\right]\right\}$$

Since the higher power terms in the series become negligible as C goes to zero, this reduces to the form

$$H_\infty = \operatorname*{Lim}_{C\to 0}\left\{\frac{g}{\sqrt{C}\ln 2}\left(\frac{1}{1+g\sqrt{C}}\right) + \frac{2g^2}{\ln 2}\left(\frac{1}{1+g\sqrt{C}}\right)\right.$$

$$\left. - \frac{g}{\sqrt{C}\ln 2}\left(\frac{1}{1-g\sqrt{C}}\right) + \frac{2g^2}{\ln 2}\left(\frac{1}{1-g\sqrt{C}}\right) + \epsilon\right\}$$

$$= \operatorname*{Lim}_{C\to 0}\left\{\frac{1}{1-g^2C}\left[\frac{2g^2}{\ln 2}\right] + \epsilon\right\}$$

$$= \frac{2g^2}{\ln 2}.$$

Thus, substituting for g, we have

$$H_\infty = \frac{1 - Q_c}{\pi\cdot\ln 2} \qquad \text{(Symmetric case)} \qquad (33a)$$

$$H_\infty = \frac{1 - Q_c}{4Q_a\cdot\pi\cdot\ln 2} \qquad \text{(Asymmetric case)} \qquad (33b)$$

Note that this represents quite a low density of information in the network; for Q_c close to zero and Q_a close to 0.25, H_∞ for both models is 0.45922 bits per connection. Note also that with $Q_a < 0.25$ the asymmetric model is capable of storing more information than the symmetric model. Thus we see that as t becomes large, the saturation of the memory is represented by an asymptotic approach to a limiting information density. This fixed amount of stored information as the number of stored stimuli increases is, of course, a direct reflection of the diminishing probability of correct recall.

In the above case it has been assumed that the original association state actually contains N_a bits of information, which will be true only if the stimulus sequence is sufficiently heterogeneous so that each A-unit may be active independently. It was further assumed that the A-units were reactivated independently by the C-system, with probability P of being correct. If the correlation coefficients ρ_{ij} are not equal to zero, this assumption is no longer accurate. In the extreme case, where only two stimuli can occur in the environment, the information content of the A-state is only 1 bit instead of N_a bits, and with perfect correlation of signals from the C-system to the A-system the N_a A-units act, in effect, as a single A-unit. Thus the analysis takes the same form as the above case, but with N_a reduced to 1. This yields limits which are only $1/N_a$ times as large as those for the heterogeneous environment. This again bears out our conclusion that memory of a diverse environment will be better than memory of a repetitious environment.

Before leaving this topic, one final exercise may prove to be illuminating. If we assume that the human brain functions in the manner of our asymmetric memory model, and that there are 10^9 neurons (or about 10 percent of the brain's population) functioning as C-units, each having 1,000 connections to A-units, and if we set $Q_c = Q_a = 0.01$, then Eq. (33b) gives us one more estimate of the information capacity of the brain. Specifically, this would predict that in its saturated condition, the brain would be capable of storing approximately 1.2×10^{13} bits of information, from a sufficiently heterogeneous environment. This fits comfortably in between the two extremes estimated by Miller and von Foerster, which were mentioned in the Introduction.

NUMERICAL RESULTS

Equation (26) has been integrated numerically for a number of cases of interest, using an IBM 7090 computer.* This has yielded estimates for the probability of a correct discriminatory response to a remembered stimulus, $P(R_x)$, for values of t ranging from 10^3 through 10^{11}, and for values of N_a and N_c ranging from 10^3 through 10^9. It was assumed that preliminary training on the discrimination of the test stimulus, S_x, was perfect, i.e., $E(u_x)/\sigma(u_x) = \infty$. MN_c (the number of input connections to each A-unit) was taken as 1,000 in all cases. This would make the A-units roughly comparable to large pyramidal cells in the cerebral cortex. Q_c was assumed to be negligibly small. N_a was assumed to be equal to N_c.

The main calculations completed to date are for the symmetric model, but preliminary calculations of performance for the asymmetric model show

*The writer is indebted to Robert Tuttle for his assistance in programming this problem, and to the Atomic Energy Commission for making its facilities at the Courant Institute available for the computation.

Go to p 50.

Under these conditions, it follows that the Z_i are normally distributed with zero means, unit variances, and correlation coefficients $\rho_{ij} = \alpha_i\alpha_j$ $(i \neq j)$.

Let $f(Z_1, Z_2, \ldots, Z_n)$ represent the joint frequency function of the variables Z_i. The cumulative distribution function is then defined by:

$$F_n(h_i) = P\{Z_i < h_i, \text{ for all } i\}$$
$$= \int_{-\infty}^{h_1} \int_{-\infty}^{h_2} \cdots \int_{-\infty}^{h_n} f(Z_1, Z_2, \ldots, Z_n)\, dZ_1, dZ_2, \ldots, dZ_n$$

We can clearly multiply this expression by $\int_{-\infty}^{\infty} \phi'(Y)\,dY = 1$ without changing its value, (where ϕ' is the normal density function). This yields

$$F_n(h_i) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{h_1} \int_{-\infty}^{h_2} \cdots \int_{-\infty}^{h_n} f(Z_1, Z_2, \ldots, Z_n)\, \phi'(Y)\, dZ_1, dZ_2, \ldots, dZ_n \right] dY.$$

Now substitute $Z_i = \dfrac{X_i - \alpha_i Y}{\sqrt{1 - \alpha_i^2}}$ for all Z_i.

Since the X_i are mutually independent, the resulting integral can be factored, and we obtain

$$F_n(h_i) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} \left[\int_{-\infty}^{h_i} \phi'\left(\frac{X_i - \alpha_i Y}{\sqrt{1 - \alpha_i^2}} \right) dX_i \right] \phi'(Y)\, dY$$
$$= \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Phi\left(\frac{h_i - \alpha_i Y}{\sqrt{1 - \alpha_i^2}} \right) \phi'(Y)\, dY \qquad (21)$$

For the particular case with which we are concerned, $\alpha_i = \sqrt{\rho}$ where $\rho =$ the value of ρ_{ij} for $i \neq j$, and the h_i are all $\pm h$, for the appropriate sets of variables. This yields

$$P(n^+, n^-) = \int_{-\infty}^{\infty} \Phi^{n^+ + N^- - n^-}\left(\frac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}} \right) \left[1 - \Phi\left(\frac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}} \right) \right]^{n^- + N^+ - n^+} \phi'(x)\, dx \qquad (22)$$

The expression $\begin{pmatrix} N^+ \\ n^+ \end{pmatrix} \begin{pmatrix} N^- \\ n^- \end{pmatrix} P(n^+, n^-)$ which appears in Eq. (7), representing the probability of any arbitrary set of n^+ proper units and n^- improper units being active, can then be written in the form

$$G(n^+, n^-) = \int_{-\infty}^{\infty} \binom{N^+}{n^+} F^{n^+} (1 - F)^{N^+ - n^+} \binom{N^-}{n^-} F^{N^- - n^-} (1 - F)^{n^-} \phi'(x)\, dx \tag{23}$$

where $F = \Phi\left(\dfrac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}}\right)$. This integrand evidently includes the product of two binomial probability functions. For large N^+ and N^- (such as we will always be dealing with) it is possible to approximate this by the product of two Gaussian probabilities as follows:

$$G(n^+, n^-) \approx \int_{-\infty}^{\infty} \phi'\left(\frac{n^+ - N^+F}{\sqrt{N^+F(1 - F)}}\right) \phi'\left(\frac{n^- - N^-(1 - F)}{\sqrt{N^-F(1 - F)}}\right) \phi'(x)\, dx \tag{24}$$

and thus, substituting in Eq. (7), we obtain

$$P(R_x) = \sum_{n^+=0}^{N^+} \sum_{n^-=0}^{N^-} G(n^+, n^-)\, \Phi[Z(n^+, n^-)]$$
$$= \int_{-\infty}^{\infty} \sum_{n^+=0}^{N^+} \sum_{n^-=0}^{N^-} \phi'\left(\frac{n^+ - N^+F}{\sqrt{N^+F(1 - F)}}\right) \phi'\left(\frac{n^- - N^-(1 - F)}{\sqrt{N^-F(1 - F)}}\right) \phi'(x) \cdot \Phi[Z(n^+, n^-)]\, dx \tag{25}$$

This can be simplified further, for large N_a, by replacing the sums by integrals, with appropriate limits, which yields

$$P(R_x) = \int_{x=-\infty}^{\infty} \int_{y=-\frac{\sqrt{N^+}}{L}}^{\sqrt{N^+}L} \int_{z=-\frac{\sqrt{N^-}}{L}}^{\sqrt{N^-}L} \phi'(x)\, \phi'(y)\, \phi'(z)\, \Phi[\psi(x, y, z)]\, dx\, dy\, dz \tag{26}$$

where

$$L = \sqrt{\frac{1 - F}{F}}$$

$$\psi(x, y, z) = Z\left[\sqrt{N^+F(1 - F)}\left(y + \frac{\sqrt{N^+}}{L}\right), \sqrt{N^-F(1 - F)}\left(z + \frac{\sqrt{N^-}}{L}\right)\right]$$
$$= Z(n^+, n^-)$$

This last form is the one which has actually been used for numerical computation.

Before going on to a consideration of numerical results, it is interesting to examine the asymptomatic behavior of $P(R_x)$, as well as several information-theoretic conclusions which follow directly from these equations.

It can easily be seen from Eq. (26) that as N_a gets large the integrand will be effectively equal to zero everywhere except in the neighborhood of the expected values of x, y, and z. Equivalently, in the form shown in Eq. (25), this means that the only terms in the summation which carry any weight are those in the neighborhood of the expected values of n^+ and n^-. Consequently, in the limit, as $N_a \to \infty$, $Z(n^+, n^-)$ can be replaced by $Z[E(n^+), E(n^-)]$, which no longer depends on n^+ and n^-. Thus the probability terms derived from the binomials in Eq. (23) can now be summed to unity, yielding

$$P(R_x) \xrightarrow[Na \to \infty]{} \int_{-\infty}^{\infty} 1 \cdot \phi'(x) \cdot \Phi[Z(En^+\ En^-)]\, dx$$
$$= \int_{-\infty}^{\infty} \Phi\left[\frac{\dfrac{FN^+}{Q_a} - \dfrac{(1 - F)\,N^-}{1 - Q_a}}{\sqrt{[FN^+ + (1 - F)\,N^-]\dfrac{N_a}{Q_a}}} \cdot \frac{E(u_x)}{\sigma(u_x)}\right] \phi'(x)\, dx$$

Substituting $N^+ = Q_aN_a$ and $N^- = (1 - Q_a)\,N_a$, this yields

$$P(R_x) \to \int_{-\infty}^{\infty} \Phi\left[\ (2F - 1) \cdot \frac{E(u_x)}{\sigma\,(u_x)}\right] \phi'(x)\, dx \tag{27}$$

where $$F = \Phi\left(\frac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}}\right)$$

Now, taking a second limit as $E(u_x)/\sigma(u_x) \to \infty$ (corresponding to a perfectly learned response to S_x), note that the argument of Φ will be $+\infty$ with $\dfrac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}} > 0$, and $-\infty$ otherwise.

But $$\frac{x\sqrt{\rho} + h}{\sqrt{1 - \rho}} = 0 \text{ for } x = -h/\sqrt{\rho}.$$

Therefore, in Eq. (27), $\phi'(x)$ is weighted by 1 for $x > -h/\sqrt{\rho}$, and by 0 for $x < -h/\sqrt{\rho}$. Consequently, we obtain the limiting performance for a perfectly trained perceptron with infinite N_a,

$$\lim_{\substack{Na\to\infty \\ \frac{Eu_x}{\sigma u_x}\to\infty}} P(R_x) = \int_{-h/\sqrt{\rho}}^{\infty} \phi'(x)\,dx = \Phi\left(\frac{h}{\sqrt{\rho}}\right)$$

This asymptotic formula suggests that the limiting performance may be quite poor if ρ is close to unity. While this is true in principle, note that in practice, large values of N_a will almost certainly be accompanied by large values of N_c. As N_c goes up, the value of h increases, so that the limit, Eq. (28), can become arbitrarily close to perfect performance regardless of ρ. Nonetheless, this formula suggests an explanation for the poor performance which is obtained in the recall of stereotyped sequences of symbols with low diversity, such as strings of binary digits. Suppose, for example, that the environment consists of only two stimuli, a square in a particular location, and a circle in a particular location, and that the perceptron is asked to record a long sequence of these squares and circles in random order. Under these conditions, for the symmetric model, q_{ij} will tend to be either 1 or 0, since there are only two meaningful A-unit sets, and any pair of A-units will either be in the same set or in opposite sets. This yields a maximum value for ρ, and correspondingly poor performance. In the asymmetric model, it may appear that this effect will not occur, but this is deceptive; ϱ_a now takes the place of q as the source of difficulty. For the highly correlated environment, those A-units which respond at all are likely to respond a large fraction of the time, resulting in large values of ϱ_a, hence low values of h and correspondingly poor performance. Thus performance will always be best in a very heterogeneous environment, with a great diversity of stimuli, and it will be poorest in a stereotyped environment, containing only a few patterns or symbols which may occur.

By substituting h and ρ in Eq. (28), it is easy to find the conditions for which the limiting performance of the symmetric model and asymmetric model are identical. This will occur when Q_a for the asymmetric model is equal to $(2q - 1)/4$ for the symmetric model. For example, a symmetric model with $q = 0.52$ (which requires Q_a close to 0.5) would have the same limiting performance as an asymmetric model with $\varrho_a = 0.01$. For sufficiently diverse stimuli the value of ϱ_a can be taken equal to Q_a, the probability that an A-unit responds to any given stimulus.

So far, we have considered only the limits for large values of N_a. As N_c becomes large, other parameters remaining constant, it is clear that the limiting performance will always become perfect (or as nearly so as the

Go to p42.

that the two are nearly identical for the condition ϱ_a (in asymmetric model) $= (2q - 1)/4$. In the limit, as was shown in the last section, the performances of the two models are identical when this condition is satisfied. In the four cases completed, q was taken as 0.50, 0.55, 0.60, and 0.75, respectively (corresponding to ϱ_a of 0, 0.025, 0.05, and 0.125 for the asymmetric case). The results are shown in Table 1. The probabilities should be good to five places, although there is a possibility of a slight error in the fifth place.

These results seem striking enough so that they can really speak for themselves. It appears that if q can be kept sufficiently close to 0.5 (or ϱ_a, in the asymmetric model, kept small enough) then the probability of correctly identifying a well-learned binary characteristic in retrospect, after having seen and recorded 10^{11} different stimuli, is about 0.994, for a network of 10^9 C-units and an equal number of A-units. This result, in itself, was sufficiently unexpected so that the writer was obliged to reconsider entirely his previous assessment of the Penfield hypothesis of complete storage of experience, and related hypotheses proposed by other investigators. Such a level of performance would permit an individual to record fifteen independent events per second (about the flicker-fusion rate) for over two centuries before the probability of correct binary identification fell below 0.99.

Table 1. Values of $P(R_x)$ for Symmetric Model

$MN_c = 1{,}000$, $E(u_x)/\sigma(u_x) = \infty$, $Q_c = 0$

Case 1: $q = 0.50$

	t = length of stored sequence								
N_a, N_c	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	10^{11}
10^3	1.0	1.0	0.99425	0.78755	0.59960	0.53179	0.51006	0.50318	0.50101
10^4	1.0	1.0	1.0	0.99419	0.78753	0.59959	0.53180	0.51006	0.50318
10^5	1.0	1.0	1.0	1.0	0.99418	0.78751	0.59960	0.53180	0.51007
10^6	1.0	1.0	1.0	1.0	1.0	0.99418	0.78754	0.59961	0.53181
10^7	1.0	1.0	1.0	1.0	1.0	1.0	0.99418	0.78754	0.59964
10^8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99419	0.78762
10^9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99420

Case 2: q = 0.55

	t = length of stored sequence								
N_a, N_c	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	10^{11}
10^3	0.99915	0.83962	0.62322	0.53953	0.51252	0.50395	0.50124	0.50039	0.50012
10^4	1.0	0.99915	0.83947	0.62315	0.53949	0.51251	0.50396	0.50125	0.50038
10^5	1.0	1.0	0.99915	0.83944	0.62315	0.53951	0.51251	0.50395	0.50125
10^6	1.0	1.0	1.0	0.99915	0.83946	0.62314	0.53952	0.51251	0.50396
10^7	1.0	1.0	1.0	1.0	0.99915	0.83944	0.62316	0.53952	0.51251
10^8	1.0	1.0	1.0	1.0	1.0	0.99915	0.83947	0.62316	0.53953
10^9	1.0	1.0	1.0	1.0	1.0	1.0	0.99915	0.83946	0.62319

Case 3: q = 0.60

	t = length of stored sequence								
N_a, N_c	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	10^{11}
10^3	0.98664	0.75956	0.58794	0.52778	0.50863	0.50272	0.50086	0.50027	0.50008
10^4	1.0	0.98703	0.75941	0.58801	0.52792	0.50876	0.50275	0.50086	0.50026
10^5	1.0	1.0	0.98703	0.75937	0.58802	0.52792	0.50877	0.50275	0.50086
10^6	1.0	1.0	1.0	0.98702	0.75939	0.58801	0.52793	0.50877	0.50275
10^7	1.0	1.0	1.0	1.0	0.98703	0.75937	0.58802	0.52793	0.50877
10^8	1.0	1.0	1.0	1.0	1.0	0.98702	0.75940	0.58802	0.52794
10^9	1.0	1.0	1.0	1.0	1.0	1.0	0.98703	0.75939	0.58805

Case 4: q = 0.75

	t = length of stored sequence								
N_a, N_c	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}	10^{11}
10^3	0.89240	0.66569	0.55871	0.51302	0.50336	0.50039	0.50000	0.50000	0.50000
10^4	0.99999	0.92084	0.67066	0.55802	0.51590	0.50481	0.50151	0.50048	0.50015
10^5	1.0	0.99999	0.92085	0.67075	0.55789	0.51602	0.40486	0.50153	0.50048
10^6	1.0	1.0	0.99999	0.92084	0.67077	0.55787	0.51603	0.50487	0.50153
10^7	1.0	1.0	1.0	0.99999	0.92085	0.67076	0.55788	0.51603	0.50487
10^8	1.0	1.0	1.0	1.0	0.99999	0.92084	0.67078	0.55788	0.51604
10^9	1.0	1.0	1.0	1.0	1.0	0.99999	0.92086	0.67077	0.55790

These probabilities are obviously attenuated considerably for more realistic values of q (or ϱ_a). Nevertheless, the probabilities shown for Case 2, where the parameters should be readily achievable, are still most impressive.

Note that as N_a, N_c, and t all get large simultaneously, $P(R_x)$ approaches a constant for a given ratio of t/N_a. This means that in a large system, to maintain a fixed level of performance, it is necessary to add a fixed number of connections for each new stimulus which the memory must accommodate. This result can also be seen from an examination of the equations in the last section. For example, if a probability of 0.999 is required, with $q = 0.55$ (or $\varrho_a = 0.025$), then one additional C-unit with 1,000 connections must be added to the network for each additional stimulus.

Such probabilities as 0.999, however, are probably too high to expect of a biological system. A complex memory is rarely defined by a single binary characteristic, occurring over a ⅟₁₅-sec period. Redundant information is almost always present, as well as the possibility of many successive "looks" at the same event, so that a relatively low probability, say about 0.6, would probably be sufficient to match human performance. Under these conditions, more than 10 new stimuli could be stored for each additional C-unit.

SELECTIVE RECALL AND PROGRAM-LEARNING PERCEPTRONS

In the memory model as it stands up to this point, the recall of a sequence occurs as a result of setting the system back to its initial state and cutting off all sensory inputs. It was suggested that the resetting might be accomplished by training the perceptron to activate an R-unit which, by virtue of strong connections to the C-system, could override its present activity state and force it into some particular starting state. Actually there might be a considerable number of such R-units, each setting up a different, independent, initial state, so that any one of a number of named sequences might be evoked on command. Thus, if the response R_h was associated to the word "Hamlet," and this was followed by a recitation of *Hamlet*, the repetition of the name "Hamlet" would tend, thereafter, to cause the perceptron to review the subsequent recitation. On the other hand, the word "Faust," associated to a different R-unit R_f, could evoke a sequence in which the dialogue of *Faust* was recorded.

Such a method of recall, while entertaining to contemplate, has several features which make it decidedly unrealistic. Perhaps the most important of these is illustrated by the fact that the word "Hamlet" occurs repeatedly within the text of the play; nonetheless no actor, having memorized the text of *Hamlet*, is likely to be "reset" to the beginning of the play whenever

Hamlet's name is mentioned. This could, it is true, be ameliorated by making the response R_h contingent upon a more complex command, such as the sentence "Begin reciting *Hamlet* from the start of Act I," but this still seems to be a contrived solution.

A second difficulty comes from the fact that recall may be triggered by many events other than a specific command or response which initiated the recalled sequence. The phenomena of free association are too well known to require elaboration here, and an acceptable model should be able to account for them.

A model which seems likely to be able to deal with both of these difficulties in a convincing way is illustrated in Fig. 4. This contains all of the component parts of the perceptron shown in Fig. 2, with several additions. The "set control network" is a set of units which may be triggered by adaptive connections from the *A*-units, provided their threshold is low enough to allow them to respond. Their threshold is itself under the control of one of the *R*-units, which can thus make it easy or difficult to set up a new state in the set-control network. The set-control units themselves are assumed to be long-persisting "On" units, possibly of the same type as are found in the *C*-system. When any of these set-control units becomes active, a strong "on" burst is transmitted to the *C*-system (possibly relayed by means of an intervening layer of short-persistence on-units, so that a continued bombardment of the *C*-system does not occur, even though the set-control network continues to hold its state). This burst of impulses arriving at the *C*-system will tend to force at least some of the *C*-units to a particular initial state determined by the transmitting set.

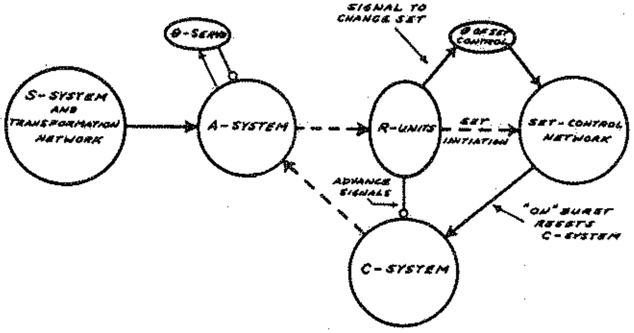

Figure 4

To be explicit, assume that the set-control network is initially silent, with none of its units active. Meanwhile, the *R*-system assumes some state, corresponding to whatever event is taking place in the perceptron's sensory system or association system. At some point, a state of the association system occurs which signals an important event, or the start of a new situation, which might call for an appropriate memory consultation. Such situations, either due to intrinsic or learned connections, are assumed to activate the *R*-unit which lowers the threshold of the set-control network (e.g., by delivering an on-burst of excitatory impulses to all of its neurons). As soon as this takes place, the set-control network is forced to assume some state which depends on the particular set of *R*-units currently active, and which have acquired the strongest connections to it. Let us say this occurs in the auditorium of a theater, at the start of the play *Hamlet*. Then the combination of *R*-units which serve to identify this situation will determine the state of the set-control network, which, in turn, forces the *C*-system to assume a corresponding initial state. The subsequent events are then recorded in the *C*-to-*A* network just as before, as long as the set-control network does not change its state. But, since the activity of the active set-control neurons is assumed to persist during this time, even though their threshold has been restored to its normally high level, the following succession of *R*-unit states is presumed to become associated to the presently active set-control state, by precisely the same sort of memory mechanism which causes the *C*-unit states to form strong connections to active *A*-unit states. Thus a large number of alternative responses occurring during the play (such as the names "Polonius," "Denmark," etc.) become associated to the same set-control state.

As a consequence of this multiple association of *R*-states to the same set-control state, any one of them, at a later time, might start the recall of *Hamlet*, from the beginning, provided the threshold of the set-control network has momentarily been lowered, permitting it to change its state. If the threshold is lowered only momentarily and then maintained at a high level for a long period, there will ensue a detailed and (presumably) accurate repetition of the play. On the other hand, if the threshold is kept at a low level, every time a trigger-event is recognized which has been associated to some other state of the set-control system, this will force an abrupt change in the recalled sequence. Such trigger events occurring in the recalled sequence itself give rise to a phenomenon akin to free association, while those which come from the external environment (while the set-control threshold is held low) give rise to a state of "high suggestibility," or "distractability."

All of this is, of course, quite speculative and unproven at the present time, but it seems likely that a model akin to that outlined here will prove to have the necessary flexibility of associative control to permit either exact recall of a sequence or free-associative recall. It seems likely, in fact, that the learning which must take place in order to repeat a long and complicated sequence correctly has little to do with the recording of experience in the C-to-A network itself, but is rather an indication of the difficulty of learning to set up and maintain the necessary set-control states without interference or distraction. This may be part of the answer to points (7) and (8) in the list which was given in the introduction.

It is rather easy to elaborate, in this heuristic manner, on possible sophistications of the basic memory model. We will indulge in only one other such speculation, however, which seems to hold promise of particularly interesting performances in the future. This is the possibility of employing one or more C-systems for the recording of experience, in the above manner, while simultaneously employing another C-system (or systems) to go through a control program, or sequence of instructions. Such a system is shown in Fig. 5.

It has been known for some time that a bias may be introduced into the association system of a perceptron by feedback from an R-unit, which leads to conditional responses to incoming stimuli (cf. Ref. 44, Chaps. 21 and 22). In the system of Fig. 5, the same principle is employed, except that instead of being determined by feedback from a currently active R-unit, the bias-condition of the association system is determined by the state of the $C^{(1)}$ network. The asynchronous operation of the C-networks makes it possible to advance either $C^{(1)}$ or $C^{(2)}$ independently of the other. Thus with the control network in one state, a whole sequence of remembered events might be fed out of $C^{(2)}$ until an event is remembered which the $C^{(1)}$ state is concerned with. For example, with $C^{(1)}$ in one "command state," the perceptron might be required to review the first act of *Hamlet* and signal (on a particular R-unit) for each occurrence of the word "king," while with $C^{(1)}$ in a different state, the perceptron might go through the identical memory-sequence, but this time count the references to soldiers, or indicate every entrance and exit of a character. Such program "instructions" can be recorded in precisely the same way that experiential events are recorded, namely, by forcing the association system to the desired bias state (by means of an external stimulus) and recording this state of the A-system in the $C^{(1)}$ memory. For this purpose, it is convenient to be able to "turn on" or "turn off" the memory-recording process, both for $C^{(1)}$ and for $C^{(2)}$, so that only the desired states will be recorded. This memory control might be carried out by additional R-units, with the special function of turning on or turning off the recording process. Thus, a whole sequence of bias states could be "read in" to the association system and stored in $C^{(1)}$, constituting a sequential program which could later be used to modify the perceptron's responses to data or events stored in $C^{(2)}$.

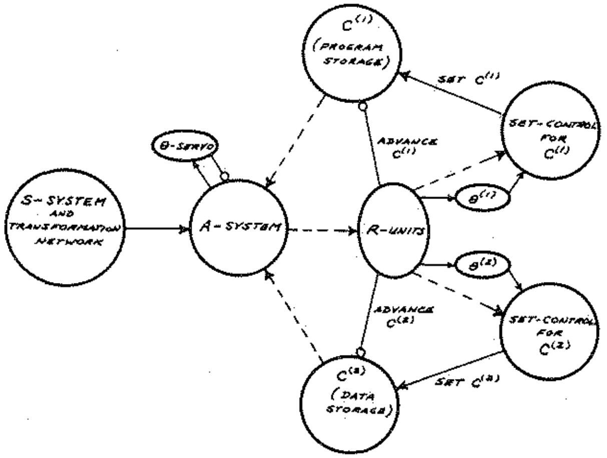

Figure 5

Without further quantitative work, it seems wasteful to speculate at greater length about these possibilities. Nonetheless, it may be worth mentioning that the writer has succeeded in "programming" such a perceptron (on paper) to perform such tasks as forming the logical product of two stored strings of binary digits, and learning an algorithm for counting in binary, which has long been a stumbling block for earlier perceptron models. Perceptual as well as logical operations can be programmed; for example, a perceptron with a movable visual system, or a tactile feeler which it can watch as it moves, can be taught to trace the outline of an object or pattern to determine whether or not it is a closed curve, or it can be taught to search for a particular object in the environment. It seems likely that the most interesting applications of such systems will be in the processing of speech and language, since many of the operations and heuristic methods which were previously applicable only in stored-program digital computers now seem to be almost within reach of our neural networks.

A POSSIBLE BIOCHEMICAL TRACE MECHANISM

In all of the foregoing it has been assumed that the modification of neural connection weights according to a γ-system equation is a tenable postulate. The choice of this particular form of "reinforcement rule" having now been supported, in part, by showing that it leads to psychological behavior of a rather anthropomorphic form, in networks with simple organizational principles, it remains to be seen whether such a choice appears plausible at the neurophysiological level.

During the past year the beginnings of a biochemical model for a gamma-system mechanism, using ordinary enzymological types of reactions, have been constructed. Most of this work will be discussed in more detail in subsequent publications, but a brief outline is presented here in order to indicate the direction of our thinking on the subject.

Since all of our proposed memory mechanisms (alpha system as well as gamma system) involve a change in "synaptic weights," let us begin by reviewing briefly what is known about the mechanism of synaptic excitation and inhibition. The reader who is unacquainted with this subject can find most of the necessary background material in Eccles,[11,12] while selected papers in the collections edited by Florey[18] and Elliott, Page, and Quastel[15] are likely to be helpful in providing missing details.

In a resting neuron, the concentration of potassium is much greater inside the cell than outside, while the concentration of sodium and chloride is much greater outside than inside. These concentration differences, which are maintained by metabolic mechanisms which are only partially understood, result in a Donnan equilibrium across the cell membrane, whereby the outside is normally about 70 to 100 millivolts positive relative to the inside. An excitatory impulse acts by partially depolarizing the cell membrane in the neighborhood of the cell body. If the membrane is sufficiently depolarized, a self-propagating spike impulse is initiated. On the other hand, if the membrane is hyperpolarized the threshold of the cell is effectively raised and excitatory impulses are less effective. It is now generally accepted that excitatory effects at the synapse are mediated by an *excitatory transmitter substance*, such as acetylcholine, which is released by the presynaptic terminal (endbulb) and which binds to a *receptor protein* in the postsynaptic membrane. This induces a change in membrane structure which appears to greatly increase the local permeability of the membrane to all species of ions, thereby causing an electrical "short circuit" which tends to depolarize the cell membrane and initiate a spike discharge. Inhibitory action at synapses is usually assumed to be mediated by an *inhibitory transmitter substance*. It seems increasingly likely that GABA is such an inhibitory transmitter,[36] although there is also accumulating evidence for the belief that some transmitters, such as ACH, may sometimes act as excitatory transmitters and at other times as inhibitory transmitters, depending on the nature of the subsynaptic membrane.[60] In any event, when an inhibitory impulse arrives at a cell, it is established that the membrane becomes hyperpolarized due to a selective increase in the permeability of the membrane to chloride and to potassium, but not to sodium. Recent evidence by Araki, Ito, and Oscarsson,[1] and by Ito, Kostyuk, and Oshima[29] strongly bears out the hypothesis that the inhibitory transmitter substance acts by causing small pores to open in the postsynaptic membrane, large enough to admit the small hydrated potassium and chloride ions, but too small to admit the larger hydrated sodium ions. For an excitatory impulse, on the other hand, it is believed that larger pores are opened which freely admit all of the ions in question. It is readily demonstrated that such a mechanism would, in fact, account for the main empirical facts of both excitatory and inhibitory transmission.

This pore hypothesis, it must be emphasized, is specific to *synaptic* transmission. Outside the region of the synapse, it is likely that rather different mechanisms control the permeability of the membrane, such as those suggested by Shanes.[52,53] These other mechanisms (which cause an increase in permeability of the depolarized membrane) are responsible for the propagation of the nerve impulse over the membrane.

We shall accept the membrane-pore hypothesis as the starting point for developing a hypothetical memory-trace mechanism. Clearly any change which permanently blocks one of the small inhibitory pores at a synapse, or which removes a previous block from a potential excitatory pore, or which widens an inhibitory pore (making it into an excitatory pore) will tend to increase the excitatory effect (or, what is often equivalent, reduce the inhibitory effect) of a synapse.

Because the asymmetric model of the proposed memory mechanism seems to operate under much more plausible biological conditions than the symmetric model (and also because the symmetric model is much more difficult to find a convincing mechanism for) we shall concentrate on defining a possible mode of operation of the asymmetric γ-system. This system requires that the following basic conditions be satisfied:

1. No reinforcement shall occur unless the post-synaptic neuron is active.
2. Assuming the post-synaptic cell is active, then an increment should accrue to the excitatory weight of each active synapse, but not at inactive synapses.
3. The sum of the synaptic weights of all connections to a given neuron must remain constant; if one synapse gains in excitatory weight, at least one other must lose a compensating amount.

Two basic mechanisms have been considered by which these conditions might be realized:

A. Active inhibitory pores to the active cell might be plugged or the active sites of the inhibitory transmitter at these pores might be blocked.

At the same time, an equal number of previously blocked inhibitory pores elsewhere in the same cell would have to be released, to satisfy condition (3).

B. Active excitatory pores in the active cell might be unblocked, or facilitated. At the same time, an equal number of previously "clear" excitatory pores elsewhere in the same cell would have to be blocked.

From the three basic conditions (stated above) which must be satisfied for the γ-system, it is possible to make some inferences about the trace mechanism:

1. The condition that the postsynaptic neuron must be active for a memory change to occur implies either
   (*a*) A critical chemical component or catalyst for the recording process mu stpass through the active cell membrane (during its period of heightened permeability), or else
   (*b*) A critical component or catalyst must be manufactured or released by the postsynaptic cell as part of the metabolic activity which follows excitation.
2. The limitation of the weight-gain to *active* synapses suggests that the mechanism of synaptic action must either unmask or create an active site for the trace mechanism to operate.
3. The conservation rule for the γ-system requires that the trace must be maintained by a substance or structure capable of metabolic normalization for the cell as a whole.

An additional condition, not directly imposed by the γ-system rules, is that the trace mechanism should be subject to disruption by the conditions which are conducive to amnesia, and should be capable of recovery in a manner consistent with empirical data.

Of the various models which have been considered which appear to satisfy these conditions, the following seems to be among the most plausible:

The memory trace depends upon four kinds of molecules which bind successively to the postsynaptic membrane. These are a *transmitter substance*, a *marker substance*, a *recorder substance*, and a *stabilizer substance*. The relationship of these molecules to one another and to the membrane structures is illustrated in a highly schematic form in Fig. 6. The diagram shows (obviously not to scale) a single inhibitory pore in a patch of subsynaptic membrane. Related diagrams can be constructed for excitatory pore mechanisms. Normally, of course, we would expect a great number of such pores to exist at every synaptic junction, so that a change in a single pore, as shown here, constitutes only a slight quantized increment to the weight of a synapse. This figure, which makes use of inhibitory pore blocking as the memory effect, corresponds to one form of mechanism (*A*) above. There are at least two possible synaptic arrangements under which this mechanism might be used:

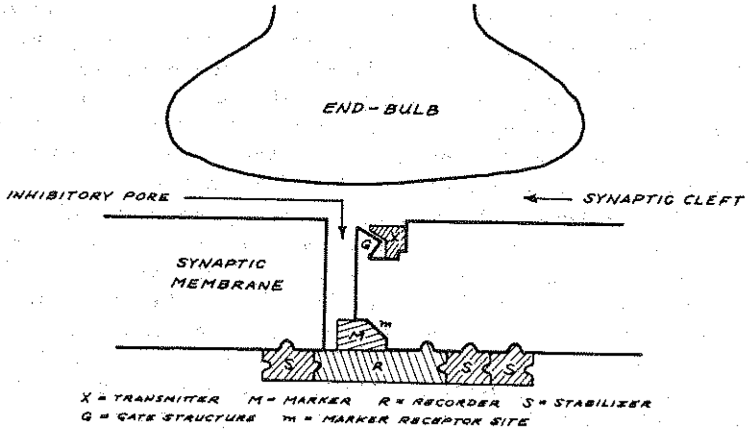
END-BULB
INHIBITORY PORE
SYNAPTIC CLEFT
G
X
SYNAPTIC MEMBRANE
M
m
S
R
S
S
X = TRANSMITTER M = MARKER R = RECORDER S = STABILIZER
G = GATE STRUCTURE m = MARKER RECEPTOR SITE

Figure 6

1. The input connections (from the *C*-network) may include both specific excitatory connections and specific inhibitory connections. Reducing the inhibitory signal component from a given *C*-state to a particular *A*-unit would be equivalent to increasing the excitatory signal component, as prescribed in the γ-system equation [Eq. (1)].
2. The connections from the *C*-units to the *A*-units may be undifferentiated, releasing a single transmitter substance which acts indiscriminately to open both excitatory and inhibitory pores. If a mixture of both types of pores exists at each synapse, then the synapse will be excitatory in effect if the excitatory pores outnumber the inhibitory pores, and it will be inhibitory if the inhibitory pores outnumber the excitatory pores to a sufficient degree. Thus, by blocking inhibitory pores or removing the blocks from them, a single synapse of this type can be transformed from an inhibitory to an excitatory synapse, or vice versa.

The origin, function, and important properties of each of the four postulated substances can now be described.

1. *Transmitter substance*. Released by the presynaptic endbulb. The transmitter has the function of opening the synaptic pore, and thus unmasking an active site for the marker substance, which must enter from outside the cell. Its reaction time is very fast, and it is quickly hydrolyzed by an antitransmitter substance, which restores the pore to its resting condition. In the diagram it is suggested that the transmitter opens the pore by a steric interaction with the receptor protein which causes a change

in the tertiary structure of the pore, causing a "gate" (G) to be displaced from the pore opening. Such a mechanism has also been suggested by Eccles,[14] but the reader should not be misled by the diagrammatic representation into thinking that gross physical deformation is the only possible blocking and unblocking mechanism; changes in the electrical distribution around the pore, possibly by means of charge transfer complexes (cf. Szent-Györgyi[56]) are in many ways a more attractive possibility.

2. *Marker substance.* Extracellular in origin, possibly released by glial cells. Enters open pores and occupies a receptor site (m in the diagram) near the intracellular end of the pore. Its required reaction rate is fast, requiring a high concentration in the neighborhood of the synapse. It must probably be removed by an *antimarker substance* within several seconds (or by a spontaneous decay process) to prevent it from saturating all available pores, and to prevent it from acting as a functional mimetic of the recorder substance, which would prevent the proper metabolic control of the γ-system normalization. The functions of the marker are: (a) to bridge the time gap between the fast transmitter substance reaction and the much slower recorder substance reactions; (b) to satisfy condition (2), above, since, by virtue of its mode of entry, it limits the subsequent process to active synapses; (c) to create an active site for the recorder molecule on the intracellular membrane; (d) although not essential, the marker itself may block the pore prior to the arrival of the recorder molecule, thus forming a temporary memory trace, with a duration of several seconds.

3. *Recorder substance.* Intracellular in origin, with the total concentration in the cell held constant under metabolic control. Free concentration is very low. It is possible to control the concentration of bound recorder by means of a production inhibitor, which is formed on the bound R molecule as a template, and tends to prevent the formation of more recorder substance by the cell. An equilibrium will be established, with the inhibitor acting as a negative feedback mechanism, tending to hold the number of bound recorder molecules (and thus the number of altered synaptic pores) constant for the entire cell. The bound recorder is assumed to be fairly stable (with a time constant of many minutes or hours) in a resting cell, but is dissociated from the membrane relatively easily during or after activity. This might be a direct consequence of the membrane changes during activity, or might be due to the production or admission through the active membrane of an *antirecorder substance.* As soon as recorder is thus removed from some of its occupied sites, during or following a period of activity, the metabolic control mechanism will tend to increase its recorder production, and the new molecules will bind to whatever sites are available, until the number of molecules bound is restored to its normal level. But only those sites marked by a marker molecule (or possibly an

unoccupied bed of stabilizer molecules) are able to bind the recorder. Consequently, the net effect of the transaction will be a shift of recorder from previously occupied synapses (each of which will lose a few molecules) to newly marked sites at the recently active synapses. Thus the functions of the recorder are (a) to maintain the memory trace, either by direct modification of the pore structure, or by protecting the previously bound marker substance from the action of anti-marker, or by interacting with the transmitter substance receptor protein, thus blocking a transmitter site; (b) to satisfy condition (1), since the redistribution can occur only after a period of activity which causes accelerated dissociation of marker from previously occupied sites; (c) to satisfy condition (3), by means of the metabolic conservation mechanism.

4. *Stabilizer substance.* Intracellular in origin. This is assumed to be a large molecule (possibly a protein) which binds to the combination of active sites provided by a recorder molecule and the synaptic membrane, or else by another stabilizer molecule and the synaptic membrane. Thus, at any occupied pore, a "bed" of stabilizer molecules will be built up, which tends to pyramid until all available sites are occupied. As a consequence of this pyramiding effect (particularly if there are more than two active sites on each recorder or stabilizer molecule) old sites, which already have a stabilizer bed established, will tend to preempt the available supply of stabilizer in preference to new sites. This would lead to a heightened stability of the earliest memory traces, and ultimately to an inability to stabilize new traces (as in senility). The free concentration of stabilizer is assumed to be at a very low level, and its reaction rate is very slow. Its functions are to protect the recorder molecules from dissociation, and to provide a "bed" in the neighborhood of the occupied pore, so that if a recorder molecule is dissociated (say as a result of convulsive activity in the nervous system), the prepared bed is more likely to recapture free recorder molecules in the future, thus permitting a gradual recovery from amnesia effects. Note that since the best prepared beds will be the oldest ones, recovery from amnesia should occur in the original temporal order, without regard to the importance of the remembered events, which checks well with the empirical data.

While obviously lacking in any sort of direct experimental confirmation at this time, we see that the above theory does, in fact, satisfy the conditions which were originally required for the asymmetric γ-system, without postulating any radical innovations in biochemistry. It is gratifying to find that such additional phenomena as the amnesia effects and stability of early memory follow directly from this model, even though they were not present in the purely logical form of the γ-system, discussed in the preceding sections. It is also tempting to consider the possibility that the hippocampus may play a role as a source of one of the extracellular components in this process, such as the marker substance, or a catalyst required for the fabrication of new marker substance. If this were the case, then the removal of the hippocampus would lead to precisely the effects observed by Milner and Penfield, which were cited in Item 22 of the list in the Introduction.

Psychogenic amnesias would, of course, involve a different mechanism, presumably an inability to reestablish the cognitive set which initializes the appropriate memory sequence in the system shown in Fig. 4. If such a set could be reestablished (which might occur by free association with any of the responses which were formerly associated to it, even if the normal trigger response had been suppressed) then recovery of the "missing memories" would be immediate and complete, unlike recovery from the physiological amnesias resulting from seizures or convulsions in the above model. The similarity of this effect to actual psychoanalytic observations is most striking.

Some of the Penfield observations on memory sequences induced by temporal lobe stimulation may also find an explanation in terms of Fig. 4. If the temporal-lobe stimulation activates a previous state of the set-control mechanism (which will be likely to occur if the states of the set-control system are mutually exclusive, so that the probability of inducing a meaningless mixture of states is reduced) then as long as the electrical stimulus is maintained, the *C*-system will be forced to recapitulate the corresponding stored sequence, without any possibility of being diverted by free association or changes in the cognitive set. Such an explanation is clearly tenuous at this point, but it serves to support the conclusion that most of the phenomena listed in the introduction, even if they are not clearly predicted by the present model, are at least not inconsistent with it.

Some of these remaining points will be considered in more detail in later reports. It is particularly important to examine the proposed biochemical mechanism from a mathematical point of view, to see the exact form of the γ-system equation which results (since this is complicated by such previously disregarded variables as decay rates, stability of old traces, etc.), and to try to obtain a more exact description of the biochemical reactions and molecular characteristics which this implies. At the same time, an empirical program has now been initiated at Cornell University to study transmitter substances and the mechanism of their action. This program, together with the work being done at many other such laboratories throughout the world, may eventually come up with evidence which will be sufficient to confirm or refute the hypotheses proposed here.

REFERENCES

1. Araki, T., M. Ito, and O. Oscarsson, "Anion Permeability of the Synaptic and Non-synaptic Motoneurone Membrane," *J. Physiol.*, 159, 1961, pp. 410–435.
2. Bartlett, F., *Remembering*, Cambridge U. P., 1954.
3. Beurle, R. L., "Properties of a Mass of Cells Capable of Regenerating Pulses," *Phil. Trans. Royal Soc. London*, B240, No. 669, 55.
4. Block, H. D., "The Perceptron: A Model for Brain Function," *Rev. Mod. Phys.*, 34, 1962, pp. 123–135.
5. Block, H. D., B. W. Knight, and F. Rosenblatt, "Analysis of a Four-Layer, Series-Coupled Perceptron," *Rev. Mod. Phys.*, 34, 1962, pp. 135–142.
6. Briggs, M. H., and G. B. Kitto, "The Molecular Basis of Memory and Learning," *Psych. Rev.* 69, 1962, pp. 537–541.
7. Brown, J., "Information, Redundancy, and Decay of the Memory Trace," in *Proceedings of Symposium on the Mechanization of Thought Processes*, H. M. Stationery Office, London, 1958.
8. Burns, B. D., *The Mammalian Cerebral Cortex*, Arnold, London, 1958.
9. Culbertson, J. T., *Consciousness and Behavior*, Brown, Dubuque, Iowa, 1950.
10. Curnow, R. N., and C. W. Dunnett, "The Numerical Evaluation of Certain Multivariate Normal Integrals," *Annals of Math. Stat.* 33, 1962, pp. 571–579.
11. Eccles, J. C., *The Neurophysiological Basis of Mind*, Clarendon, Oxford, 1953.
12. Eccles, J. C., *The Physiology of Nerve Cells*, Johns Hopkins, Baltimore, 1957.
13. Eccles, J. C., "The Effects of Use and Disuse on Synaptic Function.," in Fessard, Gerard, and Konorski (eds.), *Brain Mechanisms and Learning*, Blackwell, Oxford, 1961.
14. Eccles, J. C., "The Synaptic Mechanism for Postsynaptic Inhibition," in Florey (ed.), *Nervous Inhibition*, Pergamon, New York, 1961.
15. Elliott, K. A. C., I. H. Page, and J. H. Quastel, *Neurochemistry*, Thomas, Springfield, Ill., 1962.
16. Farley, B., and W. Clark, "Activity in Networks of Neuron-like Elements," in Cherry (ed.), *Information Theory*, Butterworths, Washington, 1961.
17. Fessard, A., R. W. Gerard, and J. Konorski, *Brain Mechanisms and Learning*, Blackwell, Oxford, 1961.
18. Florey, E., *Nervous Inhibition*, Pergamon, New York, 1961.
19. Florey, E., "Excitation, Inhibition, and the Concept of the Stimulus," in Florey (ed.), *Nervous Inhibition*, Pergamon, New York, 1961.
20. Gaito, J., "A Biochemical Approach to Learning and Memory", *Psych. Rev.*, 68, 1961, pp. 288–292.
21. Gerard, R. W., "What is Memory?" *Sci. Am.*, 189, 1953, pp. 118–126.
22. Gerard, R. W., "The Fixation of Experience," in Fessard, Gerard, and Konorski (eds.), *Brain Mechanisms and Learning*, Blackwell, Oxford, 1961.
23. Gerard, R. W., T. J. Chamberlain, and G. H. Rothschild, "RNA in Learning and Memory," *Science*, 140, 1963, p. 381.
24. Hebb, D. O., *The Organization of Behavior*, Wiley, New York, 1949.
25. Hebb, D. O., "Distinctive Features of Learning in the Higher Animal," in Fessard, Gerard, and Konorski (eds.), *Brain Mechanisms and Learning*, Blackwell, Oxford, 1961.
26. Hebb, D. O., "The Semi-autonomous Process: Its Nature and Nurture," *Amer. Psychol.* 18, 1963, pp. 16–27.

27. Hubel, D. H., and T. N. Wiesel, "Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex," *J. Physiol.*, 160, 1962, pp. 106–154.

28. Hydén, H., "A Molecular Basis of Neuron-Glia Interaction," in Schmidt, F. O. *Macromolecular Specificity and Biological Memory*, MIT Press, Cambridge, Mass., 1962.

29. Ito, M., P. G. Kostyuk, and T. Oshima, "Further Study on Anion Permeability of Inhibitory Post-synaptic Membrane of Cat Motoneurones," *J. Physiol.*, 164, 1962, pp. 150–156.

30. Jacobsen, A. L., "Learning in Flatworms and Annelids," *Psych. Bull.* 60, 1963, pp. 74–94.

31. John, E. R. "Some Speculations on the Psychophysiology of the Mind," in J. Scher, (ed.), *Theories of the Mind*, Free Press, New York, 1962.

32. Joseph, R. D., "On Predicting Perceptron Performance," *Record of IRE National Convention*, Part 2, New York, 1960.

33. Jung, R., "Neuronal Integration in the Visual Cortex," in Rosenblith (ed.), *Sensory Communication*, MIT Press, Cambridge, Mass, 1961.

34. Konorski, J., "The Physiological Approach to the Problem of Recent Memory," in Fessard, Gerard, and Konorski (eds.), *Brain Mechanisms and Learning*, Blackwell, Oxford, 1961.

35. Lashley, K. S., "In Search of the Engram," in Beach, Hebb, Morgan, and Nissen (eds.), *The Neuropsychology of Lashley*, McGraw-Hill, New York, 1960.

36. McLennan, H., "Inhibitory Transmitters—A Review," in Florey (ed.), *Nervous Inhibition*, Pergamon, N.Y., 1961.

37. Miller, G. A., "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psych. Rev.* 63, 1956, pp. 81–97.

38. Milner, B., and W. Penfield, "The Effect of Hippocampal Lesions on Recent Memory," *Trans. Am. Neurol. Assn.*, 1955, pp. 42–48.

39. Milner, P. M., "A Neural Mechanism for the Immediate Recall of Sequences," *Kybernetik*, No. 1, Berlin, July, 1961.

40. Morrell, F., "Electrophysiological Contributions to the Neural Basis of Learning," *Physiol. Rev.*, 41, 1961, pp. 443–494.

41. Morrell, F., "Information Storage in Nerve Cells," in Fields and Abbott (eds.), *Information Storage and Neural Control*, Thomas, Springfield, 1963.

42. Penfield, W., and T. Rasmussen, *The Cerebral Cortex of Man*, Macmillan, New York, 1950.

43. Penfield, W., and L. Roberts, *Speech and Brain Mechanisms*, Princeton U. P., Princeton, N.J., 1959.

44. Rosenblatt, F., *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan, Washington, 1962.

45. Rosenblatt, F., "A Comparison of Several Perceptron Models," in Yovits, Jacobi, and Goldstein (eds.), *Self-Organizing Systems-1962*, Spartan, Washington, 1962.

46. Rosenblatt, F., *A Theory of Biological Memory* (Cognitive Systems Research Program Report, Cornell University; in preparation.)

47. Roy, A., "On a Method of Storing Information," *Bull. Math. Biophysics*, 22, 1960, pp. 139–168.

48. Sandel, T. T., and N. Y. S. Kiang, "Off Responses from the Auditory Cortex of Anesthetized Cats: Effects of Stimulus Parameters," *Arch. Ital. de Biol.*, 99, 1961 pp. 105–120.

49. Schaefer, E., "Das Menschliche Gedächtnis als Informationsspeicher," *Elektronische Rundschau*, Telefunken, vol. 14, no. 3, 1959, pp. 79–84.

50. Schmidt, F. O., *Macromolecular Specificity and Biological Memory*, MIT Press, Cambridge, Mass., 1962.

51. Scoville, W. B., and B. Milner, "Loss of Recent Memory After Bilateral Hippocampal Lesions," *J. Neurol. Neurosurg. Psychiat.*, 20, 1957, pp. 11–20.

52. Shanes, A. M., "Quantitative Molecular Approach to the Permeability Changes of Excitation," *Science*, 140, 1963, pp. 51–53.

53. Shanes, A. M., "Membrane Permeability: Monolayer Relationships," *Science*, 140, 1963, pp. 824–825.

54. Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communication*, Univ. of Illinois Press; Urbana, 1959.

55. Smith, C. E., "Is Memory a Matter of Enzyme Induction?" *Science*, 138, 1962, pp. 889–890.

56. Szent-Györgyi, A., *Introduction to a Sub-Molecular Biology*, Academic, New York, 1960.

57. Von Foerster, H., *Das Gedächtnis: Eine Quantenmechanische Untersuchung*, F. Deuticke, Vienna, 1948.

58. Wechsler, D., "Engrams, Memory Storage, and Mnemonic Coding," *Amer. Psychol.* 18, 1963, pp. 149–153.

59. Widrow, B., "Generalization and Information Storage in Networks of Adaline 'Neurons.' In Yovits, Jacobi, and Goldstein (eds.), *Self-Organizing Systems—1962*, Spartan, Washington, 1962.

60. Wiersma, C. A. G., "Inhibitory Neurons: A survey of the History of their Discovery and of their Occurrence," in Florey (ed.), *Nervous Inhibition*, Pergamon, New York, 1961.