SESSION 3

PAPER 4

TWO THEOREMS OF STATISTICAL SEPARABILITY IN THE PERCEPTRON

by

DR. F. ROSENBLATT

BIOGRAPHICAL NOTE

Frank Rosenblatt, born in New Rochelle, New York, U.S.A., July 11, 1928, graduated from Cornell University in 1950, and received a PhD degree in psychology, from the same university, in 1956. He was engaged in research on schizophrenia, as a Fellow of the U.S. Public Health Service, 1951-1953. He has made contributions to techniques of multivariate analysis, psychopathology, information processing and control systems, and physiological brain models. He is currently a Research Psychologist at the Cornell Aeronautical Laboratory, Inc., in Buffalo, New York, where he is Project Engineer responsible for Project PARA (Perceiving and Recognizing Automaton).

TWO THEOREMS OF STATISTICAL SEPARABILITY IN THE PERCEPTRON

by

DR. FRANK ROSENBLATT

SUMMARY

A THEORETICAL brain model, the perceptron, has been developed at the Cornell Aeronautical Laboratory, in Buffalo, New York. The perceptron is a probabilistic system, capable of learning to recognize and differentiate stimuli in its environment. Previous reports have covered the theory of a class of perceptrons based on fixed-threshold neurons, similar to the McCulloch-Pitts model. The present paper introduces the concept of a "continuous transducer neuron", and outlines the proof of two theorems which indicate that a properly designed perceptron will be capable of spontaneously forming meaningful classifications of the stimuli in its universe, without being taught by an experimenter.

---

1. PROBABILISTIC MATHEMATICS VS. SYMBOLIC LOGIC

ONLY a few months before the Office of Naval Research began its support of the perceptron program, at the Cornell Aeronautical Laboratory, John von Neumann, one of the most outstanding advocates of the proposition that man might some day achieve an artificial device working on the same principles as the human brain, wrote the following prophetic passage *(ref. 4)*:

> "Logics and mathematics in the central nervous system...must structurally be essentially different from those languages to which our common experience refers ... When we talk mathematics, we may be discussing a *secondary* language, built on the *primary* language truly used by the central nervous system. Thus the outward forms of *our* mathematics are not absolutely relevant from the point of view of evaluating what the mathematical or logical language *truly* used by the central

nervous system is... Whatever the system is, it cannot fail to differ considerably from what we consciously and explicitly consider as mathematics."

What von Neumann is saying here deserves careful consideration. The mathematical field of symbolic logic, or Boolean algebra, has been eminently successful in producing our modern control systems and digital computing machines. Nevertheless, the attempts to account for the operation of the human brain by similar principles have always broken down under close scrutiny. The models which conceive of the brain as a strictly digital, Boolean algebra device, always involve either an impossibly large number of discrete elements, or else a precision in the "wiring diagram" and synchronization of the system which is quite unlike the conditions observed in a biological nervous system. I will not belabor this point here, as the arguments have been presented in considerable detail in the original report on the perceptron *(ref. 3)*. The important consideration is that in dealing with the brain, a different kind of mathematics, primarily statistical in nature, seems to be involved. The brain seems to arrive at many results directly, or "intuitively", rather than analytically. As von Neumann has pointed out, there is typically much less "logical depth" in the operations of the central nervous system than in the programs performed by a digital computer, which may require hundreds, thousands, or even millions of successive logical steps in order to arrive at an analytically programmed result.

Those readers who are familiar with the concept of the perceptron know that it is a model of a system which is primarily concerned with the recognition of the forms, sounds, and other stimuli which make up the ordinary physical world, as we know it through our own senses. The theory upon which this system is based is called the "theory of statistical separability". The mathematics upon which this theory stands, has much more in common with the mathematics of particle physics than with the mathematics of digital computers. The reason for this is fundamental: Boolean algebra, or symbolic logic, is well suited to the study of completely describable logical systems, but breaks down as soon as we attempt to apply it to systems on which complete information is not available. If we lack a detailed wiring diagram, but know only the statistical parameters, or probabilities of connection within a logical net, then the only way we can use Boolean algebra to determine the probable response of the system would be by complete enumeration of every possible connection diagram which meets the parametric constraints, whereupon we could actually count the number of alternatives which respond in each of the logically possible ways. In dealing with systems of any complexity, the number of possible connection diagrams becomes, for all practical purposes, infinite, so that we can not use this enumerative approach in practice, even though it may be possible in principle. Probability theory, on the other hand, is specifically designed to permit us

to make precise statements in the absence of complete information. If we know only a few parameters of a statistical distribution, we may be able to make highly accurate statements about the mass behavior of a collection of events, a logical network, or a nervous system, *even though we have never unravelled the detailed wiring diagram for any particular case.*

2. THE IMPORTANCE OF PERCEPTUAL PROCESSES FOR AUTOMATA

We have said that the perceptron is primarily concerned with the recognition of stimuli, or patterns, in its environment. In this, it is fundamentally different from any digital computer. Since computers are very good at something which people, by and large, do very badly, that is, arithmetic, they have been popularly represented as "giant electronic brains". This comparison seems to me an unfortunate one. It suggests to many people that because a computer does certain things that the brain can do, the brain must work something like a computer. I have already indicated above that I consider this position to be untenable. But in order to understand the unique capabilities of the perceptron, it might be helpful to consider the sort of thing that *can* be done by digital computers.

Computers, in general, are designed to follow rules. If we can set up rules for multiplication, we can design a computer to multiply. If we give this computer any two numbers, even if these specific numbers have not been considered by the designers of the machine, it can multiply them, and form the correct product. But this does not really satisfy our idea of original thinking, or intelligence. There has been no *discovery* involved; the correct answer simply follows from the fact that the rules of multiplication are completely universal, and apply to all numbers. Similarly, if we are ingenious enough to write a set of exact rules for minimizing the cost of some business operation, we can program a computer to minimize cost, and other such complex problems. In fact, if we can analyze the way in which people play chess sufficiently well to write an explicit set of rules for chess strategy, we can get a computer to play chess, as is being done with increasing success in recent programs.

But in all of these examples, the computer is following rules which are the result of human observation and analysis. Computers seem to share two main functions with the brain:

(a) Decision making, based on logical rules
(b) Control (as in guidance systems, automatic assembly lines, etc.) again based on logical rules.

The human brain performs these functions, together with a third: *interpretation of the environment.*

Why do we hold interpretation of the environment to be so important?

The answer, I think, is to be found in the laws of thermodynamics. A system with a completely self-contained logic can never spontaneously

improve in its ability to organize, and to draw valid conclusions from information. While there are certain trivial cases where this may appear questionable (for example, we might deliberately design a program to give wrong answers for the first ten trials and then modify itself to perform correctly from the eleventh trial on), by and large it seems to be a valid generalization. Even in the above case, the improvement after the tenth trial was anticipated and deliberately built into the program, so that it did not actually arise spontaneously. Spontaneous changes in such a system will, in general, lead to a deterioration, rather than an improvement, in its performance. On the other hand, a system which is capable of reorganizing its own logic, to correspond to a logical organization which already exists in the universe around it, takes on very different properties indeed. Such a system *can* improve (if it is properly constructed) by observing and learning from the organization of the surrounding world. The human brain is such a system. It is this ability to interpret the environment which allows the human brain to recognize and devise the logical rules which are applied by the computer. Conceptualization of the environment is the first step towards creative thinking.

So far as I know, the only machine prior to the perceptron which has shown itself to be capable of *spontaneous* improvement (as opposed to learning under the tutelage of an experimenter) has been Ashby's homeostat *(ref. 1)*. The homeostat, however, is not really a case in point here, because it is not really concerned, as we are, with the representation of meaningful information, but rather with the maintenance of an optimum state within the system.

The perceptron, as originally developed, would also have been incapable of spontaneous concept formation. The original perceptron could be taught to perceive differences between stimuli, by a process similar to that employed in training a dog. If you want to train a dog to come when called, you put him on a long leash, you say, "Come here", or "Come, Rover", and you pull him towards you. If you want him to sit down, you say, "Rover, sit", and push his tail down. In other words, we *force* the desired response. When he begins doing these things spontaneously, this indicates that we have passed on our own recognition of the difference between the words "Come here" and "Sit down", and that the dog can now perceive the distinction. This is really identical with the process that we originally studied with the perceptron.

It actually took only a very slight change in the dynamics of the perceptron to convert it into a system with very different capabilities, actually the first machine which is capable of having an original idea. It is the conversion to this new system which I wish to discuss for the remainder of this paper.

The key to this transition to a spontaneously organizing perceptron came with a recasting of our original mathematical analysis in a new and more elegant form, made possible by a revised concept of the basic unit, the

neuron. Let me begin, therefore, by presenting the rationale for this new concept.

3. THE CONTINUOUS TRANSDUCER NEURON

The logical model of the neuron which has been used almost universally in theoretical brain models during the last few decades is the "McCulloch-Pitts neuron" *(ref. 2)*. This assumes an all-or-nothing response, depending on whether the signal at some time, t, is above or below threshold. Moreover, and this is the most critical simplification assumed by the model, the cell is generally allowed to fire only at integral values of t, where time is measured in some convenient unit. Now actually, as is well known, a cell in the central nervous system does not respond in so simple a fashion. Such a cell is typically under continuous bombardment by a great number of impulses, which summate temporally as well as spatially. These impulses, eventually, may produce a region of depolarization in the membrane, and when this happens, the cell fires. As a consequence of this, the *frequency* of the cell is likely to be roughly proportional to the net, or mean intensity of the stimulation received. Lightly stimulated, the cell may respond at a very low frequency; under intense stimulation, the frequency increases. Thus, even though each impulse generated by the cell may be of the same amplitude, the cell is actually acting as a continuous transducer of the stimulus energy, if we measure the energy transmitted as a rate per unit time. Such a cell may carry considerably more information than we would suppose from a simple consideration of its discontinuous on-off properties.

The transfer function of a continuous transducer neuron, a_i, at time t, is equal to $\alpha_i(t) \cdot v(t)$, where v_i (the "value" of the neuron a_i) is a scale factor which may fluctuate with time, and which determines the relative amplitude of the output. We will have more to say about this "value" presently. A simplified continuous transducer neuron is shown in *fig. 1* (a). Note that there are an equal number of excitatory and inhibitory inputs, so that the input signal α, will be zero if stimulation is uniform over the entire sensory field. If there is a gradient of stimulation, or a localized region of stimulation such that the total excitatory component (Σx) is greater than the total inhibitory component (Σy), the input signal, α, will be positive. If the input from a given stimulus is primarily inhibitory, a will be negative.

For use in the perceptron, one additional assumption must be made about the logic of the neuron; it must be given a memory. This memory takes the form of changes in the magnitude of the "value", v. The value, v_i, of the neuron a_i, is represented by a stored quantity, which might be represented physically by an electric charge, the position of a potentiometer, the light

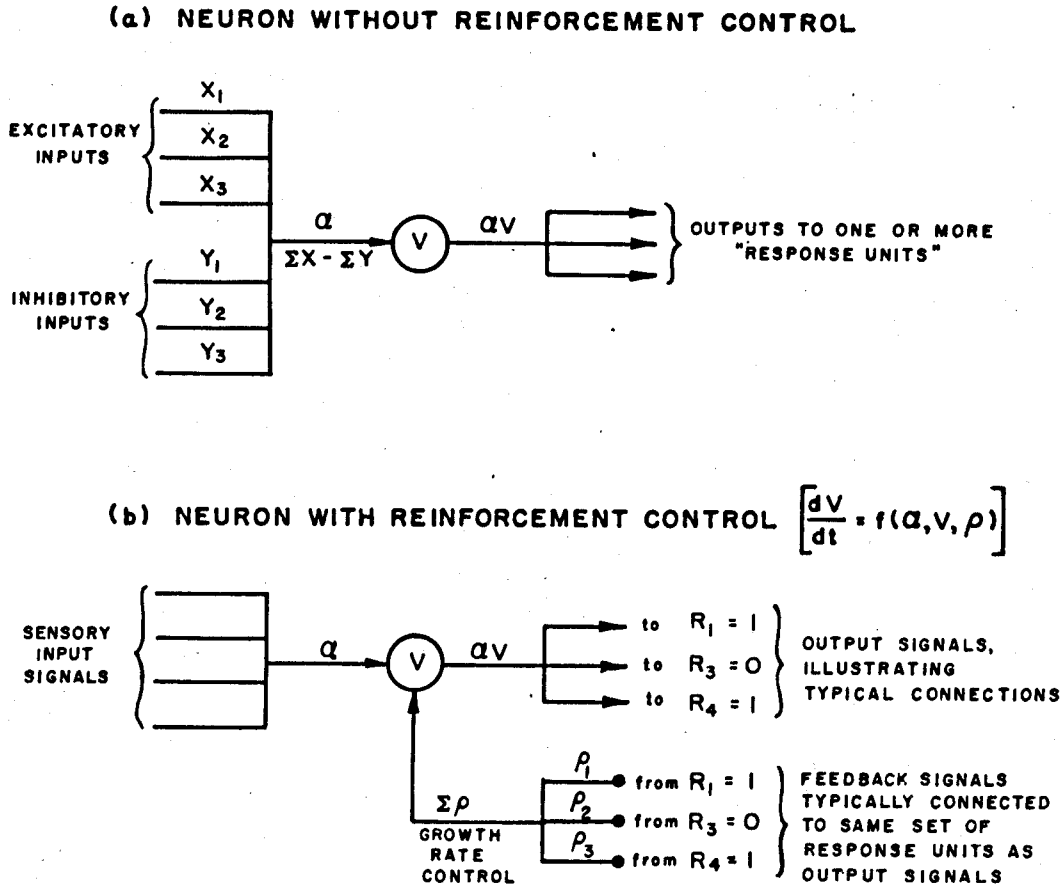


Fig. 1. Continuous Transducer Neurons

flux through a variable aperture, or any other convenient means. If we are thinking in terms of a biological analog, the "value" might be the degree of polarization of the membrane, the energy reserve of the cell, the volume of cytoplasm, or any other enduring condition which will affect the potency, amplitude, or frequency of the cell's output. The higher the "value", the more "powerful" the output of the cell, measured by whatever variable happens to be relevant for the system in question. For a perceptron, where the output function is equal to $\alpha\nu$, we assume that the amplitude of the output pulses will be proportional to the value.

The important question about the value, from the standpoint of memory storage, is how it changes, as a function of the activity of the system. *Figure 1(b)* shows a continuous transducer neuron with an additional input signal, labelled $\Sigma\rho$, which controls the growth rate of the neuron. The components of $\Sigma\rho$ (i.e., ρ_1, ρ_2, etc.) are typically feedback signals, which originate from the next logical layer of cells, which are called "Response units", or *R*-units. In the fixed threshold neurons used in the earlier perceptron systems *(ref. 3)*, the growth of the value was a function which depended on whether or not the cell was active. If the cell was active (α greater than threshold), an increment would be added to ν, which depended primarily on the presence or absence of a feedback signal corresponding to $\Sigma\rho$. In the continuous transducer system, three basic forms of growth function are of particular interest:

$$\frac{d\nu}{dt} = \alpha\Sigma\rho \tag{1}$$

$$\frac{d\nu}{dt} = \alpha\Sigma\rho - \delta\nu \tag{2}$$

$$\frac{d\nu}{dt} = \Sigma\rho(\alpha - \delta\nu) \tag{3}$$

The first of these equations is a growth function with a zero decay component. It is equivalent to (2) and (3) with δ set at zero. In the second two cases, there is a decay component proportional to the current magnitude of the value. δ, the decay coefficient, is a constant less than 1. All fixed-threshold perceptrons considered previously *(ref. 3)* share the characteristic of the first of these growth functions that the value of a neuron, or association cell, can continue to grow without bound. Note that over a set of neurons, since α is as likely to be negative as positive, the expected rate of growth will remain zero. Nonetheless, the *variance* of the value over the set of cells will tend to increase towards infinity, as time goes on, unless there is a decay function which increases monotonically with ν, as in (2) and (3). The importance of this distinction will become clear presently.

4. ORGANIZATION OF A PERCEPTRON

Let us now consider the organization of a very simple perceptron, made up of such "continuous transducer neurons" *(fig.2)*. Note that the logical depth of this system is no greater than two; the first step is from the sensory system to the association system, and the second is from the association to response system. The association system is composed of association cells (A-units) which are continuous transducer neurons. The inputs of these A-units come from origin points (which may be randomly distributed) in the sensory system, or "retina", and their outputs go to

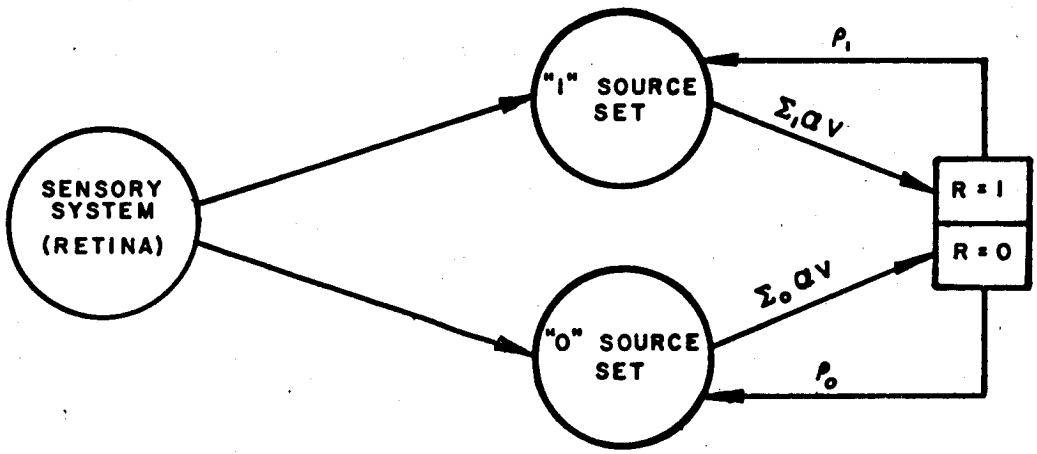


Fig.2. Organization of a Simple Perceptron

the single response unit which is shown in this system. (The organization of systems with large numbers of response units has been considered in previous work *(ref.3)*. Those A-units which tend to turn the response "on" (to the condition $R = 1$) are collectively designated the "1 source-set", while those A-units which tend to turn off, or inhibit, the response, are designated the "0 source-set". If $R = 1$, a reinforcement signal, ρ_1, is transmitted to all members of the 1 source-set. If $R = 0$, a reinforcement signal, ρ_0, is transmitted to all members of the 0 source-set. When a stimulus is presented to the visual system, the output signals from each of the source-sets are summed, and the sign of the difference, $\Sigma_1 \alpha v - \Sigma_0 \alpha v$, determines whether the response is 1 or 0. In other words, if the 1 source-set delivers a stronger signal, the response will be "1"; if the 0 source-set delivers a stronger signal, the response will be "0". Note that the

"reinforcement operator", $\Sigma\rho$, which appears in the growth rate equations (1), (2) and (3) will, in this system, be either 1 or 0, depending on the current state of the response.

Corresponding to each of the three growth rate equations, we have the three limit equations, for the value of an A-unit:

$$\nu \longrightarrow \pm\infty \tag{1a}$$

$$\nu \longrightarrow \frac{E\alpha\rho f_\rho}{\delta} \tag{2a}$$

$$\nu \longrightarrow \frac{E\alpha\rho}{\delta} \tag{3a}$$

where $E\alpha$ is the expected value of the input signal to the A-unit in question, and f_ρ is the mean frequency with which the A-unit is reinforced (i.e., the frequency with which $\Sigma\rho = 1$). The first and third of these cases will be the ones with which we are primarily concerned. It can be shown that the second case (which is more plausible for biological units, since the rate of decay does not depend on the feedback signal, $\Sigma\rho$) will merely lead to a somewhat weaker form of the system represented by the third alternative.

5. SOME IMPORTANT CORRELATION COEFFICIENTS

In the analysis of fixed threshold perceptrons *(ref. 3)*, it was shown that the ability of the perceptron to discriminate between similar forms depends on the proportion of A-units activated in common by each of the forms to be distinguished. The expected value of this proportion is called P_c. In dealing with classes of forms, it was found possible to measure the "similarity" of the classes in terms of P_c. In the case of continuous transducer perceptrons, however, where every A-unit may be "active", albeit at widely scattered frequencies, measurements based on probabilities of activation are no longer suitable. In place of P_c, we will make use of several product moment correlation coefficients to measure the similarity of the activity induced by different classes of stimuli. A class of stimuli, in this context, means any set of forms, or retinal projections, which we will regard, arbitrarily, as the same "kind" of visual object, e.g., letters of the alphabet, geometrical shapes such as squares and triangles, dogs and cats, etc. We will be concerned, particularly, with the binary discrimination of stimuli into two classes, which we will designate S_1 and S_2.

Two kinds of correlations are of particular importance in the analysis of similarity, for a continuous transducer perceptron. These correlations can be written, in abbreviated form, as $r_{\alpha\bar{\alpha}}$ and $r_{\bar{\alpha}\bar{\alpha}}$. $r_{\alpha_1\bar{\alpha}_1}$ is the correlation, measured over the set of all A-units, of the input signals, α, from some particular stimulus of class S_1, with the expected value of the input

signals from the class S_1. In other words, $r_{\alpha_1\bar{\alpha}_1}$ (which is roughly analogous to $P_{c_{11}}$, in the old system of notation) measures the similarity of any arbitrary stimulus of a class to the remainder of the class; it is a measure of the *coherence* of the stimulus classes. Note that this measure of similarity is not based on the stimuli themselves, but on the signals received by the association system. $r_{\alpha\bar{\alpha}}$, the second of the two basic correlations, is the correlation of the expected values of α for the two classes of stimuli, S_1 and S_2*. Both correlations are measured over the set of all admissible A-units, i.e., all units which meet the parametric constraints of the system.

It should be emphasized that the kinds of similarity relationships which emerge in a system of this sort will be a function of the organization of the connection system, by which signals are transmitted from sensory points to A-units. In our original studies of the perceptron, it was assumed that the origin points for the input fibers to an A-unit were randomly scattered throughout the retina, or sensory mosaic. This system, as might be expected, leads to a high degree of sensitivity to the *location* of a stimulus in the visual field, and tends to create a bias such that the influence of large stimuli outweighs the effect of small stimuli, in the development of learning and memory. An alternative system, now being investigated, calls for a polarization of the origin points for an A-unit about an arbitrarily selected line in the retinal field, such that all connections originating on one side of the line are excitatory, while all connections originating on the other side are inhibitory. The density of connections, in this system, increases in the neighborhood of the line, or "axis of polarization". Such a system is primarily sensitive to the location and direction of contours, rather than to illuminated areas per se, and the resulting measures of similarity, using the above correlation coefficients, will naturally be quite different. In order to minimize sensitivity, to the *location* of a figure, it is generally desirable to define α in such a way that negative values are eliminated. This is equivalent to setting a zero threshold for the A-unit; otherwise the expected value gain, following the translation of a figure over the field, will always be zero, if the numbers of excitatory and inhibitory connections are equal.

Despite their relativity with respect to the connection system, these "similarity correlations" ($r_{\alpha\bar{\alpha}}$ and $r_{\bar{\alpha}\alpha}$) appear to bear a definite relationship to our intuitive, phenomenological concept of "similarity". Particularly in the case of the contour-sensitive connection system mentioned above, it can be shown that classes of forms which we tend to regard as "similar" (e.g., different sizes and locations in the field of the same letter of the alphabet) will have a high value of $r_{\alpha\bar{\alpha}}$, while for dissimilar, or randomly selected forms, $r_{\alpha\bar{\alpha}}$ will be low. Similarly, if we consider $r_{\bar{\alpha}\alpha}$

---

* $r_{\bar{a}a}$ is used as an abbreviated notation for $r_{\bar{a}_1\bar{a}_2}$

between two strongly dissimilar classes of forms (such as horizontal and vertical bars, or the letters E and X) we will generally obtain a lower value than between similar classes (such as lower case "x" and upper case "X").

Now let us consider how these correlations come to be represented in the perceptron.

If any association cell is exposed exclusively to members of a single stimulus class, or if the reinforcement operator, ρ, is equal to 1 for all members of class S_1 and 0 for all members of class S_2, then the value of this cell will grow in a direction and at a rate which is determined by $\bar{\alpha}_1$, the expected value of the input signal from S_1 stimuli. Over a set of such units, therefore, as time goes to infinity, we would expect

$$r_{\alpha\nu} \longrightarrow r_{\alpha\bar{\alpha}} \tag{4}$$

where $r_{\alpha\nu}$ is the correlation of the input signals from any one stimulus (selected at random from the class in question) with the value currently stored in the A-units.

Now, if we "force" the responses of the system, so that, for example, in an environment of triangles and circles, the perceptron is made to give the response $R = 1$ consistently for all triangles, and $R = 0$ consistently for all squares, then in the 1 source-set we get

$$\begin{cases} r_{\alpha_1\nu_1} \longrightarrow r_{\alpha_1\bar{\alpha}_1} \\ r_{\alpha_2\nu_1} \longrightarrow r_{\alpha_2\bar{\alpha}_1} \end{cases} \qquad (Er_{\alpha_2\bar{\alpha}_1} < Er_{\alpha_1\bar{\alpha}_1}) \tag{5a}$$

and in the 0 source-set

$$\begin{cases} r_{\alpha_1\nu_0} \longrightarrow r_{\alpha_1\bar{\alpha}_2} \\ r_{\alpha_2\nu_0} \longrightarrow r_{\alpha_2\bar{\alpha}_2} \end{cases} \qquad (\mathrm{Er}_{\alpha_1\bar{\alpha}_2} < \mathrm{Er}_{\alpha_2\bar{\alpha}_2}) \tag{5b}$$

In other words, the values of the A-units in each source-set become correlated with the expected values of the input signals from the class of stimuli which has been "associated" to the corresponding response, and the correlation with signals from stimuli of the opposite class are expected to be smaller by a factor which varies with $r_{\bar{\alpha}\bar{\alpha}}$. Thus, if the stimulus classes are completely independent ($r_{\bar{\alpha}\bar{\alpha}} = 0$) we should expect (for the "1" source-set)

$$\begin{cases} r_{\alpha_i\nu_i} \longrightarrow r_{\alpha_i\bar{\alpha}_i} \\ r_{\alpha_j\nu_i} \longrightarrow 0 \end{cases}$$

Now consider the signals which are transmitted from the association system to the *R*-unit *(fig. 2)*. These signals are equal to the sum of $\alpha\nu$ for the source-set in question. But the correlation coefficient, $r_{\alpha\nu}$, for either source-set, has the formula:

$$r_{\alpha\nu} = \frac{\Sigma(\alpha - \bar{\alpha})(\nu - \bar{\nu})}{N\sigma_\alpha\sigma_\nu} = \frac{\Sigma\alpha\nu + [N\bar{\alpha}\bar{\nu} - \bar{\nu}\,\Sigma\alpha - \bar{\alpha}\Sigma\nu]}{N\sigma_\alpha\sigma_\nu} \tag{6}$$

where N = the number of A-units in the source-set.
If we assume that the two classes of stimuli, S_1 and S_2, produce the same distribution of signal amplitudes at the A-units, then the expected values of σ_α and $\bar{\alpha}$ will be identical for the 1-source-set and the 0-source-set. If, in addition, the same number of stimuli have been associated to each of the two responses, the variable ν will also have the same expected distribution in the two source-sets (under the experimental conditions for "forced learning", described above) so that σ_ν and $\bar{\nu}$ can be considered equal for the two cases. N is also assumed to be equal for the two source-sets. Thus, the term in brackets in the numerator of (6) reduces to a constant, which will be zero if $\bar{\alpha}$ and $\bar{\nu} = 0$. The denominator also becomes a constant, so that the correlation can be expressed in the form:

$$r_{\alpha\nu} = \frac{\Sigma\alpha\nu}{C_1} + C_2 \tag{7}$$

Consequently, the expected difference of the signals from the two source-sets, $\Sigma_1\alpha\nu - \Sigma_0\alpha\nu$, will be directly proportional to the difference of the correlations $r_{\alpha_i\nu_1}$ and $r_{\alpha_i\nu_0}$. We have already seen [in equations (5a) and (5b)] that one of these correlations will be proportional, in the limit, to $r_{\alpha_i\bar{\alpha}_i}$ and the other to $r_{\alpha_i\bar{\alpha}_j}$. The difference will be positive if the subscript i refers to the first class of stimuli, negative if i refers to the second class.

Under these conditions, therefore, we would expect that the presentation of any arbitrary stimulus from class S_1 will tend to evoke the response which has been formerly associated to S_1 stimuli, while the presentation of an S_2 stimulus should evoke the opposite response. This, actually, is the essence of the theory of "forced learning" in a continuous transducer perceptron. Predictions of the reliability of the response, as a function of the number of A-units in the system, can be made by making use of standard techniques for determining the probability of an error in the difference of two correlation coefficients, equal to or greater than the expected difference, $r_{\alpha_i\bar{\alpha}_i} - r_{\alpha_i\bar{\alpha}_j}$.

Simulation experiments, using the IBM 704 digital computer at the Cornell Aeronautical Laboratory, have substantiated the predictions of this theory. It has been possible to teach a simulated perceptron to recognize

the difference between geometrical forms, letters of the alphabet, and positions on the retina, using continuous transducer neurons. The most important question, however, still remained unanswered at this point in the program: What would such a perceptron do if, instead of forcing the desired response during a "training period", we simply turned it loose in an arbitrary environment? Would it ever, spontaneously, arrive at the desired concepts?

6. THE THEORY OF CLASS C PERCEPTRONS

The problem of spontaneous organization was originally investigated with respect to perceptrons in which the growth function of the value corresponds to equation (1). This study led to the following theorem and corollary:

THEOREM 1:

There exists a class, C, of perceptrons, which tend toward a statistically stationary state such that each binary response (R = 1, 0) becomes established as either 1 or 0 universally, for all stimuli, with an error probability, ϵ, which approaches zero as $t \rightarrow \infty$. In the terminal condition, each binary response gives 0 bits of information with regard to the current stimulus.

COROLLARY:

A perceptron in which the values of the A-units are allowed to grow without bound, and in which the sensory origin points of the A-units are randomly located, is a member of the class C.

In order to prove this theorem, it is clearly sufficient to prove the corollary for any particular case. We will consider, as a perceptron of this type, a continuous transducer perceptron with a single binary response, such as the one shown in *fig. 2*, with growth function (1). As there is no decay in the values of the A-units, they can continue to grow indefinitely, as indicated by equation (1a). The environment of this system is assumed to consist of two arbitrary stimulus classes, S_1 and S_2. Stimuli are assumed to occur in a random sequence, where the probability of an S_1 stimulus is $P(S_1)$, and the probability of an S_2 stimulus is $P(S_2)$. We must show that whatever the relationship of the stimulus classes, the condition stated in the theorem will tend to occur.

In order to develop a proof for this theorem, it will be necessary, first, to define some terms. In particular, we must consider several different methods of representing and classifying the "state of the system" at time t.

The description of the "state of the system", which gives us the information which we ultimately want, takes the form of a 2 x 2 probability matrix,

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

where the first subscript denotes the stimulus class (S_1 or S_2) and the second denotes the response ($R = 0$ or $R = 1$). Thus P_{21} = the probability that a Class 2 stimulus would evoke the response $R = 0$, at time t.

The matrix P can be classified into three cases, as follows:

P is *horizontally symmetric* if both elements of one column are greater than 0.5, and both elements of the other column are less than 0.5, i.e.,

$$P = \begin{pmatrix} .5 + \Delta & .5 - \Delta \\ .5 + \Delta & .5 - \Delta \end{pmatrix} \quad \text{or} \quad P = \begin{pmatrix} .5 - \Delta & .5 + \Delta \\ .5 - \Delta & .5 + \Delta \end{pmatrix}$$

P is *diagonally symmetric* if one pair of diagonal elements are greater than 0.5, and the other pair are less than 0.5, i.e.,

$$P = \begin{pmatrix} .5 + \Delta & .5 - \Delta \\ .5 - \Delta & .5 + \Delta \end{pmatrix} \quad \text{or} \quad P = \begin{pmatrix} .5 - \Delta & .5 + \Delta \\ .5 + \Delta & .5 - \Delta \end{pmatrix}$$

P is *neutral* if the elements of one or more rows are both equal to 0.5.

Theorem 1, stated in terms of these definitions, predicts that in a Class C perceptron the terminal state will always tend to be horizontally symmetric, and that Δ will approach 0.5 as $t \to \infty$.

In order to evaluate P at some arbitrary time, t, and to show the trend in its development through time, two other matrices are helpful. The first of these is the 2 x 2 matrix, N, where

$$N = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$$

Here the row subscript again refers to the stimulus, and the column subscript to the response. Each n is equal to the number of times that the indicated response has occurred in response to a stimulus from the indicated class. n_{21}, for example, is the number of times that a stimulus of Class 2 has evoked the response $R = 0$, throughout the history of the system up to time t.

The remaining matrix, K, is again a 2 x 2 matrix of the same form:

$$K = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix}$$

where $k_{ij} = r_{\alpha_i \nu_j} \sigma_{\nu_j}$. The elements of this matrix are proportional to the expected values of the output signals from the j-source-set, in response to an arbitrary stimulus of Class i. The significance of the correlations $r_{\alpha\nu}$ has already been indicated in the preceding section. For simplicity, we assume a perceptron so designed that $\bar{v} = 0$. The conditions assumed for

equation (7) all apply, except for the fact that σ_{ν} can no longer be assumed equal for the two source-sets, since the sets may not have been "reinforced" an equal number of times. The elements of K have, therefore, been corrected for the effect of σ_{ν}.

Now, it can be shown that $K(t)$ is a function dependent only on $N(t)$, and that $P(t)$ depends exclusively on $K(t)$. Moreover, we can show that $\{N(t)\}$ is a function exclusively of $\{N(t-1)\}$, $\{P(t-1)\}$, $P(S_1)$ and $P(S_2)$, where the last two probabilities have been defined as constants. Consequently, the matrix N is a Markovian process, and P is a function of this Markovian process.

Our general procedure will consist of showing that there exists a path by which any state of the system can progress to one of the two admissible terminal states,

$$P = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

It can be shown that, since the system is fundamentally Markovian, these states are "trapping", i.e., once the system has arrived in either of these states, by any path whatever, the probability of departing from that state is zero. The only other states that might be trapping are the diagonal states, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. The proof that the system will always, given sufficient time, tend towards one of the horizontal rather than one of the diagonal terminal states, rests on the fact that as long as t remains finite, the states will never be *absolutely* trapping, i.e., there always remains some error probability, and that this error probability tends to be greater for the diagonal than for the horizontally symmetric states.

The initial state of the system, at $t = 0$, can generally be characterized by the conditions:

$$N(0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$K(0) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$P(0) = \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix}$$

or, in abbreviated notation,

$$\begin{pmatrix}0&0\\0&0\end{pmatrix}, \begin{pmatrix}0&0\\0&0\end{pmatrix}, \begin{pmatrix}.5&.5\\.5&.5\end{pmatrix}$$

where the matrices are assumed to be written in the order N, K, P. If we now treat the occurrence of a stimulus, S_1 or S_2, as an operator on the state of the system, we can indicate the development of the system over any period of time as a tree, with a fourfold branching of possibilities at each successive moment in time (assuming time, for the sake of convenience, to be quantized rather than continuous):

$$\begin{pmatrix}0&0\\0&0\end{pmatrix}, \begin{pmatrix}0&0\\0&0\end{pmatrix}, \begin{pmatrix}.5&.5\\.5&.5\end{pmatrix} \quad \begin{array}{l} \xrightarrow{P_{11}\,P(S_1)} \begin{pmatrix}1&0\\0&0\end{pmatrix} \\ \xrightarrow{P_{12}\,P(S_1)} \begin{pmatrix}0&1\\0&0\end{pmatrix} \\ \xrightarrow{P_{21}\,P(S_2)} \begin{pmatrix}0&0\\1&0\end{pmatrix} \\ \xrightarrow{P_{22}\,P(S_2)} \begin{pmatrix}0&0\\0&1\end{pmatrix} \end{array} \qquad (8)$$

To each of these new N matrices, there again corresponds a K and a P matrix, which governs the probabilities on the next split, etc. In general, $N(t) = N(t-1)$, with one of the four elements incremented by 1 for each of the four possibilities, as indicated in equation (8), with the probabilities indicated by the expressions above the arrows. Thus, the sum of all the elements in N is always equal to t, and the sum of the elements in either column is equal to the number of occurrences of the corresponding response, in the prior history of the system.

The number of terminal branches in a tree of the form indicated in (8) will be equal to 4^t, a number which grows large far too rapidly to permit exhaustive computation of the spectrum of terminal states beyond the first few generations. The probability of any given terminal state or branch is equal to the product of the branch probabilities $(P_{sr}P(S_s))$ for each step of the sequence prior to t, where S_s is the stimulus, and R_r the response which occurred. These "world lines" of the system yield the expected values of the elements in the terminal P matrices, and the total probability that the system will terminate in a horizontal or diagonal state can be obtained by summing the branch probabilities of the terminal branches of each class.

In order to prove the theorem, which is our primary concern here, we must first show how the P matrix depends on N. The first step will be the development of the elements, k_{ij}, of the K matrix.

We have said that $k_{ij} = r_{\alpha_i \nu_j} \sigma_{\nu_j}$, which we have seen to be proportional to $\Sigma_j \alpha\nu$, the expected output signal from the j-source-set, in the presence of a stimulus of Class i. If we let $\sigma^2_{\nu\alpha}$ be the component of the variance of ν which is correlated with α, then

$$r^2_{\alpha\nu} = \frac{\sigma^2_{\nu\alpha}}{\sigma_\nu{}^2}$$

and

$$k^2_{ij} = \frac{\sigma^2_{\nu_j \alpha_i}}{\sigma^2_{\nu_j}} \cdot \sigma_{\nu_j}{}^2 = \sigma_{\nu_j}{}^2_{\alpha_i} \tag{9}$$

In other words, the elements, k_{ij}, are measures of the expected covariance of α and ν, for stimulus Class i and source-set j. Now this covariance comes from the n_{1j} increments to the value from S_1 stimuli, and the n_{2j} increments from S_2 stimuli. It is convenient, for purposes of analysis, to factor the square of the required covariance in an approximate fashion as follows:

$$\sigma_{\nu_j}{}^2_{\alpha_i} \approx r^2_{\alpha_i \bar{\alpha}_i} \sigma_{\nu_j}{}^2_{\bar{\alpha}_i} \tag{10}$$

which means that we need only find the covariance of ν_j and $\bar{\alpha}_i$, the expected value of α for the stimulus class S_i. Now each increment to the value of the source-set has an expected value equal to $\Sigma\alpha$, for all A-units in the source-set. This increment has an expected variance which we can arbitrarily set equal to unity. This variance can be divided into four components, as follows:

$\Delta_1 \equiv$ Variance which is "unique" to the particular stimulus, and uncorrelated with either $\bar{\alpha}_1$ or $\bar{\alpha}_2$

$\Delta_2 \equiv$ Variance which is correlated with $\bar{\alpha}_1$ but not with $\bar{\alpha}_2$

$\Delta_3 \equiv$ Variance which is correlated with both $\bar{\alpha}_1$ and $\bar{\alpha}_2$

$\Delta_4 \equiv$ Variance which is correlated with $\bar{\alpha}_2$ but not with $\bar{\alpha}_1$

The magnitudes of these four components will be:

$$\Delta_1 = 1 - (\Delta_2 + \Delta_3 + \Delta_4)$$

$$\Delta_2 = r^2_{\alpha_i \bar{\alpha}_i} \ (1 - r^2_{\bar{\alpha}\bar{\alpha}})$$

$$\Delta_3 \equiv r^2_{\alpha_i \bar{\alpha}_i} \ r^2_{\bar{\alpha}\bar{\alpha}}$$

$$\Delta_4 = r^2_{\alpha_i \bar{\alpha}_j} (1 - r^2_{\bar{\alpha}\bar{\alpha}})$$

The first and fourth of these components clearly makes no contribution to $\sigma^2_{\nu\bar{\alpha}}$. The second two components each make a contribution, which combine in the covariance as follows:

$$\sigma^2_{\nu_j \bar{\alpha}_i} = n^2_{ij}\,\Delta_{2(i)} + \left(\sqrt{n^2_{1j}\,\Delta_{3(1)}} + \sqrt{n^2_{2j}\,\Delta_{3(2)}}\right)^2$$

$$= n^2_{ij}\, r^2_{\alpha_i \bar{\alpha}_i}\left(1 - r^2_{\bar{\alpha}\bar{\alpha}}\right) + \left[n_{1j}\, r_{\alpha_1 \bar{\alpha}_1}\, r_{\bar{\alpha}\bar{\alpha}} + n_{2j}\, r_{\alpha_2 \bar{\alpha}_2}\, r_{\bar{\alpha}\bar{\alpha}}\right] \quad (11)$$

Multiplying through by $r^2_{\bar{\alpha}\bar{\alpha}}$, as indicated in equation (10), we obtain:

$$k^2_{ij} = n^2_{ij}\, r^4_{\alpha_i \bar{\alpha}_i}\left(1 - r^2_{\bar{\alpha}\bar{\alpha}}\right) + r^2_{\alpha_i \bar{\alpha}_i}\, r^2_{\bar{\alpha}\bar{\alpha}}\left[n_{1j}\, r_{\alpha_1 \bar{\alpha}_1} + n_{2j}\, r_{\alpha_2 \bar{\alpha}_2}\right]^2 \quad (12)$$

One point which has been overlooked thus far concerns the signs of the k_{ij}. Since k_{ij} has been squared in the course of this analysis, the sign of the element, as obtained from equation (12), might be either positive or negative. Now it must be remembered that k_{ij} is intended to be proportional to the output signal of the j-source-set, $\Sigma_j \alpha\nu$, and to $r_{\alpha_i \nu_j}$. If the values, ν_j, have been built up exclusively by increments from stimuli of the opposite class from Class i, then it is quite conceivable that $r_{\alpha_i \nu_j}$ could be negative, if $r_{\bar{\alpha}\bar{\alpha}}$ were negative. $r_{\alpha_i \nu_j}$ could also be negative, for some specific stimulus of Class i, if $r_{\alpha_i \bar{\alpha}_i}$ were negative for the stimulus in question. Under all other conditions, the signs will remain positive. Now, if we limit our discussion to perceptrons with random spatial distributions of origin points for the A-unit input fibers, it can be shown that $r_{\bar{\alpha}\bar{\alpha}}$ will always be positive. If the stimulus classes, S_1 and S_2, are totally disjunct, $r_{\bar{\alpha}\bar{\alpha}}$ may be equal to zero, but it will never go negative. If, on the other hand, the distribution of origin points is organized in some special pattern, such as the polarized contour-sensitive distribution described above, it is possible to obtain negative values of $r_{\bar{\alpha}\bar{\alpha}}$. If this should be the case, then there will be certain environments in which the perceptron which has been described will not behave as a Class C system, but will begin to act as a Class C' system (to be described in the following section) instead. Let us therefore limit ourselves to perceptrons with random spatial distributions of origin points (as indicated in the Corollary), for which the elements k_{ij} will always be positive. We should also assume a "well behaved" system in which $r_{\alpha_i \bar{\alpha}_i}$ is always positive, for any stimulus of Class i.

Now the difference between the two k's of the same row, $k_{i2} - k_{i1}$, will clearly be proportional to the difference, $\Sigma_1 \alpha\nu - \Sigma_0 \alpha\nu$, between the signals from the two source-sets, which determines the response. The probability that the response $R = 1$ will occur is a positive monotonic function of this difference. It is also clear from equation (12) that if $r_{\bar{\alpha}\bar{\alpha}} \neq 0$, the magnitude of the k_{ij} will increase monotonically with $n_{1j} + n_{2j}$, i.e., with the column sums of the N matrix. Consequently, if the ratio of the sum of the i-column to the sum of the j-column should increase without bound, it is

clear that p_{1i} and p_{2i} will tend towards unity, while p_{1j} and p_{2j} tend towards zero, which is the condition of horizontal symmetry predicted by the theorem. It is also clear that, if the elements of one row of the N matrix are equal, then an inequality between the elements of the remaining row will cause the P matrix to be horizontally symmetric.

Let us now consider what happens as soon as a stimulus of either class is presented to the perceptron [equation (8)]. It is clear that the first increment will immediately introduce a bias towards either the left column or the right column, so that equation (8), in more complete form, can be written:

$$\begin{pmatrix}0&0\\0&0\end{pmatrix},\ \begin{pmatrix}0&0\\0&0\end{pmatrix},\ \begin{pmatrix}.5&.5\\.5&.5\end{pmatrix} \longrightarrow \begin{cases} .5P(S_1) \longrightarrow \begin{pmatrix}1&0\\0&0\end{pmatrix}, \begin{pmatrix}k&0\\k&0\end{pmatrix}, \begin{pmatrix}.5+\Delta&.5-\Delta\\.5+\Delta&.5-\Delta\end{pmatrix} \\ .5P(S_1) \longrightarrow \begin{pmatrix}0&1\\0&0\end{pmatrix}, \begin{pmatrix}0&k\\0&k\end{pmatrix}, \begin{pmatrix}.5-\Delta&.5+\Delta\\.5-\Delta&.5+\Delta\end{pmatrix} \\ .5P(S_2) \longrightarrow \begin{pmatrix}0&0\\1&0\end{pmatrix}, \begin{pmatrix}k&0\\k&0\end{pmatrix}, \begin{pmatrix}.5+\Delta&.5-\Delta\\.5+\Delta&.5-\Delta\end{pmatrix} \\ .5P(S_2) \longrightarrow \begin{pmatrix}0&0\\0&1\end{pmatrix}, \begin{pmatrix}0&k\\0&k\end{pmatrix}, \begin{pmatrix}.5-\Delta&.5+\Delta\\.5-\Delta&.5+\Delta\end{pmatrix} \end{cases} \tag{13}$$

Each of the resultant first generation matrices is of horizontally symmetric form. (In the event that $r_{\overline{\alpha}\overline{\alpha}} = 0$, the first generation matrices would be neutral, but this can be regarded as a limiting case for the Class C perceptron, in which horizontally and diagonally symmetric states remain equally probable, no matter how far t is extended.) It is now clear that, *whichever stimulus is the next to occur*, the probability is greater than 0.5 that the same response will occur which occurred originally; and if this does happen, then it is evident that the same column of the N-matrix which was incremented before will be incremented again, leading to an increase in Δ, and strengthening the tendency towards the horizontal form. Thus, we find that a matrix which is already a horizontally symmetric case tends to remain so, and that the application of further stimuli will always tend to increase the existing tendency. In the limit, as Δ approaches 0.5, it is clear that the probability of ever gaining an increment in the "weak" column of the N-matrix goes to zero, and consequently, the state can be considered strongly trapping.

We must still consider, however, the eventuality that, before becoming "trapped" in a horizontal state, sufficient responses should occur in each column so that the system goes into a diagonally symmetric state. Consider, for example, the second generation matrix $N = \begin{pmatrix}1&0\\0&1\end{pmatrix}$, for which we would have:

$$P = \begin{pmatrix}.5+\Delta&.5-\Delta\\.5-\Delta&.5+\Delta\end{pmatrix}$$

Note, however, that the K-matrix, in this case, takes the form $\begin{pmatrix} k & k \\ k & k \end{pmatrix}$, with a nonzero element in each cell, whereas in a second generation horizontal case, $K = \begin{pmatrix} k & 0 \\ k & 0 \end{pmatrix}$. Moreover, it is clear from equation (12) that each k_{ij} in the diagonal case will be less than the dominant k's in the horizontal case, since the magnitude of the k's tends to increase with the square of the column sums of the N matrix. Consequently, the row differences, $k_{i2} - k_{i1}$, which determine the p_{ij}, are bound to be greater for the horizontally symmetric cases than for the diagonally symmetric cases; in other words, the horizontal cases are "stronger", or more binding, than the diagonal cases, and the probability of an "atypical" response (one which goes counter to the current tendency of the P matrix) is greater for the diagonal than for the parallel case. This argument can clearly be extended for any generation matrix. Consequently, no matter how far the matrices progress towards a perfectly trapping condition, the horizontal cases will always be more strongly trapping than the diagonal cases.

It remains only to be shown that any diagonal state of the system can be transformed into a horizontally symmetric state, in order to prove the theorem. Consider the arbitrary diagonal state, $N = \begin{pmatrix} n' & n \\ n & n' \end{pmatrix}$, where the primed elements indicate large integers, and the unprimed letters indicate smaller integers. To each of the large elements corresponds some $p_{ij} < 1$, and to each of the small elements corresponds some $p_{ij} > 0$. Consequently, on the subsequent cycle, $(t + 1)$, there is a nonzero probability that $n \rightarrow n + 1$. This yields a new matrix, which is subject to the same argument. It is therefore clear that after $n' - n$ such events, either of the small elements (n) might grow to the magnitude of the large elements (n'), which is a sufficient condition for a horizontally symmetric state (since the n's of one row are now equal, and the n's of the other row unequal).

This completes the proof for Theorem 1. The behavior of such a Class C system can be most clearly represented by an equilibrium diagram *(fig.3)*, in which heavy arrows indicate strong tendencies, and light arrows indicate weak tendencies. In general, as time goes to infinity, all systems will eventually assume a horizontally symmetric form, with continually decreasing probability of escaping to either a neutral or a diagonal condition.

7. THE CLASS C' THEOREM

Theorem 1 indicates that the type of perceptron which had been considered at the outset would be incapable of "spontaneous organization". In fact, such a system, even if it has been deliberately taught to associate opposite responses to two classes of stimuli, is likely to degenerate to the terminal condition described by the theorem, if it is subsequently left on its own,

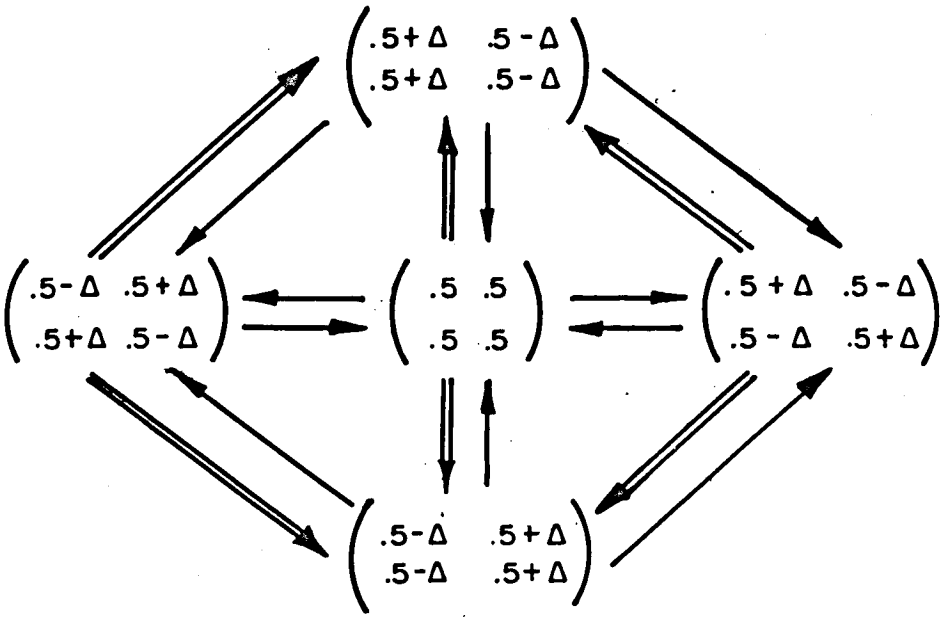


Fig.3. Equilibrium diagram for a Class C System
(States of the System are represented by P Matrices)

without further training or supervision by a human operator. This discovery led to a search for a system which would not be subject to the Class C effect, and which would tend, instead, to arrive at a "useful" division of its environment, without human intervention. The following existence theorem states the conditions with which we are specifically concerned:

THEOREM 2

There exists a class, C', of perceptrons, which always tend toward a statistically stationary state, such that for each binary response (R = 1, 0) the environment will be dichotomized into the two stimulus classes, S_1 (whose members evoke the response R = 1) and S_2 (whose members evoke the response R = 0). The dichotomy which is established will, in general, be characterized by a high degree of similarity between stimuli of the same class, and marked dissimilarity between stimuli of opposite classes. In the limit, each response of such a system contains 1 bit of information, with respect to the current stimulus.

Actually, there are many systems which meet the conditions of this theorem. The simplest of these, mathematically, is one in which the output signals of each source-set of A-units are divided by the current value of σ_ν, thus making the elements of the K-matrix (defined above) proportional to $r_{\alpha\nu}$. A system of much greater practical interest, however (which, moreover, converges to its terminal condition more rapidly), results from the growth functions defined by equations (2) and (3). As indicated in (2a) and (3a), the values of the A-units tend towards a limit which is proportional to the expected value of the input signal times the associated magnitude of the reinforcement operator, which is assumed here to be 1 or 0. Hence, we have the following corollary:

COROLLARY:

A perceptron in which the values of the A-units decay at a rate proportional to their current magnitude is a member of the Class C'.

With the proof of the Class C theorem as a model, we are now prepared to undertake the rather more subtle proof of the Class C' theorem. As before, it will be sufficient to prove the corollary, for any particular system, in order to prove the theorem. Actually, this proof involves two distinct points:

1. We will prove that the system always tends to dichotomize the environment into *some* two classes. To prove this, it is sufficient to show that, given *any* two stimulus classes, S_1 and S_2, the system will tend towards a diagonally symmetric terminal state in preference to a horizontally symmetric terminal state.
2. We must then prove that the dichotomy which is most strongly preferred by the system, in any environment, will correspond to our concept of "good similarity and dissimilarity", i.e., that it will tend toward a dichotomy which maximizes $r_{\alpha\bar{\alpha}}$ and minimizes $r_{\bar{\alpha}\bar{\alpha}}$.

The analysis will be carried out for a perceptron having the same organization as before, and characterized by the growth function stated in equation (3).

As in the case of the Class C perceptron, the state of the system will be described in terms of three matrices. The matrix P has precisely the same meaning as before, and can be classified, in the same manner, into horizontally and diagonally symmetric cases, and neutral cases. If we try to use the N matrix as a basis for calculating the state of the system, however, we will run into serious difficulties, for N is no longer Markovian. In the Class C perceptron, it makes no difference in what sequence the contributions to the elements n_{ij} occur; a given number of reinforcements in a specified category will always have the identical effect, regardless of the time at which they occur. In the Class C' system, however, this is no longer the case. The elements of the N matrix no longer unambiguously determine the P matrix, and consequently, the probability of each subsequent N matrix can

no longer be predicted as before. Fortunately, we can replace N with an alternative matrix, which is Markovian, and which we will call V. The V matrix is a 2 x 2 matrix with elements v_{ij}, where, as before, the row corresponds to the stimulus class, and the column to the response. The element v_{ij} can be interpreted as a measure of the residual value, at time t, which is due to the past reinforcements of the j source-set by stimuli of Class i. In accordance with equation (3), the elements of V will decay by a fraction δ, whenever the source-set j is reinforced (i.e., whenever the response $R = R_j$). The column sums of V correspond to the total values of the elements in the j source-set at time t due to all previous stimuli. It is clear that if the proportional decay, δ, is applied to each component of the sum individually (i.e., to the elements v_{1j} and v_{2j}) the net effect will be the same as if the decay had been applied to the column sums at each step, so that the analysis of the source-set value into two components does not introduce any error in the stochastic process. Specifically, the elements of each successive V matrix can be obtained by the rule:

$$v_{ij}(t) = \begin{cases} v_{ij}(t-1) & \text{if } R(t) \neq R_j \\ (1-\delta)v_{ij}(t-1) + 1 & \text{if } R(t) = R_j \end{cases} \qquad (14)$$

Now we have indicated that the terminal values of the A-units in a Class C' perceptron will be proportional to the expected values of α, measured over the set of stimuli to which the A-units in question have been exposed. The variance of the value, σ_v^2, will therefore tend towards a statistically stable magnitude. If we assume that the initial value distribution over the set of A-units has the same variance as this terminal distribution, then we would actually not expect the variance to change at all, as a result of reinforcement of the A-units; in effect, we will be changing the *correlations* of the A-unit values without changing the moments of the distribution. Consequently, the covariance matrix, K, which was used in the preceding analysis, can be replaced with the correlation matrix R, which consists of the coefficients $r_{ij} \equiv r_{\alpha_i v_j}$. These correlation coefficients are clearly equal to the covariance which can now be obtained in terms of the V matrix just as it was before from the N matrix) divided by σ_v. We now require an explicit expression for σ_v. Specifically, the variance σ_v^2 can be divided into five components as follows:

$\Delta_0 \equiv$ residual of initial noise component, present at time $t = 0$.
$\Delta_1 \equiv$ variance correlated with $\bar{\alpha}_1$ but not with $\bar{\alpha}_2$.
$\Delta_2 \equiv$ variance correlated with $\bar{\alpha}_2$ but not with $\bar{\alpha}_1$.
$\Delta_3 \equiv$ variance which is correlated with both $\bar{\alpha}_1$ and $\bar{\alpha}_2$.
$\Delta_4 \equiv$ variance which is "unique" to the individual stimulus (corresponding to Δ_1 in the previous analysis)

The magnitude of these components is given by the expressions:

$$\Delta_0 = \frac{\bar{\nu}_0^2}{\bar{\alpha}^2}(1-\delta)^{N_j}$$

$$\Delta_1 = \nu_{1j}^2\, r_{\alpha_1\bar{\alpha}_1}^2 (1 - r_{\bar{\alpha}\bar{\alpha}}^2) + \nu_{2j}^2\, r_{\alpha_1\bar{\alpha}_2}^2 (1 - r_{\bar{\alpha}\bar{\alpha}}^2)$$

$$\Delta_2 = \nu_{2j}^2\, r_{\alpha_2\bar{\alpha}_2}^2 (1 - r_{\bar{\alpha}\bar{\alpha}}^2) + \nu_{1j}^2\, r_{\alpha_2\bar{\alpha}_1}^2 (1 - r_{\bar{\alpha}\bar{\alpha}}^2)$$

$$\Delta_3 \equiv r_{\bar{\alpha}\bar{\alpha}}^2 \left[\nu_{1j}\, r_{\alpha_1\bar{\alpha}_1} + \nu_{2j}\, r_{\alpha_2\bar{\alpha}_2}\right]^2$$

$$\Delta_4 = \nu_{1j}\,(1 - r^2_{\alpha_1\bar{\alpha}_1}) + \nu_{2j}\,(1 - r^2_{\alpha_2\bar{\alpha}_2})$$

where $\bar{\nu}_0$ is the magnitude of the initial mean value, at $t = 0$, and N_j is the number of times that the response $R = R_j$ has occurred. From these components, we can readily obtain the approximation:

$$r_{\alpha_i\,\nu_j} \approx r_{\nu_j\bar{\alpha}_i}\, r_{\alpha_i\bar{\alpha}_i} = r_{\alpha_i\bar{\alpha}_i}\sqrt{\frac{\Delta_i + \Delta_3}{\Sigma\Delta}} \tag{15}$$

In terms of these r's, it is easy to approximate the elements of the P matrix, through the use of Fisher's z-transformation:

$$\begin{cases} p_{i1} = \phi\left(\dfrac{z_{i1} - z_{i2}}{\sqrt{\dfrac{2}{N_{A_r} - 3}}}\right) \\ p_{i2} = 1 - p_{i1} \end{cases}$$

where $$z_{ij} = \frac{1}{2}\ln(1 + r_{\alpha_i\nu_j}) - \frac{1}{2}\ln(1 - r_{\alpha_i\nu_j})$$

$\phi(x)$ = normal probability integral from $-\infty$ to x

and N_{A_r} = number of A-units in a source-set.

We must now define the concept of a "saturated state" for a C' system. Such a system will be called "saturated" when the sum of the squares for each column of the R matrix is maximum, i.e., when

$$\left(\frac{r_{\alpha_1\nu_j}}{r_{\alpha_1\bar{\alpha}_1}}\right)^2 + \left(\frac{r_{\alpha_2\nu_j}}{r_{\alpha_2\bar{\alpha}_2}}\right)^2 = 1 + r_{\bar{\alpha}\bar{\alpha}}^2 \tag{17}$$

This expression represents the upper limit for the correlations, as $r_{\nu\bar{\alpha}} \to 1$ for the two elements of the column, taken jointly.

Now, it is clear that every increment to the values of the A-units will bring the system closer to this saturated condition. Moreover, *the degree of saturation can only increase;* once a system has become saturated, it can never become unsaturated, since whichever stimulus occurs, and whatever the sequence of events, every new increment will only increase the tendency of $r_{\nu\bar{\alpha}}$ to approach unity. If we now examine the saturated state, it is clear that in the limit a horizontally symmetric system will become neutral as the two columns saturate, while a diagonally symmetric system will tend to remain diagonally symmetric. Thus, the only acceptable terminal conditions, for a Class C' system, are either neutral or diagonally symmetric. In the neutral terminal state, each of the r_{ij} elements is equal to $r_{\alpha\bar{\alpha}}\sqrt{\frac{1+3r^2_{\bar{\alpha}\alpha}}{2+2r^2_{\bar{\alpha}\alpha}}}$, while in the diagonal terminal state, the strong elements go to $r_{\alpha_i\bar{\alpha}_i}$ and the weak elements go to $r_{\alpha_j\bar{\alpha}_i}$. Consequently, the diagonal terminal states tend to be "trapping", and the neutral states have a low probability. A Class C system can now be recognized as a limiting case of the Class C' system, in which the time for saturation goes to infinity, due to the fact that σ_ν never approaches a limit.

This completes the first part of our proof, i.e., we have shown that, given any two classes, S_1 and S_2, the perceptron will prefer a diagonal terminal state to a horizontally symmetric state. The equilibrium diagram which characterizes this type of system is shown in *fig.4*. It is now necessary to show that the particular dichotomy which is formed in an arbitrary environment will tend to be one which corresponds to our criterion of similarity. In order to do this it will be sufficient to show that the net terminal error probability (considering as an "error" any response which is contrary to the bias of the terminal P matrix) will be minimum for a dichotomy which meets our similarity criteria. If the system terminates in a state in which the error probabilities are not minimum, then it will always tend (given infinite time) towards a state in which the error probabilities are still smaller, i.e., it will tend towards a more "strongly trapping" terminal condition. Let us, therefore, examine the various possible dichotomies, in an arbitrary environment, from the standpoint of their error probabilities.

Consider, first, an environment consisting only of two stimuli (not two *classes* of stimuli), S_1 and S_2. Under these conditions, $r_{\alpha\bar{\alpha}}$ for each stimulus will be equal to 1, and $r_{\bar{\alpha}\alpha}$ is equal to $r_{\alpha_1\alpha_2}$. Here there is only one dichotomy possible, so there is no doubt about which will be selected. As long as the two stimuli do not produce identical signals, (i.e., as long as $r_{\alpha_1\alpha_2} \neq 1$), the perceptron should tend to give a 1-response for one stimulus,

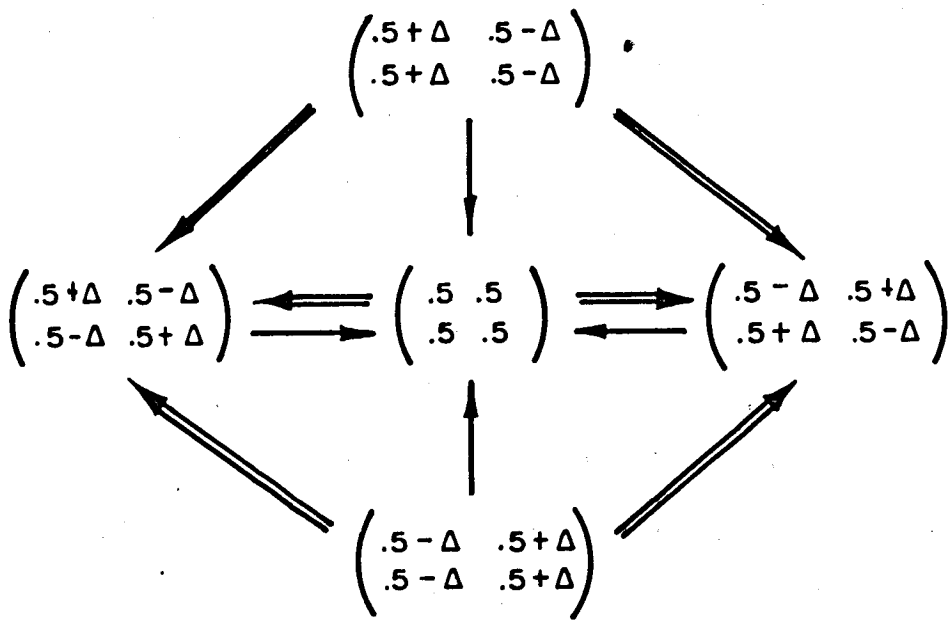


Fig.4. Equilibrium diagram for a Class C' System

and a O-response for the other. Now, if there are more than two stimuli, we can represent the state of the system unambiguously, without presupposing any particular division of the environment into classes, by adding a row to the V, R, and P matrices for each additional stimulus. Thus, for n possible stimuli, the P matrix becomes:

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \\ \vdots & \vdots \\ p_{n1} & p_{n2} \end{pmatrix}$$

This matrix can be classified as horizontally symmetric if all elements of one column are less than 0.5, and all elements of the other column are greater than 0.5. If for any row, $p_{i1} > p_{i2}$, and for any other row, $\rho_{i1} < \rho_{i2}$, the matrix is diagonal. All of the development which previously applied to the calculation of the P matrix from V and R can similarly be

restated in terms of this expanded matrix. Equation (16) still applies to the calculation of the p's, equation (14) to the ν's, and a suitable extension of equation (15) (in which the variance contributions from each element in the appropriate column of the V matrix are included) permits us to calculate the r's.

Now, the total error probability for such a system at time t will be

$$P_\epsilon = \sum_i g_i P(S_i) \tag{18}$$

where g_i is the "weak" probability (i.e., the probability which is less than 0.5) from the i row of the P matrix, and $P(S_1)$ is the probability that the particular stimulus, S_i, will occur. It is clear from the general dynamics of the Class C' systems, as discussed above, that the system will tend to keep changing until it arrives in a maximally stable terminal condition, and this maximally stable condition will be one in which P_ϵ, as defined by equation (18), is minimized.

Now consider any class, S, of stimuli. The probability, $P(S)$, that some member of this class will occur, is equal to the sum of the probabilities, $P(S_i)$, that the individual members of that class will occur. Consequently, if the system is in state P at time t, and if we define the two stimulus *classes*, S_1 and S_2, so that S_1 is the class of all stimuli for which the response $R = 1$ is preferred at time t, and S_2 is the class of all stimuli for which $R = 0$ is the preferred response, we can write [from equation (16)]:

$$P_\epsilon(t) = g_1 P(S_1) + g_2 P(S_2) \tag{19}$$

But we can write explicit expressions for g_1 and g_2 (the weak diagonal elements of a terminal P matrix) in terms of equation (16), i.e.,

$$g_i = 1 - \phi\left(\frac{Z(r_{\alpha_i\bar{\alpha}_i}) - Z(r_{\alpha_j\bar{\alpha}_i})}{\sqrt{2/N_{Ar} - 3}}\right)$$

where the z's are defined as for equation (16). Consequently, we have

$$P_\epsilon = \sum_{i=1,2}\left[1 - \phi\left(\frac{Z(r_{\alpha_i\bar{\alpha}_i}) - Z(r_{\alpha_j\bar{\alpha}_i})}{\sqrt{2/N_{Ar} - 3}}\right)\right] - P(S_i) \tag{20}$$

which is the function which must be minimized in order to find the most stable classification system for the environment in question. An examination of this expression shows the following:

1. The terminal condition depends on the frequency of stimuli in the two classes; a highly coherent class which is very small, or the members of which are all highly unlikely to occur, will be less likely to occur (other things being equal) than a larger class, or a class which includes more frequent stimuli. This means that the system will tend to favor dichotomies which divide the environment evenly between classes.

2. Other things being equal, the system will favor dichotomies such that the intraclass correlations, $r_{\alpha\bar{\alpha}}$, are large, and interclass correlations, $r_{\overline{\alpha\alpha}}$, are small. This is the similarity condition which we set out to establish.

Since we have shown that the error function, [equation (20)], will be minimum under the conditions which best satisfy our similarity criteria, we have effectively proven Theorem 2. The predictions of this theorem, as well as those of Theorem 1, have been successfully demonstrated using the IBM 704 computer to simulate the performance of a perceptron in a simple perceptual environment. In the first experiment to be successfully completed, a 500 A-unit perceptron spontaneously learned to distinguish the class of squares on the left from the class of squares on the right, in an environment in which squares were allowed to appear in random positions anywhere in the right or left halves of a visual field, being excluded only from positions in which they would overlap the center line. After being exposed to 100 squares chosen at random from this environment, the perceptron exhibited the P matrix (based on 100 test-stimuli of each class):

$$P = \begin{pmatrix} 0 & 1.00 \\ .94 & .06 \end{pmatrix}$$

in which the net error probability is only 0.03.

8. SIGNIFICANCE OF THE CLASS C' PERCEPTRON

In discussing the Class C' system, I have taken pains to avoid any reference to the concept of *entropy*. A number of theorists have felt that a system which is self-organizing, in the sense that the Class C' perceptron appears to be, is in contradiction to the second law of thermodynamics. Actually, of course, if we concern ourselves simply with the physical state of the system, entropy is clearly increased, as the perceptron is an energy consuming device. From the standpoint of *information*, on the other hand, we may still ask the question whether the total amount of information has somehow been increased as a result of the perceptron's organizing process.

Consider an environment of eight stimuli, $S_1 \ldots\ldots S_8$. At the outset, the response $R = 1$ might denote the presence of any of the stimuli S_1, S_3, S_5, or S_7, while in its terminal condition, $R = 1$ might indicate the presence of S_1, S_2, S_3, or S_4. Can we legitimately say that the information given by the response $R = 1$ in the terminal state is greater than the information given in the initial state? In any absolute sense, probably not. On the other hand, if we recognize that the first four stimuli are all triangles, while $S_5 \ldots S_8$ are all circles, we can legitimately say that the information carried by the terminal response is more meaningful than the information carried by the initial response. "Meaningful", in this sense, means that, while $R = 1$ conveys one bit of information in either case, the information

in the terminal condition has become correlated with certain aspects of the environment which we wish to have reported, whereas in the original state, the information conveyed by $R = 1$ was of no interest or utility.

This argument is, admittedly, a sketchy one, and may contain loopholes. In any case, the fact that such questions arise (and the difficulty in answering them convincingly), seems to indicate a certain ambiguity in our concept of "information". The fact that "information" is not equivalent to "meaning" has often been noted before. The performance of the Class C' perceptron seems to make this distinction even more apparent.

Be that as it may, it seems clear that the Class C' perceptron introduces a new kind of information-processing automaton: for the first time, we have a machine which is capable of having original ideas. As an analog of the biological brain, the perceptron, or, more precisely, the theory of statistical separability, seems to come closer to meeting the requirements of a functional explanation of the nervous system than any system previously proposed. The neuroeconomy, and the number of necessary constraints required in order to specify a system of this type, seems to be well within the bounds of biological plausibility. So far as we have been able to ascertain, no known facts about the central nervous system have been violated by our assumptions, even though at times we have used simplification and short cuts which are not available in a biological organism. Wherever we have made such assumptions, as in the simplification of a "response unit" to a single two-stage element, the biological equivalent can be clearly indicated. Thus, while we have not proven the validity of statistical separability as the explanation of an organic brain, this explanation now seems to be the most plausible of the available alternatives.

As a machine, the future of the perceptron will depend heavily upon our ability to achieve an efficient and inexpensive A-unit. Studies of components and circuitry are currently in progress, and success seems imminent. We will not speculate here upon the possible applications of such a device. The perceptron is in its infancy, and it would be a mistake to rush it too abruptly towards an adolescence which can still scarcely be foreseen.

As a concept, it would seem that the perceptron has established, beyond doubt, the feasibility in principle of nonhuman systems which may embody human cognitive functions at a level far beyond that which can be achieved through present day automata. The future of information processing devices which operate upon statistical, rather than logical, principles, seems to be clearly indicated.

REFERENCES

1. ASHBY, W. ROSS, Design for a brain, *Wiley & Sons, New York,* (1954).
2. McCULLOCH, W. S., and PITTS, W., "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Math. Biophysics,* 1943, 5, p. 115.
3. ROSENBLATT, FRANK, The Perceptron - A theory of statistical separability in cognitive systems, *Cornell Aeronautical Laboratory, Inc. Report No. VG-1196-G-1,* January 1958.
4. VON NEUMANN, JOHN, The computer and the brain, *Yale University Press, New Haven,* (1958).

APPENDIX TO THE PAPER BY DR. F. ROSENBLATT ∅

Since the draft of the preceding paper was written, a much improved method of analysis has been developed, which should really supersede much of the foregoing material. This method is applicable to networks of "conventional" on-or-off neurones, rather than to neurones of the continuous transducer variety. Each A-unit, a_i, is characterized by a threshold, θ, and a variable output strength, v_i, which serves as the memory variable. If the algebraic sum of the excitatory and inhibitory input signals from the retina is greater than the threshold, an A-unit "fires", and delivers its output, v, to the R-unit to which it is connected. If the measure of the total input signal to the A-unit is α, then we can define the "activity function", $\alpha*$, as

$$\alpha* = \begin{cases} 1 \text{ if } \alpha \geq \theta \\ 0 \text{ if } \alpha < \theta \end{cases}$$

The output signal from every A-unit will thus be equal to $\alpha* v$, and we will assume that the R-unit is "activated" by a net input signal greater than zero, and "suppressed" by a net input signal less than zero. Specifically,

$$R = \begin{cases} 1 \text{ if } \sum_i \alpha_i* v_i \geq 0 \\ 0 \text{ if } \sum_i \alpha_i* v_i < 0 \end{cases}$$

where the summation is over all A-units, a_i.

Now let P_{a_i} = the proportion of A-units responding to stimulus S_i, and let $P_{a_{ij}}$ = the proportion of A-units responding *both* to stimulus S_i and to S_j. The expected values of these proportions are known functions, which have been previously described *(ref. 1)*. While a small perceptron may show appreciable deviations from these expected values, in a very large perceptron, the proportions of responding units should be very close to the expected values. In order to eliminate from consideration the variability of different perceptrons, the following remarks will be restricted to the assumption of a very large, or infinite, perceptron, in a finite universe of stimuli, S_1, S_2, ..., S_n.

If a perceptron is exposed to some stimulus, S_i, and its association system is reinforced, then the total amount of value gained by the set of A-units responding to S_i will be some function of P_{a_i}, which depends on

∅ Added after the Symposium

the particular rules of reinforcement which are applied. For example, if we simply have a rule that the active units gain some increment of value, Δv, and inactive units are unaffected then the total amount of value gained by the set of A-units responding to S_i will be kP_{a_i}, where k is a constant of proportionality. Such a system is called an "alpha system". Now suppose we present stimulus S_i, and apply the above rules for reinforcement. Each of the A-units thus reinforced is likely to respond not only to S_i, but to one or more other stimuli as well. Let us now examine the effect of the reinforcement of S_i upon the sets of A-units responding to each of the other possible stimuli. The measure of the value change in the set of units responding to S_j, as a result of having reinforced S_i, will be called the "generalization coefficient", g_{ij}. In the alpha system, considered above, g_{ij} will clearly be equal to $kP_{a_{ij}}$. If we arbitrarily define our units of reinforcement so that $k = 1$, then we have for the alpha system that $g_{ij} = P_{a_{ij}}$.

In a different system, which we have called a "gamma system", the reinforcement rules specify that the total value, measured over the entire set of the A-units, must remain equal to zero. In this case, if the active units gain a quantity equal to P_{a_i}, then this same quantity is subtracted uniformly over the association set as a whole, so that, again assuming the proportionality constant, k, to be equal to 1, we have for the gamma system: $g_{ij} = P_{a_{ij}} - P_{a_i} P_{a_j}$.

Now if there are exactly n possible stimuli, there will be an n by n matrix of generalization coefficients, g_{ij}. The *rows* of this matrix represent all of the "contributions" from a reinforcement of S_i to the sets of A-units responding to each possible stimulus, S_j (where S_j may be identical with S_i). The *columns* represent all of the possible contributions to the set of A-units responding to S_j, from reinforcements of each possible contributing stimulus, including itself. Let us designate the row vectors of this matrix G_i, and the column vectors G_j^*. Let us represent the expected frequency of occurrence of each of the n stimuli by a positive scalar number, f_i. The set of n frequencies, for the n stimuli, can be represented by a vector, F. The scalar product, FG_j^*, will then be equal to the expected rate of change for the total value of the set of A-units responding to S_j, as a result of being shown each of the possible stimuli at its expected frequency, each one being positively reinforced. Thus, over a period of time, t, assuming a zero decay rate, we would expect the value measured over this set of A-units to grow to tFG_j^*.

Let us now consider the effect of the reinforcement operator, ρ, which is set equal to +1 if the response $R = 1$ is to be "positively reinforced",

and -1 if the response is "negatively reinforced", for a given stimulus. Assume that this reinforcement is applied consistently for each stimulus, i.e., whenever stimulus S_i occurs, its responding A-set is either positively reinforced or else negatively reinforced on each occasion. Then the generalization coefficient, g_{ij}, must be multiplied by the sign of ρ, and the column vector now consists of elements $\rho_i g_{ij}$. If we sum the elements of each such column vector, multiplying each element by its appropriate frequency, f_i, then the array of column sums forms the vector:

$$V = \sum_i f_i \rho_i G_i = (v_1, v_2, \text{---} v_n)$$

Each element of this vector, v_j, represents the expected rate of change of the value of the A-set responding to S_j, as a result of exposure to a random sequence of stimuli from the universe in question. If v_j is positive, the stimulus S_j will tend to turn on the response $R = 1$, while if v_j is negative, it will induce the response $R = 0$.

Let us first consider the simple alpha system, with no decay. The generalization coefficients for this system, g_{ij}, are simply equal to $P_{a_{ij}}$, and consequently are all positive. Suppose S_i occurs, at time t_1. This immediately gives us the value vector, $V(t_1) = G_i$. Since every element of this vector is positive, it follows that whichever one of the n possible stimuli occurs at time t_2, the response $R = 1$ will occur. But the occurrence of this response means that we will again add an all-positive vector to V, and, in fact, it is clear that no element of V can ever become negative. Consequently, this system behaves as a "Class C perceptron", in accordance with our first theorem.

Now consider the gamma system, for which $g_{ij} = P_{a_{ij}} - P_{a_i} P_{a_j}$. It can be shown that for a perceptron in which the A-units are randomly connected to the retina, and each A-unit receives a fixed number of excitatory and a fixed number of inhibitory input connections, it will always be true that

$$P_{a_{ij}} = P_{a_i} P_{a_j}$$

provided that the area of the intersection of the two stimuli is equal to the product of their normalized areas (i.e., $C_{ij} = R_i R_j$, where C_{ij} is the common area, and R_i and R_j are the retinal areas of the stimuli S_i and S_j, with the area of the retina taken as unity). If stimuli can occur with equal probability in any retinal position, then their expected intersection will indeed be equal to this value of C_{ij}. If all intersections are equal to their expected value, the corresponding value of g_{ij} will be zero. If, however, two stimuli have a common area, C_{ij}, greater than this expected value, g_{ij} will be > 0, and if the stimuli are disjunct, or have a common area less than the expected value of C_{ij}, it follows that g_{ij}

will be negative. Even though the expected value of C_{ij} for a class of randomly placed stimuli is equal to zero, it can be shown that, due to nonlinear characteristics of the $P_{a_{ij}}$ function, the expected value of g_{ij} over the class of stimuli may be greater than zero. This is actually a necessary condition to permit a response to generalize consistently over the class of stimuli.

To analyze the simplest possible case, consider a stimulus universe consisting of two disjunct stimuli S_1 and S_2. From what has just been said, the generalization coefficients g_{11} and g_{22} will clearly be positive, and g_{12} and g_{21} will be negative. Consequently, each stimulus will generalize positively to itself, and negatively to the other stimulus, and the perceptron will necessarily behave in the fashion of a Class C' system, described in Theorem 2. It should be noted, however, that no decay has been postulated in this system, and none is necessary. Even though the magnitude of the values of the A-units, and consequently the variance of the value, will grow without bound, the system remains well-behaved, so that the Corollary of Theorem 1 is disproved, at least for this special case.

Due to the fact that the expected value of g_{ij} for stimuli of opposite classes, in an environment in which every placement of the stimulus on the retina is equally probable, is equal to zero, it is desirable in some cases to add a constant loss rate, ϵ, to the value dynamics. In this case, every A-unit always loses a slight decrement of value, in addition to any other changes upon which the loss may be superimposed. This gives us a modified gamma system, with the generalization equation:

$$g_{ij} = P_{a_{ij}} - P_{a_i} P_{a_j} - \epsilon$$

If a decay rate, δ, is also incorporated in the system, it is clear that in the absence of any reinforcements, the system will stabilize with all units at a level in which the values are proportional to $-\frac{\epsilon}{S}$, at which point the decay rate (which is now positive) will exactly balance the rate of loss. This system has the advantage of adding a slight negative interaction between stimulus classes, which otherwise might develop a positive relationship. Again, due to the nonlinearity of the $P_{a_{ij}}$ function, this negative interaction tends to affect "well separated classes" more strongly than classes of "similar" stimuli, and thus generally helps in establishing a desirable separation.

The performance of an infinite perceptron, in which g_{ij} is obtained from the above formula, is illustrated in *fig. 1*. The reinforcement operator, ρ, was set equal to +1 for $R = 1$, and to 0 for $R = 0$, i.e., the system was reinforced only for $R = 1$, and was left unchanged for $R = 0$. In this experiment, two classes of stimuli were assumed: one class consists of horizontal

SPONTANEOUS ORGANIZATION OF INFINITE PERCEPTRON IN ENVIRONMENT OF 4 x 20 VERTICAL & HORIZONTAL BARS

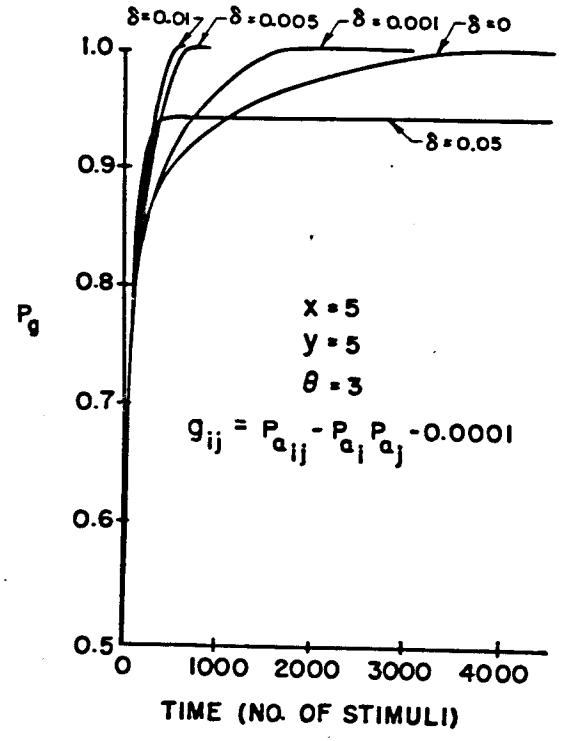


Fig.1

EXPECTED WAITING TIME TO PERFECT PERFORMANCE, AS A FUNCTION OF DECAY RATE (MEANS OF 10 RUNS)

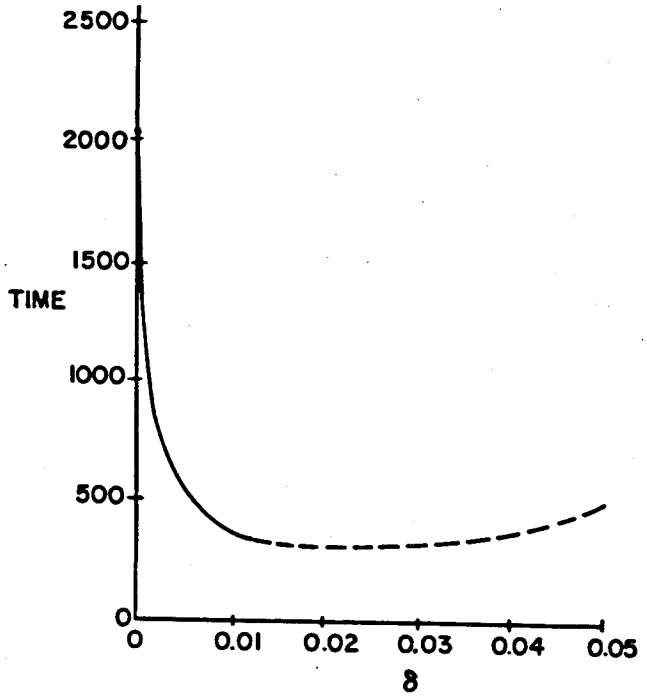


Fig.2

bars, covering an area of 4 x 20 retinal units, and the other class consists of vertical bars of the same area. The retina is a 20 x 20 mosaic, and is assumed to have a torroidal connectivity, i.e., if a stimulus image is shifted off of one edge of the retinal space, it re-enters at the opposite edge. This guarantees a uniform coverage of all retinal points by each class of stimuli. The first set of curves shows the probability of correct (i.e., consistent) generalization of the response $R = 1$ to one class of stimuli, and $R = 0$ to members of the other class, for different values of the decay rate, δ. Note that even with a zero decay rate, the system eventually discriminates perfectly between the two classes. As the decay rate increases, performance gradually improves, up to a point where, with $\delta > 0.01$, the perceptron begins to "forget" too rapidly, and performance becomes unstable. This condition is also reflected in *fig. 2*, which shows the expected waiting time to perfect performance, as a function of δ.

REFERENCE

1. ROSENBLATT, F., The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain; *Psychol. Review*, **65**, 1958, 386.

DISCUSSION ON THE PAPER BY DR. F. ROSENBLATT

MR. E. A. NEWMAN: Several years ago Mr. Day, of the N.P.L., and I thought up a device which I will say a little about. This assumed a matrix of photocells as a sensing network, random connected to an association network which in turn was connected to a response unit. There was feedback between the input and output of the association network via the response unit, arranged to reinforce association cell activity by reducing gate thresholds should the response to a stimulus be the desired one. Since in England we have not the large amounts of programming effort available to our friends abroad we never programmed the device.

We would have liked our device to possess properties, which, as far as we could see it did not have. The first is that if presented with a slowly changing shape it would infer that all the shapes covered by the transitions were related, the second was that if presented with a set of shapes all varying but slightly from each other, it would infer they were related, even if other shapes were perceived in between. What it did do was to relate in a random way some items out of a set of coincident ones, afterwards putting special stress on this accidental selection, and to be more likely to relate inputs which followed in time but had no continuity of spatial pattern than those that had. This I suggest is not quite what one wants to do.

Most patterns that interest us contain a great deal of special connectivity and continuity. The information needed to specify that a pattern is continuous, and not a snow storm is very large indeed. In our system information about spatial order was thrown away in the radom connection to the association network and had to be learnt. An association network which was connected to the photocell matrix in a spatially ordered manner would be better in this respect.

Our device took note of too much information. The only interesting part of a pattern is its edges, and these only if they alter in time. What we would have liked our device to do was to react to changes in pattern in space time. Apart from any other aspect of the matter, a device which takes note of other than space time changes rapidly uses up all its storage.

Our device reached different end states for different kinds of pattern, but these end states were patterns of dots which had to be recognised. There is not much point in knowing that a device gives a 1:1 correspondence between input and output state, if one cannot recognise the output states. I think the device has at least something in common with the perceptron. Would Dr. Rosenblatt explain how this latter device overcomes the limitations given above.

DR. SHUEY: I have four questions. I would like to say, first, that I think Dr. Rosenblatt's work has contributed a great deal to the analysis of networks of the neural type. My questions do not refer to the new material Dr. Rosenblatt has just presented, but to the material in the printed paper and other published material which people in both England and the United States have seen.

First, some of the initial reports made a great point of random distributions and probability. I have a question as to whether randomness and probability are philosophical essentials or merely a very significant feature of this particular analysis. I do not wish to de-emphasize the importance of analysis. I am merely raising the point as to whether probability is a philosophical necessity in networks of this type.

Second, it seems to me that the detailed connectivity either in a precise fashion, or in a statistical fashion, should be determined by the class of objects or environment that you want the device to recognize. I think I can illustrate my point quite simply if I take a pattern recognition system with three levels. Let the first level contain N binary receptors. Let the second level of binary elements have one element for each of the 2^N input states. Let the third level of binary elements have one element for each possible association that can be formed between input states; there are 2^{2^N} such associations. A specific input can belong to many associations. In other words, given a set of 2^N points, there are 2^{2^N} subsets that I can form from these points. 2^{2^N} soon becomes a very large number as N is increased and it is, of course, even larger in a non-binary system. I do not believe that you want a system that will be able to make all possible associations. I think the possible associations should be determined by what you want the machine to do.

It seems to me that, if you assign the original connections at random, you are in a sense going to make a machine that is equally sensitive to all classes of patterns. I do not believe you want to do this any more than you as a person want to be able to differentiate snowstorms on a TV screen. The differentiation of individual snowflake patterns is in general of no interest to you.

I think one has to place some constraint on the connection. In the perceptron, you do not have all possible connections. You can change the gain of the connections, but you cannot change the connections. If with a given number of elements and connections, I am allowed to change the connections, I shall have a much more flexible system. There is a parallel in the nervous systems of animals. If the connections of dendrites and axons are in part determined by environment, the biological machine will be more flexible than if environment can determine only the strength of those connections. I understand, although I am not a physiologist, that this is an unanswered question as far as the cortex is concerned.

In the example I took earlier, the 2^{2^N} possible associations is deceiving. This is not necessarily the number of elements needed to form that number of associations. The associations might be formed by the state of another system. In a binary system with N_1 elements, it is only required that $2^{N_1} \geq 2^{2^N}$. Possibly, the perceptron does this type of thing.

My third question is related to one of Mr. Newman's points. In the experiments which have been done, if you continuously distort one object into another object, does the probability of making a correct selection smoothly go over or does it jump as a step function?

Fourth, can you make any statement relating how the perceptron discriminates to how we as individuals discriminate? I realize that this is a speculative field.

DR. A. M. UTTLEY: I want to say something about this 2^n.

We know that a set of n binary inputs has 2^n logical states, 2^{2^n} logical functions of those functions, and so on indefinitely. But each of our "outputs" to the external world provides a set of proprioceptive "inputs" to the analyser, in fact, some of the 2^n inputs and *not* one of the 2^{2^n} logical functions which, anyway, is just 0 or 1 depending on the input state.

Consider now the problems of a computer learning to produce an output O_1 which is a logical function of n inputs, I_1, ... I_n. Let there be an input S_1 to indicate success in choosing O_1 at random. Also let the random choice of O_1 be fed back as an input to the computer. We now have a machine with $(n + 2)$ inputs and 2^{n+2} units which can solve the problem. If the machine has, in addition, to learn a second output O_2 which is another logical function of the inputs then there must be 2^{n+4} units, and so on.

MR. STAFFORD BEER: I should like to address my remarks to the written paper, rather than to any of the chimerical perceptrons which have been floating about outside it.

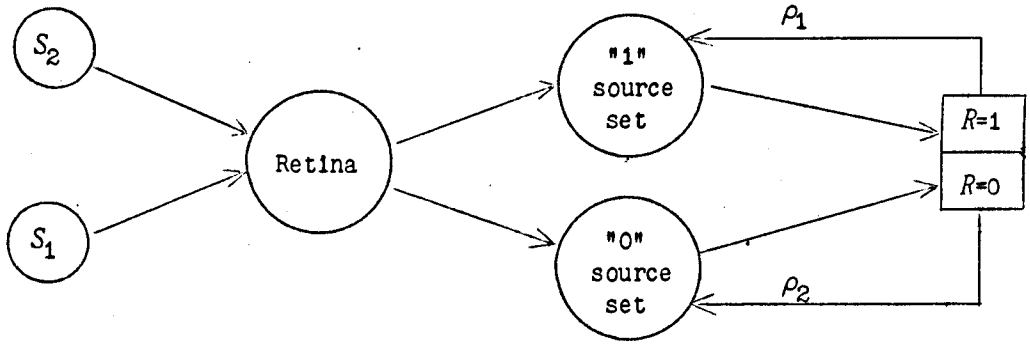
My criticism begins in the first section where Dr. Rosenblatt says "symbolic logic or Boolean algebra", and later on the same page "Boolean algebra or symbolic logic". That is, he presumably takes these terms as synonymous whereas they are not. Surely Boolean algebra is a small sub-set of the class of symbolic logics. There is a whole range of these, most of which are non-Boolean. For example, there is the predicate calculus.

The confusion the author has made leads him to a false dichotomy between Boolean algebra and statistical mathematics as tools. This excludes the most useful semantic tools at our disposal. For example, the kind of logic used by Ashby in this work is an algebra of classes, the precise virtue of which is that the classes need only be specified as bounded in dimensions that are of interest, and that their behaviour is entirely probabilistic. This gives us the virtues of statistical methods, without the lack of logical distinctness which seems to characterize this paper. I would

concede that this might be just a question of terminology (if it is, my point is merely captious) were it not for the disastrous consequences which seem to follow.

For example, an A-unit (one of the association cells) is conceived as a set of ν - values. These may be, it seems to me, *either* a point-set covering the neurons in the A phase-space, *or* a set of values through time taken up by one neuron, *or* again the values taken up over time by the phase averages. By spurning the kind of logic which would distinguish between these possibilities, Dr. Rosenblatt has left himself open to a charge of circularity. For he proposes two theorems which clearly purport to demonstrate ergodic properties in the perceptron, whereas the ergodicity is probably subsumed in the logically loose definitions of the ν - set from which he starts.

From that example, I shall now go on to generalize the charge of circularity. In Theorem 2, Dr. Rosenblatt distinguishes dichotomously between two classes of stimuli. I shall amend his diagram by doing the same. All stimuli must belong to one class or the other. This is an arbitrary division of the set of all stimuli, but one which I am entitled to make; note that I say nothing of the criteria by which a particular stimulus is identified as belonging to S_1 or S_2. I say only that any stimulus can be regarded as belonging to one of two classes and as feeding into the retina thus:-



This does *not* say what effect these stimuli have on the system.

The possibility is given in this system that events take place which cause the response to go to "1". Therefore we can collate these events, and assert that a class C of them exists if and only if the responses go to "1":-

$$\exists\ !\ C \equiv R \rightarrow 1$$

Equally, there exists another class of events if and only if the responses go to "0". We shall extend class C to include these too, and assert the existence of the class in which events take the response either to "1" or to "0", but not both:-

$$\exists\ !\ C \equiv R \rightarrow 1 \wedge 0$$

What information is given in the system about the role of the stimuli? In the system as shown there is nothing to show what kind of "switch" there may be across the retina; there is nothing to show that a member of either stimulus class will influence an "1" source set more than a "0" source set. So either kind of stimulus is capable of evoking either response. The account of class C may therefore be extended to:-

$$\exists\,!\ C \equiv S_1 \vee S_2\ (R \rightarrow 1 \wedge 0) \qquad \ldots\ldots \quad (1)$$

Since the system is no more restricted than this, a number of logically weaker proportions can be asserted. The one which interests me is this:-

$$\exists\,!\ C' \equiv S_1\ (R \rightarrow 1)\ S_2\ (R \rightarrow 0) \qquad \ldots\ldots \quad (2)$$

That is to say that since the first kind of stimulus is capable (expression 1) of evoking the unit response, then the cases of this happening can be collected as a sub-set. Similarly, since the second kind of stimulus is capable of evoking the zero response, these too can be collected as a sub-set. There is no reason why these two sub-sets should not be considered together as the class C′ (expression 2).

These two classes, C and C′, certainly exist in this system, although the second could be empty, and although there are many others. What do we know of them? For example, how much information can they each transmit? Class C transmits zero bits about a specific stimulus, because it may evoke either response. Class C′ transmits one bit about a specific stimulus, because that stimulus can only evoke *one* of the responses.

Looking at the diagram, how can this happen? What mechanism must be inferred at the retina to account for this information flow? If zero bits are transmitted from either S_1 or S_2, the retina must be acting as a kind of random switch. It is in fact uncoupling the stimulus system (S) from the response system (R). If one bit is transmitted, on the other hand, then in some sense S and R are coupled together. This coupling cannot be asserted categorically on the evidence; we can however say that a tendency must exist at the retina to couple (S_1R_1) and (S_2R_0), otherwise the system would not transmit.

Now this inference can be quantified. The conditional probability of the response given the stimulus must, for the class C, tend to 0.5; because there is nothing controlling the "switch". But for the class C′, the conditional probability of the response must tend to 1.0. That is:-

$$\exists\,!\ C : Pr\ (R|S) \rightarrow 0.5$$

$$\exists\,!\ C' : Pr\ (R|S) \rightarrow 1.0$$

The nature of the logical correspondence between S and R follows from this. In the first case, C, it is strictly unknown. But is is important to recognize that it *could* be a "one-many" correspondence. In the second case,

C′, however, it must tend to be a "one-one" correspondence, because the variety in the right-hand part of the system must now be limited by a function of the input α.

What limitations do these inferences about logical correspondence produce in the A-set? In the first case, C, the variance of the value of ν can very well increase. Since, as was shown, the *S-R* correspondence could be "one-many", or at least might include "one-many" transformations, this variance can go on growing indefinitely. In the C′ case, however, the variance of the value of ν must tend to a limit, because the variety is limited as $f(\alpha)$.

So the chain of logical properties propagates from the existential propositions with which we began. Now there is only one thing really going on in this simple version of the perceptron. This is given by the transfer function of the neuron α_i at time t, which the author tells us is equal to $\alpha_i(t).\nu(t)$. What is important about the system is the way it grows. The α input is the same for both classes of events considered, and so the growth function in which we are interested concerns the ν (t) element, and is expressed as $d\nu/dt$. In both the classes of events here considered, this function will involve the α input and the totality of feedback information ρ_1 and ρ_0 from the response sets to the source sets: that is, $\alpha\Sigma\rho$. Consider class C of events. There is an assumption of equal likelihood in the response alternative. Therefore the ρ feedbacks will frustrate each other in the long run, and the growth function $\alpha\Sigma\rho$ will permit the A-set ν - value system to increase indefinitely as prescribed by the inferences drawn above. In the class C′ case, however, this system tends to a limit, and $\alpha\Sigma\rho$ will not satisfy the conditions. If the ν - value system is to be limited while ν grows, the growth function must include a decay function arranged to increase monotonically with ν. Thus the growth function for class C′ may be expressed as $\alpha\Sigma\rho - \delta\nu$.

This completes *a possible description* of the Rosenblatt system. I say "a possible" because there are many more possible kinds of behaviour; I say "description" because this is nothing to do with proof - it is to do with displaying the system as given. I think the whole contents of the author's "theorems" are displayed by my description. A perceptron as defined is simply going to behave like this. It is also going to behave in other ways, because we could choose to define it in other ways. So the paper does not actually demonstrate anything. And since other kinds of behaviour are possible, no doubt the author has observed them. This may explain why his written accounts of the perceptron seem to be about different things, and why his verbal presentation today seemed to bear no resemblance to the written paper.

What then is the logical status of the two announcements which the author labels "theorems"? His method seems to be this. He asserts a

proposition about possible stochastic behaviour in the long run called a theorem. He asserts that another proposition is a corollary of this, and then has a discussion from which it emerges that the corollary class is not empty. Therefore the theorem class is not empty. This, it seems to me, tells us nothing more than that in the long run every configuration of which the system is capable will occur, and that Rosenblatt finds two of these configurations interesting. The "theorems" are, I would say, assertoric existential propositions about the system; and you neither prove nor disprove such descriptions. They are simply there, and they may or may not be interesting. What you *can* show about such propositions is that they are internally consistent, and this is I think precisely what I have done.

It follows, in my opinion, that the elaborate mathematical edifice from sections 5 to 8 in the paper is an array of tautologous remarks about the existential description given in sections 1 to 4. But as it does not actually say anything, there was no need to write it. What it does do is to build up an ethos of potency around the perceptron. On this account, someone might unfortunately be mesmerized into believing that this device is the ultimate unit of the brain that lies behind intelligence.

I feel sure that Dr. Rosenblatt would not want to make such an exaggerated claim as that; but I think that there are tendencies to exaggerated claims to be found in the paper. For instance, one which ought really to be mentioned is that, apart from the Homeostat, the perceptron is the first machine "to show spontaneous improvement". This is not so. But perhaps Dr. Rosenblatt rightly assumed that, after all, everyone here must certainly have heard of Dr. Uttley's work.

To sum up, I am saying that what Rosenblatt has shown is not that the perceptron is the necessary explanation of certain time trends observable in the brain, nor that it is a sufficient explanation. He has shown that, if we select various sub-sets of the stochastic behaviour of this worthwhile device, we shall find them interesting. And for this I, for one, would like to thank him.

Dr. W. K. TAYLOR: I think that a number of contributions to this Symposium have some similarity to Dr. Uttley's work on the classification of signals in the nervous system *(ref. 1)* and on probability computers *(ref. 2)*. Dr. Rosenblatt has introduced a difference that I described in connection with my work on the simulation of nervous systems *(ref. 3)*. This difference can be illustrated by considering an Uttley type model neuron unit with n binary inputs and a threshold m. The unit gives an output of "one" only if the sum of the "one" inputs is greater than or equal to m, otherwise the output is zero. Rosenblatt's unit, however, is equivalent to the one that I described at the Third London Symposium. The input signals were effectively continuous variables represented by pulse repetition rates and the output

pulse repetition rate was proportional to the algebraic sum of the input signals, inhibitory inputs being given a negative sign. It was shown that a suitable network of the units could learn to classify shapes if the transmission strength (Rosenblatt's "value") of the connections increased with use. There was no need to introduce a separate "growth rate control" input. This latter postulate does not appear to have a physiological basis whereas recent electron-microscope pictures of the brain *(ref. 4)* tend to support the hypothesis that the area of contact at each synapse, and hence the transmission strength, is controlled by the impulses arriving at the synapse.

One thing that seemed to be missing from Dr. Rosenblatt's paper was an estimate of how many shapes a perception with a given number of units could be expected to classify. How many A-units would be required to recognise a typewritten alphabet, for example?

Finally, Dr. Rosenblatt suggests that if he could find suitable hardware for constructing A-units, he could actually build a perceptron. I would like to ask him why he considers this to be worth while when he has already shown that he can simulate its operation on a digital computer.

DR. J. McCARTHY: I would like to raise an issue which concerns all of the neural work on perception - that is the work on perception which is based on nerve models, and I believe it applies also to the pandemonium and some of the work on speech recognition which was described this morning. The problem of perception can be divided into parts, at least two of which are discrimination and description. Now all of the work which has been described has been on the problem of discrimination; that is, a finite number of classes of stimuli are discriminated, and in the training process at least one example from each class is presented. However, much of our own perception can be described as description, whereby we can perceive something and generate a description of it, and we may be able to do this without ever having seen the thing before. I would like to give an example from letter recognition. All of us can tell A from B, and the previous speakers have made it plausible that they can train various devices to tell A from B, but we can also do the following. We can take this figure (the Russian letter) and describe it by saying, for example, that the figure consists of

REFERENCES

1. UTTLEY, A.M.; The Classification of Signals in the Nervous System, *R.R.E. Memorandum No. 1047*, 1954.
2. UTTLEY, A. M.; Temporal and Spatial Patterns in a Conditional Probability Machine, (Automata Studies, Edited by Shannon, C. E. and McCarthy, J., *Princeton University Press*, 1956.)
3. TAYLOR, W. K.; Electrical Simulation of some Nervous System Functional Activities, Third London Symposium on Information Theory, Edited by Cherry, C., *Butterworths*, 1956.
4. GRAY, E. G.; *Journal of Biophysical and Biochemical Cytology*, In Press.

four line segments, one of which is horizontal and the other three project up from it, one from each end and one from the middle. We can describe this without ever having seen it before.

I have done some work on the question of generating descriptions of visual images, but this work has not had a neural basis; it was a proposal for a programme which takes a digital representation of the picture, walks around it, so to speak, and attempts to determine the line segments.

Some kinds of description can be obtained as sequences of discriminations. In particular, we can get descriptions of speech or writing in this form, because of their sequential nature provided we can make discriminations for each letter. This constitutes a description, but I think it can be shown that, in the case of two-dimensional data, it is not possible to get a description system entirely out of discriminations.

DR. L. M. SPETNER: I would like to ask Dr. Rosenblatt if what he called $r_{\alpha\bar{\alpha}}$, the correlation between the α inputs of the association cells and the average of α, is really a good measure of similarity of stimuli. In some special cases it may be all right, but it is not clear that it is sufficiently general. The use of such a criterion, particularly in the C' type perceptron, means that the machine will tend to classify its environment according to its own construction. For example, the experiment quoted on page 448 of Dr. Rosenblatt's paper, which showed that a class-C' perceptron correctly divided all squares into those which were on the left and those which were on the right, seemed to be a direct result of the way the sensory and association cells were connected. I suspect that this perceptron might do a much worse job if it were asked to distinguish on some basis other than position. It seems to me that organisms do not operate this way. That is, they do not really go off on their own and begin classifying things in a vacuum; but rather they act more like a perceptron whose response is being forced. I do not mean that all organisms always have a formal teacher, but rather, I believe there is a continual reinforcing of responses through some kind of feedback from the environment to the organism.

I would like to think of the perceptron as a self-organizing transformation from a stimulus to a response. Each stimulus that may be applied to a perceptron can be considered as a point in a multi-dimensional Stimulus space. We might consider the response somewhat more generally than Dr. Rosenblatt does by saying that a response will consist of a *set* of numbers in each of the various response cells. Then any particular response can be considered as a point in a multi-dimensional Response space. (See *fig. 1*). The association cells then would provide a mapping of the Stimulus space into the Response space. Now it may well be that stimuli which I wish to consider similar may be very far apart from each other in Stimulus space, but if the perceptron is to organize its environment into categories of

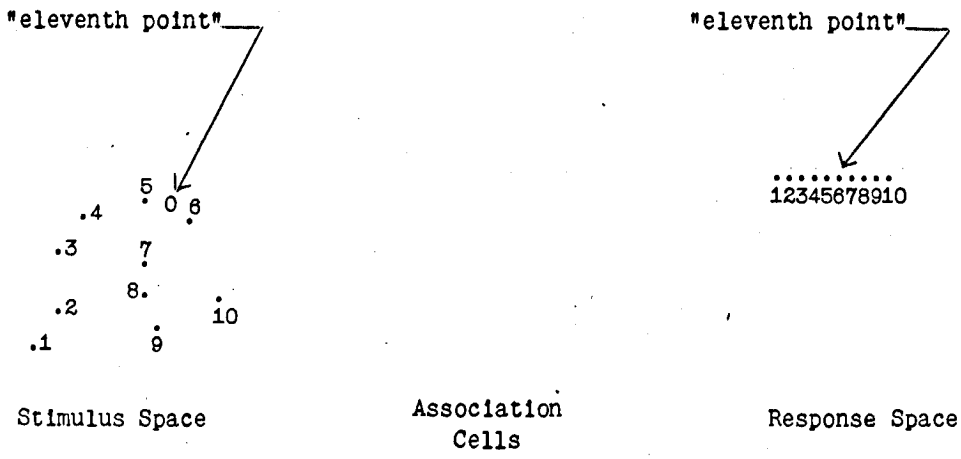


Fig.1

similarity which I choose to designate, then I shall force similar stimuli to have responses which are close together in Response space. I believe similarity of stimuli is really a rather arbitrary concept; at one time, I may wish to consider all triangles as similar and at another time I may want to consider all objects on the left side of the field of view as similar, whether they are triangles or not. Thus it is clear than arbitrary stimuli which I may wish to consider similar might well be very far apart from each other in Stimulus space. This allows for the concept of similarity to be imposed by an arbitrary experimenter rather than having it tied into the construction of the machine. Viewing the perceptron as a transformation from Stimulus to Response space also leads quite naturally to the concept of generalization. Referring again to the figure, let ten stimuli be represented by the ten points shown in the stimulus space. If I wish to consider these stimuli as similar, then I force their responses to lie in the restricted portion of Response space as shown, so that the perceptron shall learn the responses that correspond to these particular stimuli; in that it can reproduce each response for given each stimulus. Now we imagine that we present to the perceptron an eleventh stimulus that lies somewhere in between two that we have already given it in stimulus space. If our mapping has been sufficiently continuous, then it will appear as shown in the correct portion of Response space, and hence, we have reduced the concept of generalization to interpolation or extrapolation, provided one has achieved a continuous type of mapping.

DR. F. ROSENBLATT (in reply): I will try to touch very briefly on each of the points which have been raised.

The system which Mr. Newman described does indeed seem to bear a considerable resemblance to the perceptron, although the nature of the reinforcement function is somewhat different. Several of the comments which have been made here seem to reflect a common misimpression that a reinforcement function which changes thresholds is equivalent to a reinforcement function which changes the output of a neurone. This is not the case; by reducing the threshold, we are making a cell responsive to a greater variety of stimuli and thus changing the information transmitted by the cell, while in changing the output, we are changing the *weight* of the transmission channel of which the cell is a part, without in any way changing the signification of the information transmitted. A difficulty in systems which operate by changing thresholds is that as the threshold goes down, the cell becomes responsive to a greater and greater variety of stimulus events, and in the limit responds equally readily to any input, thus conveying no information at all. While such effects may not have been the source of the difficulty in Mr. Newman's device, they are apt to present an initial handicap to any system operating by general threshold reduction, rather than by the strengthening of specific transmission channels. I would predict that Dr. Taylor will encounter the same difficulty when he goes to larger models of his own system, if he continues to reduce the thresholds without a lower bound.

With regard to Mr. Newman's problem of recognizing slowly changing shapes as identical, I must say that the perceptron does indeed have such a tendency. For example, if we change a triangle gradually into a square by a series of progressive distortions, the perceptron will sometimes recognize the two stimuli as members of different classes, and sometimes will place them in the same class, depending on the frequency distribution of the various possible intermediate states. If there is a bimodal distribution, with more "good" triangles and "good" squares than distorted figures, then we can design a perceptron which will spontaneously arrive at the desired classification. If, on the other hand, all transformations (including those which are halfway between triangle and square) are equally likely, there is no logical basis for concluding that there are "two kinds of things" here, except by an arbitrary convention. How far we have to broaden the cleft between modes of the distribution in order to make the system discriminate reliably is something which we are now beginning to investigate. We are also examining the nature of the probability transition which Dr. Shuey asks about in his third question. In some cases, it seems possible to design a perceptron which will break up a continuum (say between squares on the extreme left and squares on the extreme right of the field, or between circles and ellipses) into a number of discrete domains, assigning a separate response to each. We find that the domain over which the

perceptron generalizes is governed in part by its design parameters, and in part by the particular environment of stimuli in which it is located.

The problem of the serial order in which events occur does not seem, in the perceptron, to present as serious a problem as Mr. Newman has suggested. While there is a bias, in some systems, to associate consecutive events, a perceptron which is designed so that the values of inactive A-units lose an amount which just balances the gain of the active units (what we have called the gamma-system) may actually exhibit interference between successive stimuli, if they are sufficiently different from one another.

With regard to Mr. Newman's comments on the importance of boundary or contour detection as a means of concentrating on the important input information, I am in full agreement. In addition to the networks which Dr. Taylor has described *(ref. 1)*, an approach to this problem has been considered in the first report on the perceptron program *(ref. 2)*.

Dr. Shuey has asked about the philosophical necessity of randomness in the perceptron. I do not believe randomness to be a philosophical necessity, but rather a practical one. In analyzing a nerve net about which we have complete information — a knowledge of all of its connections, the state of every unit, etc. -- it is clearly possible, in principle, to perform an exact logical analysis of its response to every logically possible input configuration, count those cases in which the response is the desired one and those in which it is not, and thus determine exactly the "probability of correct response" for the finite environment in question. In practice, this can hardly be recommended. Biological brains and physical environments being what they are, it is unlikely that we shall ever have the complete information which is required for a procedure of complete enumeration, and, moreover, this would still not enable us to talk about brains as a class of systems, but only one particular brain having the particular connections described. To talk about brains as a class we would further have to enumerate every admissible system of connections. It seems to me that the probabilistic approach is our only recourse at this point. It offers the further advantage that the necessary conditions for a class of systems to work *in general* (even though individual models may fail) can be stated in terms of a simple set of rules of construction, or statistical parameters, rather than in terms of a detailed and unmanageable wiring diagram. If we grant that we can never, in practice, attain complete information about such systems and their environments, then the question of whether probability is philosophically essential becomes entirely academic: if we are to analyze the systems at all, we really have no choice in the matter.

Concerning the relationship of perceptron discrimination to human discrimination (Dr. Shuey's fourth question), I think it is premature to expect any close correspondence between the "psychology" of any of our brain models and that of human subjects; too many ingredients are still missing in the perceptron. One important difference between the learning curves

observed for spontaneously organizing perceptrons and those of humans, however, is worth noting: The learning curves for the perceptron are convex, whereas those for a human subject under similar circumstances would certainly be concave. That is, if a perceptron is required to distinguish horizontal from vertical bars, in a spontaneous learning experiment, it quickly learns to classify the first sixty or seventy percent of the cases, and takes longer and longer to establish the correct response for the few stimuli still unclassified. A human subject, in such a problem, once he has achieved an "insight" would undoubtedly jump to 100% accuracy immediately thereafter.

I am sure I will not be able to do full justice to Mr. Beer's entrancing remarks concerning my paper. I find that I am almost mesmerized into believing them. Perhaps, however, I can arrive at *a possible description* of what Mr. Beer says that I am trying to say. I say "a possible" because I may have missed some subtlety in his analysis which makes it more relevant than it appears; I say "description" because this has nothing to do with a disproof, but with a set of fundamental misunderstandings of my paper evidenced by his remarks.

First of all, let me express my complete agreement with Mr. Beer about the lack of rigor in the paper. Had a more rigorous analysis been possible at the outset, I would have been spared the necessity of pointing to mistakes in my own work in order to defend its logical status, as I am about to do. Mr. Beer would, perhaps, have been happier with his whole analysis had the "corollaries" been labelled "lemmas" and placed before the theorems instead of after. This would actually have been preferable, as the "corollaries" are actually independent of the "theorems", although the reverse is not true -- the corollaries were used to establish a case in point, on the basis of which the more general existence theorems might be asserted. Mr. Beer suggests that this analysis is a rationalization of behavior which I must have "observed" previously in the perceptron. This is not true. Neither of the types of behavior described in the paper were observed until after the analysis was completed, at which time a new simulation program was written. It then turned out that while we did indeed obtain the predicted behavior most of the time, it was by no means infallible. Therefore, the analysis, far from being tautologous, must actually be *wrong* if interpreted rigorously. This led to the revised analysis which I outlined this morning, and which does not represent a completely new set of chimerical perceptrons, as Mr. Beer suggests, but rather a new way of looking at the same kinds of systems that we have been discussing all along.

Now what is the substance of Mr. Beer's argument? First of all, it should be noted that Mr. Beer mistakenly takes the terms "Class C" and "Class C′" to refer to classes of *stimuli*, whereas they were intended to refer to classes of *perceptrons*. Secondly it seems to me that he has systematically inverted the roles of premises and conclusions, from the original paper -- an error for which the admittedly peculiar relationship of "theorems" and

"corollaries" is probably responsible. He then proceeds to try to deduce a set of characteristics for the A-system, and particularly a system of value-dynamics, which is consistent with one of the two assumptions that (1) the perceptron, in the Class C case, effectively "uncouples" the responses from the stimuli, so that the probability of the response R=1 is 0.5, regardless of the stimulus, *or* (2) the perceptron, in the Class C′ case, has a switching mechanism (unspecified in Beer's description) located, of all places, in the retina, which is specifically designed to evoke the response R=1 for one class of stimuli, and R=0 for the other class. It should be noted, first of all, that Mr. Beer's "Class C" perceptron does not in any way correspond to our own "Class C" system. The Class C perceptron, in the original paper, is one which tends towards a terminal condition in which either (a) every stimulus evokes the response R=1 *or* (b) every stimulus evokes the response R=0. The reason there is no information transmitted in this case is not that the responses are random, but rather that they are inevitably the same, whatever the stimulus happens to be. Actually, as our analysis shows, increasing the variance of the values to infinity does not lead to random responses, as Mr. Beer infers, but rather to the perfectly consistent responses characterizing our Class C case, *provided the indicated rules of reinforcement are observed.*

I find myself unable to recognize the source of Mr. Beer's difficulty in understanding the nature of the v-set. Each A-unit, a_i, is characterized by a value, v_i, which is assumed to be initially zero, and which changes through time in accordance with some clearly specified rules. Again, the *idée fixe* that these rules are a *consequence* of some preexisting classification system, rather than the cause for the stochastic development of a classification system in an initially random machine, must be responsible for the confusion.

I think Mr. Beer's assertion that the classification scheme must somehow be built into the retina suggests that he has missed the principle point of my paper. Let me repeat, therefore, that the C′ perceptron arrives at a classification scheme *independently* of any scheme which previously exists in the mind of the experimenter. The classification arrived at by the C′ type of perceptron does *not* depend on the experimenter, nor is it built into the retina; the experimenter is forbidden to intervene in any way which might help this machine to arrive at the "right" decisions. Uttley's models do seem to represent a possible first step in this direction, and should not have been omitted from my original review of the field. Actually, I think Taylor's system comes at least as close to being a sponaneous classification device, in this case, as Uttley's, although in Taylor's model, as I understand it, the system is actually forced to learn the "proper" associations by presenting an "unconditioned stimulus" together with the "conditioned stimulus" during learning — a procedure which is logically analogous to our "forced learning" procedures with the perceptron *(ref. 2)*.

With regard to Dr. Taylor's remarks, I don't believe either of us can claim to have invented the threshold, but I agree that in setting his threshold equal to the number of input connections to a cell, Dr. Uttley is somewhat straining the biological credibility of his system. Clearly, our neuron models are very similar, as they are both based on the same fund of established physiological knowledge. The one important difference which I see in our neuron models is in the choice of a suitable memory variable -- in Dr. Taylor's case the threshold, and in my own case, the strength or "value" of the output signal, I have already indicated my reasons for this choice of variable. I would like to refer Dr. Taylor to Appendix V of my original paper *(ref. 2)*, in which it is shown that a reinforcement of the cell as a whole is logically equivalent to a reinforcement of particular connections, for an optimally designed perceptron. I agree, however, that at present the reinforcement of specific connections seems more plausible than the reinforcement of the cell as a whole, and I am particularly grateful for the reference to the electron-microscope work on synaptic areas, which I look forward to reading.

The "vocabulary size" of a perceptron, which Dr. Taylor asks about, is in part a function of the level of reliability required in the perceptron's responses. To give some idea, however, a system of 1000 A-units should be able to distinguish all letters of the alphabet with a better-than-chance probability of being correct in each case (say, somewhere between 0.6 and 0.9, depending partly upon whether or not the letters always appear in the same position, whether they may be rotated, etc.). The reasons for building a hardware model of the perceptron are (1) the simulation program is slow and inflexible, particularly with respect to the variety of stimulus material which can be presented; (2) while it is easy to change parameters in the simulation program it is much harder to introduce new qualitative constraints, which we are now interested in studying. The hardware model, with patch boards to permit flexible interconnections of the units, will make this considerably easier. A third point is that we are interested in gaining some experience in the design of components which might ultimately be used in a practical system, rather than a purely research model.

I think Dr. McCarthy may have an important point in his distinction between description and discrimination. As he is using the term, *description* implies the statement of *relations among parts*, and this is something that the perceptron is quite incapable of doing, in its present form. I think Dr. McCarthy might be interested in the work of Grimsdale and associates at the University of Manchester *(ref. 3)*, who have also attempted to develop a digital computer program for the description of visual forms.

Dr. Spetner raises the question of whether $r_{\alpha\bar{\alpha}}$ is actually the best measure of similarity to use in the perceptron. I am sure that better measures of similarity could be constructed, if this were actually what we had in mind. However, we are not trying to construct a system which will

embody some predetermined similarity function, or equation. Rather, we are constructing a system which must (1) be composed of simple, neuron-type units, and (2) be connected in a relatively free, unconstrained fashion. I do not know whether such a system will really do what we would like it to, but this is precisely what we are trying to find out. $r_{\alpha\bar{\alpha}}$, therefore, is not a measure of similarity which we have deliberately set out to use, but rather an analytic property of this type of system. The concept of a response space which maps the "similarity relations" of the stimulus space is an interesting one, which is clearly basic to the whole field. It seems to me that this describes a desirable end result, towards which the perceptron is a tentative first step.

REFERENCES

1. TAYLOR, W. K., Electrical Simulation of Some Nervous System Functional Activities, in Information Theory (Third London Symposium), Ed. by Colin Cherry, *Butterworths, London*, (1955).
2. ROSENBLATT, F., The Perceptron - A theory of Statistical Separability in Cognitive Systems, *Cornell Aeronautical Laboratory Report* # VG-1196-G-1, January, 1958.
3. GRIMSDALE, R. L., SUMNER, F. H., TUNIS, C. J., and KILBURN, T. A System for the Automatic Recognition of Patterns, Proc. *I.E.E.*, 1959, 106, Part B.