J. Theoret. Biol. (1961) **2**, 204–235.

Outline of a Theory of Thought-Processes and Thinking Machines

E. R. CAIANIELLO

Istituto di Fisica Teorica dell'Università di Napoli
(Sottosezione di Napoli dell'Istituto Nazionale di Fisica Nucleare)

(*Received* 9 *December* 1960)

Thought-processes and certain typical mental phenomena are schematized into exact mathematical definitions, in terms of a theory which, with the assumption that learning is a relatively slow process, reduces to two sets of equations: "neuronic equations", with fixed coefficients, which determine the instantaneous behavior, "mnemonic equations", which determine the long-term behavior of a "model of the brain" or "thinking machine". A qualitative but rigorous discussion shows that this machine exhibits, as a necessary consequence of the theory, many properties that are typical of the living brain: including need to "sleep", ability spontaneously to form new ideas (patterns) which associate old ones, self-organization towards more reliable operation, and many others. Future works will deal with the quantitative solution of these equations and with concrete problems of construction—things that appear reasonably feasible. With a transposition of names, this theory could be applied to many sorts of social or, more generally, "collective" problems.

1. Introduction

A. LEVELS OF APPROACH

Attempts at a quantitative understanding and analysis of thought-processes, with or without the explicit aim of devising machines that should reproduce functions typical of the living nervous system, date as far back as Ramon Lull's syllogistic wheels. They have become a recognized and major part of scientific investigation since N. Wiener's celebrated enunciation of the principles of Cybernetics; herein lies indeed clearly, much more than in specialized studies of circuitry or of information theory, the heart and scope of this new science, which aims at synthesis as well as analysis.

The investigation of the mechanism of thought has been undertaken with a variety of methods, ranging e.g. from the study of systems that should mechanize the operations of Aristotelian logic without any requirement of similarity to living structures, to the faithful electronic reproduction of populations of hundreds or thousands of neurons. We shall benefit

from all these discussions in that they permit us to reduce the verbal presentation of our own concepts to a bare minimum, since they have made abundantly clear with what cautions and restrictions one should accept for example the very expression "mechanical thought"; otherwise we shall restrict our treatment exclusively to the presentation of our approach to this problem, as we feel that in such a field judgment is passed better *a posteriori* than *a priori*, on the ground of concrete results—which are yet to be borne by any theory, including ours—than of mere opinion.

The present outline of a theory of thought-processes is the result of about three years of discussions with people who have been working with the same premises in various fields of neuroanatomy, mathematics and theoretical physics; it also reflects, of course, the evolution of our own ideas through many discussions with guests and hosts. Our main guiding principle has been the conviction, strengthened by these discussions, that the human brain, tremendous in its complexity, yet obeys, if one looks at the operation of individual neurons, dynamical laws that are not necessarily complicated; and that these laws are such as to engender in large neuronal assemblies collective modes of behavior, to which thought-processes are correlated. A convenient formulation of these laws appears therefore as the primary objective of a research of this nature; it can only be achieved by trial and error by the process, familiar in the physical sciences, of abstracting what seems relevant into a simplified model of the real thing. The present work is one such trial; its novelty is not, of course, in the concepts just mentioned, which are as old as physics itself, although they have not yet gained general acceptance among neurophysiologists, but in the attempt made here to give them a precise and quantitative formulation.

Constant resort to neuroanatomy and neurophysiology, which is the keystone of our approach, appears necessary at two different levels: the "elementary level", which studies the individual neurons and the connections, or synapses, between neurons; the "integrative level", which studies the structure and function of specially connected assemblies of neurons, which may act as a whole and play in the nervous system a rôle similar to that of specialized organs in the body. The integrative level compares with the first as the physics of matter does with that of the atom, and is of course as essential to the understanding of the functions of a brain or of a thinking machine; we firmly believe, against the opposite views which we have heard expressed, that a study at the elementary level is as essential to the second as one type of physics is to the other.

We shall have very little to say here about the integrative approach, in which many more investigations are needed before a satisfactory state of knowledge is achieved, except that our equations, once the appropriate connections among neurons of a given assembly are introduced into them,

will permit the quantitative study of its collective behavior as a whole. It is our belief that the subsystems of a brain are quite different, in structure and complication, from the standard circuits of electronics, and that there will be a great deal to learn in this respect from neuroanatomy; also, that a thinking machine built for some special purpose may well need organs, or subsystems, organized quite differently from those of the animal brain, although the same elementary laws will be valid. Our equations are also intended to provide a useful tool both for theoretical study and for experimentation in this respect.

Because of the lack of definite knowledge and of general agreement among specialists on many facts of neuroanatomy and neurophysiology on one hand, and of the great wealth of available observational material on the other, we think it best to present our views as the direct description, reduced to bare essentials, of a *model* of the brain forgoing the detailed analysis of anatomical data from which, in fact, our considerations stem.

B. THE MODEL

By "model" or "machine" we mean exclusively a device *that can actually be built*, and which operates according to mathematical equations that are *exactly known and numerically solvable* to any wanted accuracy. Although this necessarily implies drastic schematizations and simplifications, it is hoped that the features essential to thought-production are retained by the model; successive approximations to reality will require improvements in the structure of the machine and in its operational laws, but at each step one must know exactly what is being done. Without a complete mathematical control of the situation, a machine may perhaps think, but one would hardly know why or how.

Mathematically, our model consists of two sets of equations: the "neuronic equations", which describe the instantaneous operation of the machine; the "mnemonic equations", which describe the growth of memory into it. From these equations it is possible to predict and study the "mental" phenomena which are typical of such a machine: learning, forgetting, re-integration, conditioning, analysis of patterns and spontaneous formation of new patterns, self-organization into reliable operation. An exact mathematical definition is given of each of these phenomena; that they do actually take place is shown, qualitatively but rigorously, from the form of both sets of equations; methods for the quantitative solution of these are in part already available and will be discussed in a future work. Likewise, although we are actively engaged also in the study of the concrete aspects of the question, we shall limit the present report to an outline of the mathematical theory.

C. NORMAL PHYSIOLOGICAL MODEL

Finally, we wish to emphasize that our machine does not purport to realize necessarily *an anatomical model* of the brain, that is, there need be no one-to-one correspondence between the anatomical neuron and the basic unit of the machine; we are concerned here only with the description of a *physiological model*, in which, as a whole, it is irrelevant whether the functions of a single neuron are taken up by a single unit or by a group of units in the machine, or vice-versa. Likewise, one could reproduce the functions of a circuit containing electronic tubes of various descriptions in terms, say, of a model circuit containing only triodes. We wish to emphasize also that our model intends to simulate the physiology of neurons in their *normal* condition in the living tissue, and not at all the various reactions they exhibit when tortured in the physiologist's laboratory: most of the latter will be as irrelevant to the study of the collective behavior of neuronal assemblies, as is the detailed knowledge of the radiation spectra of Na and Cl ions to the determination of the crystalline structure of the NaCl salt.

2. Symbols

$1(x)$ = unit step function
Σ = Stieltjes integral or summation
h, k, i, r = indices denoting integers (subscripts of superscripts)
R, N = fixed integers
$a_{hk}^{(r)}, A_{hk}^{(r)}, b_{hk}$ = real numbers (coupling coefficients)
s_h = real numbers (thresholds)
t = time variable
τ = a fixed "time quantum"
$u_h(t)$ = piece-wise-constant functions (-0 or 1 in any quantal interval of time)
ν = class of all functions $u_h(t)$
$c_i(t)$ = a constellation of neurons at time t
$n_i(t)$ = number of neurons of $c_i(t)$
E = a fixed set of neurons
$M_E, \overline{M}_E$ = classes of solutions of eqs. (2) relative to E.
$G(t)$ = group of transformations under which eqs. (2) are invariant
$\mathcal{N}$ = configuration space
$P(t)$ = representative point of system in $\mathcal{N}$
$\mathcal{F}$ = a functional space built over $\mathcal{N}$
$S(t)$ = a frame in $\mathcal{F}$
$\Theta_i, \Theta_{ij}, \ldots$ = patterns presented to, or constructed by, the machine

$f(a_{hk}; \lambda)$ = secular equation with variable λ
$\bar{\lambda}_h$ = eigenvalue of $f(a_{hk}; \lambda) = 0$
ρ_{hk} = small random variation of a_{hk}
ρ, $<\delta\bar{\lambda}_h>$ = average values of ρ_{hk} and of corresponding variation of $\bar{\lambda}_h$

3. Neuronic and Mnemonic Equations

A. GENERAL REMARKS

1. The present considerations aim at simplicity, rather than at formal elegance; many restrictive assumptions are therefore made which could easily be relaxed, gaining thereby a greater apparent generality in our equations but, in reality, only complication which is better avoided at this early stage. The most evident is the fact that we use throughout summations instead of integrations, although Stieltjes integrals would be in many cases more appropriate to a faithful description of the anatomical situations of interest.

Instead of considering the actual speed of propagation of the neuronic discharge along dendrites and axones, we neglect the first and lump the second together with the synaptic delay into a single time-unit τ; this is a better approximation than it may seem, because in the brain, as is well known, speed in axones is proportional to diameter and, although less generally, diameter to length. We schematize this situation by assuming that a neuron which receives a pulse (either does not fire or) fires after exactly τ sec; or, more generally, that τ denotes some conveniently small "time quantum", of which the neuronic delay times are (not necessarily equal) multiples (our neuronic equations (2), although apparently designed to describe only the first situation, also cover the second).

2. We shall base our treatment on two sets of equations: the *neuronic equations* (N.E.) which have constant coefficients and determine the instantaneous behavior of the system, and the *mnemonic equations* (M.E.) which account for the semi-permanent or permanent changes in the structure of the system caused by its past operation. This is, again, an artificial simplification of the actual situation, which is better described by retaining only the first set of equations, with coefficients taken as "slow" functions of time and past neuronic activity. The approximation thus made is analogous to the Born-Oppenheimer approximation of molecular physics, which consists in studying first the motion of the (much faster) electrons as if the nuclei were fixed, and then the behavior of the latter. It is justified physiologically by the experimental observation that electroshock, or concussion, cancels all memories of things learnt within a previous time interval of minutes or more, while memories acquired before

that time remain unimpaired: this makes it reasonable to assume that the brain takes about that much time to change the dynamical phenomena which we consider here to be the carriers of functional, short-range memories, into semi-permanent or permanent alterations. That the latter actually exist is proved by the fact that they are not suppressed by hybernation or artificially provoked cessation of all neural activity.

We may call this the *adiabatic learning hypothesis* (A.L.H.): the degree of adiabaticity of learning in the brain can be estimated roughly from the remark just made, with the conclusion that the engramming of permanent or semi-permanent memories takes roughly a time of 10^4 to 10^5 sec or more. The determination of the duration of semi-permanent memories in the brain is a task for experimental psychology, and is not discussed here.

The mathematical advantages of uncoupling the actual equations of neural activity into two distinct sets by means of the A.L.H. will be evident: by considering all constants frozen, the resulting N.E. are solvable notwithstanding their utter non-linearity, and in any case their very form leads immediately to many interesting qualitative conclusions, as we shall show later.

3. It is perhaps relevant to emphasize that the equations which we shall take as the basis of our treatment do not certainly contain, in themselves, any striking novelties. They are about what any neurophysiologist would write at once, should he wish to arithmetize, say, the kind of logic that is usually associated with neuronic circuits, or to formulate some reasonable guess about the growth of memory.

What we consider to be the essential point in our whole theory is, rather, the fact that arithmetization is considered here as the *necessary first step*: once equations are written, then, and only then in our opinion, the real groundwork can begin. Furthermore, equations alone mean very little to a mathematician; the detailed prescription of the type of information which is wanted from the solutions of a given equation constitutes a "problem", the formulation and solution of which is, in all cases, the most relevant question. We shall therefore be concerned here essentially with the formulation of problems which arise from these equations and are central to our theory of thought-processes; in so doing, we shall meet interesting and novel mathematical situations, the quantitative study of which is well under way and will be reported in the future. The qualitative discussions of Sections 4 and 5 will suffice for our present purposes.

B. NEURONIC EQUATIONS

1. We take as the basic component of the machine—which for convenience we call a "neuron", although its functional relation to living

neurons need not be 1 : 1—a discriminator with a large number of inputs (dendrites) and a large number of outputs (branching axones). Signals can only travel *unidirectionally*, with infinite speed, from the output of a neuron to the input of the neurons connected to it; when a signal reaches a neuron it is annihilated, unless enough signals arrive with it to cause the neuron to fire a pulse, after a delay τ, simultaneously in *all* its outputs. The intensity of these pulses may vary with the "anatomy" of the neuron, i.e. number of inputs, outputs, location in the machine, etc.; such pulses may be attenuated during propagation, or other phenomena may occur, as is discussed later. As a matter of formal convenience, we normalize all pulses to unit strength and account for larger or smaller strengths by giving suitable values to the coupling coefficients. Finally, a neuron will fire only if the total sum of afferent pulses is greater than its threshold. All coupling coefficients and thresholds are considered to be constant (adiabatic learning approximation).

We define the function:

$$1(x) = \begin{cases} +1 \text{ for } x > 0 \\ 0 \text{ for } x \leq 0 \end{cases} \tag{1}$$

let $u_h(t)$ denote a function belonging to the class U of piece-wise-constant functions which are either constantly 0 or constantly 1 in any of the intervals $l\tau$, $(l+1)\tau$ (l integer ≥ 0); we take then as fundamental equations for the description of the instantaneous behavior of our machine (neuronic equations, N.E.):

$$u_h(t+\tau) = 1\left[\sum_{k,r} a_{hk}^{(r)} u_k(t - r\tau) - s_h\right] \tag{2}$$

The meaning of the coefficients $a_{hk}^{(r)}$ and s_h is stated below; the anatomy of the machine at a given instant is described entirely by their values. (Taking $\rho_h\tau$ instead of τ at l.h.s. of (2) would not change the structure of these equations: an obvious re-naming of their coefficients would lead back to the form (2).)

2. s_h, usually > 0, is the threshold of the neuron h; the neuron h fires at time $t+\tau$ if its *excitation* at time t (given by the sum in (2)) is greater than s_h.

$a_{hk}^{(0)}$ ($k \neq h$) is the *coupling coefficient* that transfers the pulse originating from neuron k to neuron h; it contains the *total effect* of the first on the second, *regardless* of the number of synapses between k and h and of the intensity with which the stimulus coming from k reaches h along each pathway. When $a_{hk}^{0} \neq 0$, we say that there is a (unidirectional) *direct channel* between neuron k and neuron h, which causes a *facilitation* $k \to h$ if $a_{hk}^{(0)} > 0$, or an inhibition $k \dashrightarrow h$ if $a_{h}^{(0)} < 0$.

The rôle of the coefficients $a_{hk}^{(r)}$ ($h \neq k$) and $a_{hh}^{(r)}$ is quite different:

$a_{hk}^{(r)}$ ($h \neq k$; r integer > 0) is $\neq 0$ only if it is required that the actual mechanism of stimulation be such that the effect of the pulse from k may reach h, or last on h some time $> \tau$ after k has ceased firing; this would be the case if stimulation were due, say, to some transmitter substance released at the synaptic junction, which would be re-absorbed only after a time $> \tau$. Such a mechanism would account for latency and be related to the well-known dependence of pulse frequency on intensity of stimuli. It may not be a bad approximation, in a model, to take $a_{hk}^{(r)} = 0$ for $h \neq k$, $r > 0$, except perhaps for input elements.

The coefficients $a_{hh}^{(r)}$ express instead the memory that the neuron h retains of each of its firings (in the brain, for about 100τ sec). For all we know, the characteristic observed shape of the neuronic discharge (as well as many other things) may well be only the result of biological necessity, and to ask that it be closely reproduced in a thinking machine might prove as binding as demanding that moving objects be built with legs rather than wheels. We shall want in any case $a_{hh}^{(r)} \ll 0$ for all values of r from $r = 0$ until $r\tau$ becomes greater than the absolute refractory time of the neuron; for the latter and higher values of r it may be convenient to follow different prescriptions, according as one wishes to study the actual behavior of the brain on this model, or instead to construct a thinking machine for some special purpose.

3. As an example (among the many that might be produced) of the fact mentioned earlier that our N.E. might be a poor description of the anatomy and yet give a faithful description of the physiology of a nervous system, we consider here the situation that would arise if, in a nerve, or bundle of fibers, the electrotonus due to axones which are carriers of pulses should induce firings in other axones of the same nerve which originate from neurons that *have not fired.*

This possibility was not contemplated when writing the N.E. (2). A model which reproduces also this new type of behavior must lead to equations such that signals can be either *transmitted directly* from neuron k to neuron h, or *induced* into the channel $k \to h$ by the firing of some neighbouring neurons; the neuron h must not be able to discriminate whether the pulse it receives through that channel has a direct or induced nature. Taking for simplicity $a_{hk}^{(r)} = 0$ for $r > 0$ ($h \neq k$), we obtain clearly the wanted equations by replacing $\sum_r a_{hk}^{(r)} u_k(t - r\tau)$ in (2) with

$$a_{hk}^{(0)} u_k(t) + \sum_{k_i \neq k} b_{hk_i}^{(0)} u_{k_i}(t) \tag{3}$$

where $\sum_{k_i \neq k}$ means sum over the neurons k_i, neighbours of k, the axones of which can act in this way on the channel $k \to h$, and $b_{hk_i}^{(0)}$ are some suitable coefficients.

It is then evident that, renaming the coefficients, one finds again N.E. of type (2). The same can be said for inter-dendritic interference.

C. MNEMONIC EQUATIONS

1. There is sufficient evidence to prove that memory in the brain is due both to functional processes and to reversible and irreversible alterations of its micro-structure. Very little, if anything, is known for certain beyond this, so that we are forced to rely upon "plausible" hypotheses if we wish to assign the specific laws which determine semi-permanent or permanent physico-chemical changes. We shall not hesitate to do so for the sake of concreteness; we wish however to emphasize that the qualitative analysis of thought-processes which is the purpose of this work does not require precise knowledge of these laws, but only that they share some very general features, which may be assumed with much greater reliability.

Here lies a substantial difference between the brain and the thinking machine: the latter, which is obviously not restricted by the severe limitations of biological necessity, may have mnemonic devices and laws much more efficient than those of Nature, while giving rise to thought-processes (as described by the N.E.) of the same type. We feel also that, as regards memory growth and contrary to the situation that arises in the study of the N.E., a thinking machine of this sort might be of greater use to neurophysiology than vice versa; observations performed on models, which can be built with mnemonic laws changeable at will, might help to shed light on the quantitative aspects of biological phenomena which are extremely difficult to observe directly.

Thought-processes in a portion of the cortex may be ascribed either to excitation of neurons which would be otherwise mostly at rest, or to inhibition of the activity of neurons which would be otherwise unceasingly firing. To the first one would associate mnemonic mechanisms which make firing easier with the progress of learning (this we may call a *facilitatory*, or *positive*, type of memory); the opposite with the second (*inhibitory*, or *negative memory*). Both types offer interesting possibilities for machine construction; since they obey essentially the same kind of N.E., we refer here throughout only to the first type.

The so-called "genetic", or "anatomical", i.e. permanent inherited memory, corresponds clearly in our description to the fact that, as we shall see, some (actually most) of the coefficients which couple neurons together must be taken initially, and kept throughout, vanishing. Our mnemonic laws will therefore be chosen so that if a coupling coefficient vanishes initially, it stays forever so, while its modulus may grow to maximum value from any given initial non-vanishing value.

In our model, thought-processes will be represented by non-trivial

solutions of the N.E.; the machine can also "think" therefore if all coefficients in the N.E. stay forever frozen, i.e. if the machine cannot learn or forget, provided these coefficients have convenient values. The present framework can thus account, as it should, for a clear distinction between "instinctive" and "intelligent" behavior. It is natural to suppose that genetic patterns determine the laws according to which cells duplicate, branch out and anastomize, rather than the actual ultimate detailed anatomy of a tissue (thus, a "gene" carrying the instruction "add + 1" would suffice to generate all integers from zero, while an infinite number of "genes" would be obviously required if each integer should have its distinctive "gene"); then even a few mutations may determine the appearance of neural structures quite at variance with previous patterns, from which the evolutionary laws can secure the selection of the fittest, that is those which possess the most favorable neuronal couplings. Our definition of thought comprises thus two types of performance for which we use the conventional terminology: "instinct", which is learnt genetically, and "intelligence" which arises when these couplings can change during the life of the individual.

2. The quantities s_h, $a_{hh}^{(r)}$ and $a_{hk}^{(r)}$ $(h \neq k)$ were seen to play quite different rôles. When assigning their variation with time, we refer henceforth to a machine rather than to the living brain, for the reasons mentioned before.

It is apparent from (1) and (2) that the maximum learning capacity of the machine is already reached by assigning suitable variations only to $a_{hh}^{(r)}$ and $a_{hk}^{(r)}$. Once the mnemonic laws are given for these, changes induced in the s_h appear as the best way of controlling the operation of the machine. We shall return on this point in Section 4 and consider here the s_h as quantities the values of which do not change because of mnemonic laws, but, if at all, through some different mechanism.

The coefficients $a_{hh}^{(r)}$ have already been discussed in B, 2, p. 210; for the purposes of the present discussion we may assume.

$$a_{hh}^{(r)} = \begin{cases} -\infty & 0 \leq r \leq R \\ 0 & r > R \text{ (integer)} \end{cases} \tag{4}$$

For $h \neq k$ a convenient law is (for positive, or facilitatory $a_{hk}^{(r)}$):

$$\frac{da_{hk}^{(r)}(t)}{dt} = \{\alpha^{(r)} u_k(t-\tau) u_h(t) - \beta^{(r)} \mathrm{I}[a_{hk}^{(r)}(t) - a_{hk}^{(r)}(0)]\} a_{hk}^{(r)}(t) \mathrm{I}[A_{hk}^{(r)} - a_{hk}^{(r)}(t)], \tag{5}$$

where $\alpha^{(r)} \gg \beta^{(r)} > 0$, $A_{hk}^{(r)} > 0$, and it is imposed that $a_{hk}^{(r)}(t)$ be continuous, with $a_{hk}^{(r)}(0) \leq A_{hk}^{(r)}$.

For the sake of concreteness we take (5) as the mnemonic equations (M.E.) of our machine; we also neglect inhibitory (negative) couplings, to which (5) is immediately extended in an obvious manner. We may suppose here, for simplicity, that only coefficients with $r = 0$ survive, and that all $A_{hk}^{(0)} = A$ and all $a_{hk}^{(0)} = a$. We have already emphasized that all that we actually need are M.E. that admit solutions having the same qualitative behavior as those of (5); these we proceed to discuss briefly.

3. We write $a_{hk}(t)$ for $a_{hk}^{(0)}(t)$. The M.E. (5) describe a situation in which $a_{hk}(t)$ never becomes smaller than $a > 0$, nor greater than A. When the latter value is reached, it is retained for ever: the information is engrammed permanently. This is perhaps an oversimplified view of the real situation in the brain; it could be, though, easily modified.

$a_{hk}(t)$ increases if, and only if, the neuron h, which is connected by a direct channel to neuron k, fires at time $t + \tau$ and has received a pulse at time $t - \tau$ from the latter. It decreases slowly afterwards ($\alpha \gg \beta$), until the same situation repeats. Only if a series of such rises occurs, without excessive delays in between, can $a_{hk}(t)$ reach the engramming value A.

There is ample choice of mechanical devices which can reproduce qualitatively this behavior. If it is desired that the machine exhibit a behavior typified by (3), coefficients like $b_{hk}^{(0)}$ might be given constant values, not subject to mnemonic phenomena.

4. Qualitative Discussion

A. OPERATIONAL DEFINITION OF "THOUGHT"

1. We propose now to show that, as was mentioned in the Introduction, a machine that works according to the N.E. (2) and the M.E. (5) will exhibit phenomena which are typical of a nervous system, provided of course the number of its elements is sufficiently large and the initial values $a_{hk}^{(r)}(0)$ of the couplings among these ("genetic memory") are conveniently chosen (e.g. so as to prevent "epilepsy": cf. C, 4, p. 221).

The most obvious features of the N.E. are non-linearity and unidirectionality of pulse transmission; their solutions describe therefore in any case states of excitation (or "motions", or "modes") that "travel" unidirectionally from neuron to neuron and interfere nonlinearly whenever they meet. This interference is either instantaneous or nearly so, as it happens when summation of pulses at the synapsis of a neuron causes its firing (as described in the r.h.s. of (2)); or delayed, as it happens when pulses, which would otherwise cause the firing of a neuron, cannot do so because they reach that neuron when it is still inhibited by a previous firing, due to different pulses.

2. *We define a thought-process*, operationally, *as a solution of the N.E.*, or, equivalently, as the corresponding "motion" in the machine. It is convenient further to qualify this definition, so as to meet obvious objections.

We may disregard as "trivial" and not consider as "thoughts" solutions that correspond to (say accidental) firings of neurons at a given time, such that no other neurons are induced into firing thereby and all activity ceases immediately afterwards. Any "thought" implies thus the passing of at least one neuronic channel.

3. For any given set E of neurons, all motions of a given duration can be classed either into a set M_E, the motions of which cause at least one neuron in E to fire at least once, or into a set $\overline{M}_E$ of the remaining possible motions. There is thus (and in many ways) the possibility of establishing operational distinctions between "types of thought"; should, for instance, a portion of the machine correspond to the central and one to the autonomic nervous system, the name "thought" could be further restricted thereby to the solutions of the N.E. which affect only the neurons of the first. If, in a different partition, E is the set of neurons the firing of which is associated somehow with consciousness (e.g. because they control a loudspeaker, or some prescribed feed-back mechanism), then all motions of M_E can be taken as representing the "conscious activity", all those of $\overline{M}_E$ the "subconscious activity" of the brain.

It is interesting to remark that, in the latter instance, because of the various possibilities of interference discussed before between the motions of $\overline{M}_E$ and those of M_E, each type of activity influences the other. "Psychoanalysis" reduces for this machine to a simple and well-defined mathematical problem.

B. PROBLEMS CONNECTED WITH THE N.E.

1. The N.E. clearly contain, as special cases, the description of all logical networks of the kind beautifully analyzed in the pioneering work of McCulloch and Pitts. Should their solution be attempted by the obvious method of iteration, they would, for these cases, give just as much—or as little—information as can be gathered from the standard logical switch-board analysis; there is here a clear analogy with the Darboux (better than the Cauchy) problem of the theory of differential equations.

The systematic algebraization of logic, which is the real content of the N.E. (with frozen coefficients), permits us to pose for them much more general questions, which may be treated with a variety of mathematical tools; the logic of the system is seen to play a rôle so to speak similar to that of the constraints which limit position and mobility of a dynamical system; an appropriate treatment of the N.E. will permit, as with the

equations of motion of dynamical systems, the search for those long-range collective solutions which, in our scheme, form the basis for a useful analysis of thought-processes.

2. We consider first of all the N.E. with frozen coefficients, in keeping with the A.L.H. Their quantitative discussion poses some interesting and novel mathematical questions, and will probably require the introduction of techniques *ad hoc*; on the other hand, it is evident that in simple cases, such as may correspond to situations involving very few neurons, the N.E. may be solved on inspection. It is also clear that straightforward combinatorics can give useful information on the possible types and multiplicities of the solutions of interest, as defined below; and that this can be translated at once into the customary language of "excitation probabilities", etc. While deferring to future reports for detailed studies on these matters, on which work is in progress, it is fully sufficient for our present purposes to formulate the "problems" which we envisage as most relevant in study of the N.E., and to discuss them briefly at a qualitative level.

The first obvious, and obviously important, remark is that the N.E. are not uniquely determined; their formulation (2) is perhaps deceptively simple. Because of the definition of the function $1(x)$, there is a whole group G of transformations which change a given set of N.E. into an *equivalent* one—having, that is, exactly the same solutions, although not necessarily the same form (thus, $1(x) = 1(2x) = 1(x^3) = 1(\sin x)$, etc.). This fact was already used in the discussion of the threshold values s_h made in Section **3**, C, 2; it shows, for instance, that matrix algebra should be used with caution in handling these questions.

For the same reason, "suitably small" changes of the $a_{hk}^{(r)}$ and s_h will not change the solutions of a set of N.E.: this adds credit to the reliability of the A.L.H. and provides what we may call the *first criterion of stability* of the machine.

3. We shall soon specify what types of "input" and "output" seem most appropriate for a machine of this sort; we are now interested in the "spontaneous" activity of the machine, which we define as that which takes place in it when, at a given time t_0, the machine starts from any given state of excitation and no input pulses are fed into it for $t > t_0$.

In a linear network—it is convenient, for purposes of comparison, to refer to a system of harmonic oscillators with linear couplings—such activity is naturally analyzed in terms of eigensolutions, eigenfrequencies, harmonics; the behavior of a single element is in general not periodic, but simple periodal analysis will resolve it into a sum of periodic normal modes, which have a collective character and may be defined as the motions of quasi-particles (this remark already suffices to eliminate as

illusory any attempt at deciding on the existence of periodic motions in the brain through observations performed on one or few neurons). In a non-linear system things become much more involved; e.g. one finds in general, besides harmonics (multiples of a fundamental frequency), also subharmonics (multiples of a fundamental period).

The extreme schematization which is expressed by the form (2) given here to the N.E. has the evident consequence that one can only expect subharmonics; if there are periodic solutions of the N.E., these are *reverberations*, i.e. transfers of excitation from neuron to neuron which may reach anywhere into the machine and, after the closing of suitable multichannel paths, repeat with a periodicity which is, obviously, some integral multiple of τ.

The consideration of reverberations is central to our approach. There are tremendous numbers of them even in the simplest conceivable models; their types, paths, multiplicities are determined by the coefficients $a_{hk}^{(r)}$ and s_h of the N.E., and change therefore, because of the M.E., with learning and forgetting.

4. The first mathematical problem is therefore the determination of all the solutions that correspond to reverberations, or free ("spontaneous", "autogenic") modes compatible with the N.E. The minimum duration of a reverberation is clearly determined by the refractory period of the neurons through which it travels; if we assume that "normal" activity (i.e. without special stimulation) of the neurons in the brain uses the total period of the pattern of spike-afterpotentials ($\sim$ 100 msec), then the maximum possible frequency (reverberations involving $\sim$ 100 neurons) is about 10 cycles/sec, which coincides with the frequency of the α-waves of the E.E.G. If we assume further that stimulation may force the neurons of the brain into using a refractory time intermediate between the absolute ($\sim$4 msec) and the total time, then this maximum frequency increases and the number of neurons necessary for the smallest permissible reverberations decreases. We do not wish to draw any conclusions at this early stage from these remarks, which may be a gross oversimplification of reality; we only state here that they are not in disagreement with observation.

If thresholds and couplings have the values that are observed in the brain, then reverberations certainly involve several tens of neurons. Reverberations, furthermore, should last for ever in an ideal machine, a conveniently long time in a real machine and in the brain. From the first remark it follows that one cannot expect to observe direct evidence of prolonged autogenic activity in a portion of the cortex in ordinary conditions: this would require innumerable microelectrodes stuck into as many neurons for an experiment to be feasible. One would expect, how-

ever, from our theory, that if thresholds are sufficiently lowered artificially or the intensity of stimuli increased, then also a very small number of aptly chosen neurons should suffice for a prolonged autogenic reverberation to take place. In a brilliant series of experiments A. Féssard (Symposium on Memory, Naples, 1960) has demonstrated, by using tetanic potentiation, that this actually happens: he recorded reverberations among only four neurons which would last minutes. We regard his results as a crucial, if only partial, confirmation of our theory, which was developed while we were still unaware of his work.

5. Reverberations, as all other motions, interfere non-linearly unless one reverberation never affects in any way the neurons of another, i.e. as we shall say, is *disjoint* from the other. At this point the analogy with a linear network breaks down completely, much to the advantage of our machine, which possesses many more essentially distinct modes of behavior than a linear system. It is still possible to classify *all* possible spontaneous reverberations, for instance according to periodicity, multiplicity (i.e. degeneracy), etc.

The next mathematical problem that arises is the study of the evolution of the state of excitation which was present in the machine at time $t = t_0$, as was discussed in 3, above. It may either coincide with a configuration of excitations which characterizes at t_0 a reverberation, and thereafter continues its periodic behavior; or, more often, *decay*, into a reverberation, or *develop* into a reverberation, or produce *catastrophic behavior*, i.e. lead to total (or nearly total) simultaneous excitation ("epilepsy") of the neurons, which may decay immediately afterwards into rest (cf. the N.E.).

Excluding for the time being the last dramatic alternative, we find here the most interesting situation, as close an analogue as is possible with a non-linear system to harmonic analysis. In a frozen state of knowledge, out of the machine as many distinct responses can be evoked as there are distinct excitable reverberations, or modes; each of these we may identify with a "pattern" which the machine knows genetically, or has learnt; the "initial configuration" is the pattern which is presented to the machine; the set of (one or more) disjoint reverberations to which the latter gives rise (depending upon the value of the couplings) is the *analysis of that pattern* performed by the machine which corresponds to the state of knowledge it has learnt until that moment.

Apart from learning, we have here the counterpart to what we regard as the essential activity of the mind, the ability to analyze a situation, or shape, or pattern, into a set of already classified patterns. No single element acts as a classifier since the total response of the machine is required for this analysis.

6. The situation described above is manifestly an extreme simplification. The next mathematical problem is in fact that of studying the evolution in time of the total state of excitation of the machine when its "input" is subject to continued external stimulations.

It will also be expedient, of course, whenever dealing with very large assemblies of neurons, to distinguish between "traveling" and "stationary" solutions. We have been considering thus far only the latter, but it is clear that, as soon as distinctive special-purpose "organs" are built into the anatomical structure of the machine, our previous considerations should be restricted mostly to the latter, with pulses travelling from organ to organ as among the boxes of a diagram.

It is also to be expected that there will be a maximum duration, and a maximum complexity, beyond which reverberations cease to be significant for pattern-analysis. This assumption, or requirement, will greatly facilitate the mathematical study of the problem formulated in this section.

7. We have thus far taken, for the sake of simplicity, a perhaps too realistic view of reverberations as modes which are actually connected to fixed chains of neurons. This is not certainly the case when one considers the normal modes of linear networks, and it is therefore of interest also to investigate the possibility of resolving actual motions, which do not have manifest periodicity, into "normal" periodic collective modes (cf. 3, above: the Lissajous figures of linear problems are an example of this behavior); any such latent periodicity would be easily revealed by observations made upon populations of neurons (e.g. with the E.E.G.).

Questions of this nature, and many others, suggest themselves in a quantitative investigation; they need not be considered here in further detail.

C. RÔLE OF THE M.E.

1. In the preceding section we have focused our attention on the operation of the machine when all coupling constants and thresholds are kept fixed, and have found that *reverberations* play a central rôle in its most typical activity, which is *pattern-analysis* in a very general sense. All such statements presuppose already, of course, the existence of favorable conditions, as are expressed for instance by the assumption (A, 1), which prevent epileptic, or catastrophic, behavior; it was also implicitly assumed that the machine is indifferent to the "meaning" (referred to any standards) of what it knows genetically or has learnt during its past activity. While we can reasonably expect that careful engineering and a long series of painstaking adjustments would in the end produce devices capable of some useful performance solely by virtue of conveniently chosen N.E., we are

much more interested in machines that can adjust themselves to prescribed tasks by means of some learning mechanism; this should also give the machine a tendency to organize itself into increasingly reliable operation, so as to compensate for minor flaws in the accuracy of its elements.

The M.E. provide, to a large extent, the answer to these questions, as we shall now show. In the course of the same discussion it will become apparent, however, that a machine of this sort is not realistically conceivable unless at least two additional controlling devices are not also explicitly included; the first we identify tentatively with the thalamus, the second, with more assurance, with the reticular system of the brain. The necessity of devices of this sort, if not already suggested in the brain by anatomical and physiological evidence, is made imperative in the machine by the structure of the N.E. and M.E.

Mentioning a "thalamus" takes us, of course, one step nearer to the "sentient" machine than we wish to stay for the time being; we shall therefore restrict this part of our discussion to barest essentials, pointing only to what is relevant for purely "rational" thought.

2. There are many mathematical ways of representing the overall situation and evolution of the machine, each suited to some special purposes. We mention here briefly a few which take the instantaneous state of each neuron as the object of interest.

If the number of neurons in the N.E. is N, then a solution of the N.E. at time t is representable by means of a one-column matrix with N rows, the element of row h being given by $u_h(t)$; or one can define, equivalently, an N-dimentional *configuration* (or *neuron*) *space* $\mathcal{N}$, which has N axes, on the hth of which the abscissa is $u_h(t)$. The state at time t of the machine is thus represented by the *point*, or *matrix*, or *vector*, $P(t) \equiv \{u_h(t)\} \equiv \tilde{u}(t)$; its evolution in time by the (discontinuous) motion of the point $P(t)$.

All trajectories in $\mathcal{N}$ are invariant under the transformations of the N.E. which belong to the group G defined in B, 2, provided of course that at each time t one takes G as it is determined by the M.E.: now, $G = G(t)$.

$\mathcal{N}$ contains at most 2^N points; a trajectory in $\mathcal{N}$ is a polygonal joining some, or all, of these points. A reverberation is represented in $\mathcal{N}$ by a *closed polygonal* (and lasts at least as long as the coefficients in the N.E. stay frozen).

$P(t)$ changes in $\mathcal{N}$ (i) because it describes the evolution in time of a solution of the N.E., (ii) because the N.E. themselves change, due to the intervention of the M.E. The A.L.H. permits the study of simple phenomena by separating step (i) from step (ii): it allows, that is, that they be performed alternately. Step (ii) becomes necessary as soon as, because of M.E., the N.E. undergo a transformation which does not belong to G.

A qualitative discussion is better stopped here (see, however, Section **5**, A, 2); it should already be clear from what little has been said on this subject, though, that the introduction of spaces of functionals on $\mathcal{N}$ will be of the highest conceptual importance, because then everything becomes again linear, group theory may be resorted to as a valid tool of analysis, each "pattern" is easily made to correspond to a point, and problems such as those of language translation or study of emotive behavior can receive a precise mathematical formulation.

3. The pattern-analysis described in B, 5, presupposes, clearly, that the machine has already formed, either genetically or by learning, some typical responses (or modes, or patterns), in terms of which a pattern presented to it is analyzed. Very little, if anything at all, can be expected from a machine with fixed constants built entirely at random: the most likely thing to occur in such a case is that, unless the experimenter arranges the connections of the machine in a way that is equivalent to giving it a genetic memory, the only resulting effect will be a total loss of information. Also a machine endowed with ability to learn will give, at best, a poor performance, unless the controlling devices mentioned above (1) are included into it, as we shall soon discuss. The best procedure, or at least by far the most economical, appears to be in any case that of borrowing as much as possible from anatomical and physiological information.

Before proceeding further we need say a few words about the kind of "input" and "output" which seems appropriate to a machine of this nature. The notions of input as "that which comes before" and of output as "that which comes after" the machine proper are clearly out of the question; we are interested in what occurs *at all places* and *at all times* in the machine, and a relatively small number of terminal plugs could never tell us readily this much. Adequate inputs and outputs are instead devices out of which (inputs) an afferent lead goes to *each* neuron of the machine, or into which (outputs) an efferent lead comes out of *each* neuron; or, more economically, these leads connect input and output terminals with a *large* number of neurons spread throughout the machine, and connected with all the other neurons so that no relevant information on its behavior is lost. Anatomically, this seems to correspond to some regions of the brain stem for the afferents, for instance, to the thalamus. It is clear that such a device is what can be best desired for many sorts of feed-back operations; to this point we shall return briefly in connection with the learning mechanism.

4. We have excluded, in the discussion made in B, 5, the possibility that a stimulus presented at the input may produce catastrophic, or epileptic behavior. In any real machine constructed with a very large number of elements and connections among these, however, and even with very good

engineering and planning, catastrophic behavior is the very first thing to be expected.

Even assuming ideal starting conditions, the intervention of the M.E. will soon change the values of the $a_{hk}^{(r)}(t)$. The maintenance of a reverberation presupposes that, after its cycle is completed, the same situation repeats identically. Unless the ratios of the numbers and weights (as given by the $a_{hk}^{(r)}(t)$) of output vs. input terminals are kept within very critical limits (cf. the operation of a nuclear reactor), the most likely event to occur is that the initial situation does not repeat exactly after one cycle of reverberation is completed, but more, or fewer neurons are excited than the correct number. In either case a process very similar to a chain reaction might take place immediately, that is, a very fast excitation of all neurons, or a very fast extinction of all activity; more generally, totally uncontrollable phenomena would occur as a rule rather than as an exception.

Whenever learning is involved, and in any case whenever design is not ideally perfect, a controlling mechanism (like the cadmium bars in reactors) is necessary to prevent any such possibility. We saw in Section **3**, c, 2, that we can use, without any loss in the learning capacity of the machine, the neuronic thresholds s_h for this purpose; this is also, clearly, the best choice from a practical point of view.

Any machine that works according to the N.E. and the M.E. necessitates therefore a mechanism which, upon receiving information on its local and general activity at a given time, may alter the thresholds s_h of the neurons, so as to avoid catastrophic conditions at later times. We have described here the function of the reticular system of the brain, as was made clear by the profound physiological investigations of G. Moruzzi & H. W. Magoun. The existence of such a mechanism will provide the *second criterion of stability* for our machine.

It will also be necessary, of course, to resort to the methods of which Nature avails herself in the brain: most of the $a_{hk}^{(0)}(0)$ will be taken as vanishing, in such a way that a neuron be connected through its efferent terminals mostly to rather distant neurons. Thus, excitations spread out and tend to stay below the epileptic thresholds; if couplings were only with "nearest neighbours", epileptic waves would be the only mode of operation of this machine. Such a choice of initial coupling coefficients will provide the *third criterion of stability*. We may include into this the action of inhibitory couplings also.

5. We can now discuss the operation of the machine from the point of view of the M.E.; learning and forgetting will play an equally important rôle. The cortex of the brain has many, more or less specialized input and output "areas"; we shall refer, generically, only to "input" and "output" and refrain from using the current terminology of "sensory", "motor"

and "associative" zones, which presupposes a more detailed structural knowledge than is needed for a qualitative discussion.

We also forgo such obvious things as the convenience that special input devices be constructed so as to perform a preliminary analysis of the figures, or patterns, "shown" them (fed into them; we refer, for concreteness, to a visual input); thus, if a set of homothetic triangles is shown, one may require that the input device transmit to the direct input of the machine only the image of a standardized triangle, plus an information on the value of the homothety parameter. Devices of this sort are not hard to conceive as special organs or extensions of the machine itself, to which they are linked by additional N.E. with fixed coefficients (learning is undesirable at this level). We refer hereafter only to the direct input of the machine, and assume that any such simplification has already been performed somehow.

Without a thalamus, the machine can only learn *by repetition* from habit. Suppose it starts as a *tabula rasa*, i.e. with all couplings having values $a_{hk}^{(r)}(0)$; suppose also, for the sake of simplicity, that a figure may be presented to it any number of times, but each time only for a very short duration. Each presentation will stimulate into firing a number of input neurons which we assume to be sufficiently large to initiate a collective activity, which spreads into the machine as described in A, 1 and B, 3 and 5. Unless the collective motion thus induced is strongly favoured by the genetic memory (i.e. the $a_{hk}^{(r)}(0)$; this ought to be the case for such things as the infant's sucking reflex), nothing much should happen after the virgin machine sees a figure Θ_1 (say, a triangle) for the first time; reverberations will be evoked but, because of the low values of the coefficients $a_{hk}^{(r)}(0)$, a great many pulses are required to cause the firing of each single neuron, so that periodic, or nearly periodic, modes may be expected to involve a great many neurons and to be quite slow. The activity thus induced will be rather *diffuse*, not yet quite specific; it will last for some time $\gg \tau$ and the coefficients will start changing slowly because of the M.E. The pattern of this motion will be altered as soon as this change requires step (ii) as described in 2, p. 220: the increase of the coupling constants will tend to favor the creation of pathways through which reverberations are facilitated, i.e. have shorter cycles and require less neurons. Normal biological disturbances, or mechanical variations, spontaneous (e.g. due to the variations of the coefficients caused by the M.E.) or imposed, e.g. by the reticular system, will cause this state of motion eventually to cease; the chance that a single direct channel be used as many times as is required by the M.E. for a permanent engramming of a facilitation is very low; what alterations have been caused by the sight of Θ_1 will be slowly forgotten by the machine.

Clearly, however, the thing is quite different if Θ_1 is shown many times in succession to the machine: then the semi-permanent changes induced by the M.E. will accumulate, until permanent changes are induced that definitely facilitate the "most convenient" reverberations evoked by Θ_1, i.e. those that are quickest and involve the least number of neurons; these we regard as specific to Θ_1. Permanent engramming is favored if the intervals between the exposures of Θ_1 are made shorter, disfavored otherwise; it may however take a very long time (days, or months, or years in an animal), or never occur at all: semi-permanent memories may decay very slowly indeed, and there may also be more stages, with various decay times, before permanent engramming obtains, than is assumed in the form (5) of the M.E.

It is evident that if a series of random figures is shown in succession to the machine, it will learn nothing, except perhaps a response meaning only that "a figure is being shown" (see later for a further discussion of this fact); but if among these random figures a given one Θ_1 is shown repeatedly enough, the machine will "learn" only Θ_1, together with the general response mentioned before.

This discussion is manifestly incomplete; diminishing response because of assuefaction, for instance, is not considered. The reason we do not wish to push it farther is that it appears too easy, rather than too difficult, to answer such questions at a qualitative level—which may be as dangerous as inconclusive. For instance, both the action of inhibitory couplings (which we have arbitrarily disregarded here for the sake of brevity, although their importance cannot be doubted) and of external inhibitions, e.g. from the reticular system may be invoked to explain assuefaction. Even with these restrictions, we hope that the present discussions of the subject may suffice to provide a convenient basis for further elaborations.

6. Learning by repetition, as described in 5 above, does not appear very satisfactory. One may say that the virgin machine has no more reason to wish to learn a triangle than the infant child. The same mechanism, however, makes it easy to explain "learning by punishment and reward", or *conditioning of the first kind* (as we shall say) and to account why the latter is much faster than the first.

The machine must, of course, be sentient to some degree to know whether it is being punished or rewarded. This will mean here that the thalamus has in-built criteria (homeostatic devices) which enable it to "like" or "dislike" what it records. Suppose that this is the case, and that the thalamus can either suppress, or create a state of excitation in the neurons. Then, even when a stimulus is presented only once at the input, the thalamus may evaluate the situation through its homeostats and determine either a quick suppression of it (or of a part of it, e.g. that

which produces some specific motor pulses), or that it be reinforced and maintained until permanent engramming is achieved, much more quickly in this way, clearly, than by "external" repetition at long or random intervals.

A thinking machine may perhaps do without this device and method of learning, which becomes of utmost importance only when survival has to be fought for.

7. We have discussed in 5 (p. 222) the situation that occurs if a given figure, or pattern, Θ_1 is presented repeatedly at the input. We consider now what can happen if two distinct patterns Θ_1 and Θ_2 are used.

If the machine has already learned Θ_1, i.e. if it responds to Θ_1 with specific reverberations, when Θ_2 is presented for the first time the situation is not the same as at the origin of time; some facilitations are already formed in the connections, so that the machine, as its first reaction, will show all of the reverberations which Θ_2 may evoke in common with Θ_1, plus extra motions which will slowly be changed into a response distinctive of Θ_2. In other words, if the machine knows already Θ_1, and sees Θ_2 for the first time, its only immediate response will be to tell how much Θ_2 has in common with Θ_1; later on it will learn also Θ_2. Thereafter, it will be able to analyze likewise Θ_3 in terms of Θ_1 and Θ_2; and so on.

This is, at its simplest, the mechanism of pattern-analysis, as it develops with the evolution in time of the machine—its "education". The analysis the machine is capable of performing at time t is determined, because of the A.L.H., solely by the N.E., taken with the values their coefficients have at that time.

8. The most typical and distinctive characteristic of the human mind is, in our opinion, its ability to abstract what is "common" to two, or more, situations or patterns, and to retain the result of this operation as a new pattern, which is entrusted to the memory as if learnt from the outside. Just as pattern-analysis was seen to be the fundamental operation of a machine described by N.E. with frozen coefficients, we proceed now to show that *abstraction* is the other fundamental operation performed by a machine which obeys N.E. and M.E. as well. The same discussion will clarify also the exact meaning to be given to the word "common" used above.

At any time t intermediate between the time t_0 at which a pattern Θ_1 is activated into the machine (e.g. a triangle Θ_1 is shown at the input) and the time t_1 at which the collective motion aroused by that act dies out ($t_1 - t \gg \tau$; no other such activations are supposed to take place, and the mechanism described in 7 above is excluded), there will be $n_1(t - t_0)$ neurons which actually fire, and form the *constellation* (set) $c_1(t - t_0)$ engendered by the activation of Θ_1 at t_0. There will be also, in addition, a

constellation $c_1'(t - t_0)$ of $n_1'(t - t_0)$ (in general, $n' > n$) of "penumbral" neurons which *do not fire*, but receive a *subliminal facilitation* from the activation of Θ_1 at t_0.

Suppose now that at some time $t_0'(t_0 < t_0' < t_1)$ another pattern Θ_2 is activated (a different triangle Θ_2 is shown at the input). The machine may have already learnt Θ_1 in a permanent or semi-permanent way, or not. Define likewise $n_2(t - t_0')$ and $c_2(t - t_0')$, etc., and neglect (this is incorrect and could easily be avoided in this discussion without altering its *qualitative* conclusions) the alterations caused by the non-linear interference of these motions at times between t_0' and t in the constellations $c_1'(t - t_0)$ and $c_2'(t - t_0')$; suppose that at least for some t between t_0' and t_1 the intersection of the sets $c_1'(t - t_0)$ and $c_2'(t - t_0')$ is not void and contains a constellation $c_{1,2}(t - t_0')$ of neurons which (would receive only subliminal stimuli from *either* motion, but) *fire* because the non-linear summation of the stimuli from *both* motions exceeds their thresholds: then, $c_{1,2}$ cannot be distinguished from a state of excitation such as would be produced by the presentation (not necessarily at the direct input which this, we recall, has the same structure as the rest of the machine) of a pattern $\Theta_{1,2}$. $\Theta_{1,2}$ can be aroused, clearly, only if Θ_1 *and* Θ_2 are shown at the input, the second after the first, and will result quite differently, in general, from $\Theta_{2,1}$, if the temporal development of both motions and the interference effects between them are treated without the illegitimately oversimplified assumptions which were made here for short.

Although presented here in mere outline on purpose, this line of reasoning makes it evident that, because of the non-linearity of the N.E., whenever a pattern is activated while the response to a previous one has not yet died out, the machine can *abstract* something which is "common" to both (the structure of the machine decides what meaning this word should have) and then adjust itself through the M.E. so as to memorize it, permanently or not, not differently from its normal behavior in response to any other pattern.

"Patterns of patterns" of any sort, in any number, may be formed and learnt in this way: chains of abstractions can take place without limitations other than those imposed by the complexity and structure of the machine. We recognize here, in full, the mechanical analogue to the faculty of abstraction of the human mind.

9. Abstraction alone is not enough: it would be highly uneconomical to remember all single instances, once the general concept is grasped. Our machine takes, in this respect, good care of itself. Suppose that, say, a sufficiently large number a of random triangles Θ_i have been shown to the machine, which has learnt them semi-permanently; it has also formed the responses $\Theta_{1,2}$, $\Theta_{2,1}$, $\Theta_{1,2,3}$, ..., $\Theta_{1,2,3\ldots a} \equiv \Theta$ which are common to all sub-

sets of triangles (ordered or not) and memorized these semi-permanently. This memorization is accompanied by a process of facilitation; whereas, before memorization, Θ could be evoked only if *all* the triangles were shown to the machine, after facilitation, the $a_{hk}^{(r)}$ that appear in the N.E. which describe the behavior of the neurons of the constellation $c_{1,2,\ldots,a} \equiv c$, have increased their values considerably; suppose, for maximum simplicity, these values to have become so large that each single motion Θ_i gives now to the neurons of c, instead of a subliminal stimulus, an excitation above threshold (if more than one, but less than a, Θ_i were involved at this stage, our argument would only take a few more steps). Then each time *any* triangle Θ_i is presented to the machine, the common response Θ (which can convey, clearly, only the "general concept" of triangle) is always evoked, i.e. the neurons of the "common" constellation c fire and their channels are facilitated some more; on the average, the couplings of each constellations c_i will be susceptible to increase for only $1/a$ of the total time, while those of c will be exposed to facilitating actions for *all* the time—until permanent memory is achieved. Even if at the beginning, say, c_1 had been markedly facilitated, the presentation of more and more patterns which have "something in common" with it (as decided by the machine) will cause the specific response to c_1 to fade into oblivion, while the "abstract", or "common" response is evoked as the first thing, after a convenient learning period.

If we consider now only two stimuli Θ_1 and Θ_2 and assume that Θ_1 *causes* already, say, because of the genetic structure of the machine, *a direct response* (e.g. the food-salivation reflex of Pavlov's dogs), while Θ_2 is ineffective (cf. bell-ringing), then the same mechanism can be clearly extended to account first for the formation of the response $\Theta_{2,1}$ (bell before food, $\neq \Theta_{1,2}$), then, for the fact that Θ_2 alone comes to provoke the same effect as Θ_1 or Θ_2. The temporal behavior given by M.E. of type (5) is perfectly suited to describe the available evidence, which might be used for a determination of the numerical values of some of the constants which appear in (5).

This type of conditioning is manifestly different from that described in 6 above, which requires the intervention of the thalamus: we shall call it *conditioning of the second kind*, or "by information" (the bell just tells the dog that food is coming).

10. We discuss, finally, *re-integration*, which we define as the fact that our machine shall, after learning a pattern, respond to the "incomplete" presentation of that pattern as if it were complete (cf. the familiar oversights of the proofreader). That this must be the case can be seen now quite trivially: after a number of facilitations have occurred, a smaller number of input stimuli will be required to cause the firing of neurons

than were necessary when that pattern was presented for the first time. This also accounts for the children's alterations of new words, which are reduced to combinations of already familiar words; in part, for the fact that a cue suffices to evoke a long string of memories, e.g. verses, etc.

5. Concluding Remarks

A. TIME EVOLUTION OF PATTERN-ANALYSIS

1. The preceding discussion has been, at various places, restricted to situations in which a stimulation, or presentation of a pattern, at the input occurs at given instant, after which it ceases while the machine starts its analysis of it; that is, we have chosen to consider only the "free" modes of motion of the machine, of which the stimulation sets the initial conditions, rather than the evidently prevalent situations in which the machine will perform "forced" motions under the *continued* influence of stimulations which persist and may vary with time.

This simplification is obviously convenient for the purposes of a purely qualitative discussion, as it permits separate examination of the various features of the operation of the machine; nor can forced motions be adequately described without a quantitative analysis of the solutions of the N.E. and M.E. We wish to point out here, however, that this simplifying assumption is, in all likelihood, a much better approximation of reality than it may seem at first.

We can consider a continued stimulation as the presentation to the machine of a *time-series of patterns*; its analysis by the machine is therefore a *serial* operation. Our observation is, that a machine such as the one envisaged by us, with a very large number of elements, can actually transform that serial operation into a *parallel* operation.

Let the patterns presented at the input (or anywhere by the machine to itself: cf. Section **4**, C, 8) at time $0, \tau, 2\tau, 3\tau \ldots$ be $\Theta_0, \Theta_1, \Theta_2 \ldots$. We recall (Section **3**, B, 2) that each neuron, as described by the N.E., has a *refractory* period which is $> \tau$, and may be $\gg \tau$; we take it here to be $R\tau$, according to the schematization expressed by (4). Θ_0 causes a set S_0 of neurons to fire at time 0; all neurons of S_0 then stay dead for $R\tau$ sec, while other neurons are excited by them at rather distant places (3rd criterion of stability, Section **4**, C, 4) and then remain dead in turn while stimulating other distant neurons, etc. When Θ_1 is presented at time τ, all neurons of S_0 *cannot respond*, and another set S_1 *disjoint* from S_0 fire and excite likewise distant neurons, etc.; and so on for Θ_2 at time 2τ, etc.

All the patterns presented from t_0 until $t = R\tau$ are therefore registered *in parallel* by the machine, which can thus, for example, abstract a con-

cept meaning "motion" with the same mechanism by which it abstracts one meaning "triangularity".

The extent to which this happens is determined, in this simple example, by the value of R. If the reverberations which are significant for pattern-analysis have periods $< R\tau$, then the simplifying assumption made in Section **4** is quite good. In any case, the conversion of serial into parallel operation, and *vice versa*, will emerge as an obvious and important feature in any quantitative discussion of the N.E.

2. It is convenient at this stage to return briefly to the matters mentioned in Section **4**, C, 2, so as to summarize the essentials of the operation of the machine into a few mathematical concepts.

Our machine learns, by the process of abstraction (by virtue of the M.E.), *to perform pattern-analysis* (by virtue of the N.E.). This sentence contains all that is most relevant in our theory; mathematically, it can be expressed as follows.

At a fixed time t, the N.E. have frozen coefficients: A.L.H. A pattern Θ presented at or into the machine at t evokes, in general, a very large number of disjoint modes, or reverberations: this is the "Fourier" analysis performed by the machine, each mode corresponding to a point $\mathcal{Q}$ or an axis in an appropriate functional space $\mathcal{F}$ on $\mathcal{N}$ (the quotes call to mind the profound difference from linear Fourier analysis).

We may also say that each point $\mathcal{Q}$ of $\mathcal{F}$ corresponds to one of the *basic concept*, or *words*, which the machine has learnt until t: the N.E. contain, implicitly, all the *knowledge* or *vocabulary* of the machine, in terms of which each new pattern is translated by the machine. (This vocabulary we expect to have a surprisingly different structure from those of Western languages, in which a word is a "point", and a sentence or definition is a "surface constructed point-by-point"; more akin to Chinese or Japanese, in which a word is a "plane" and a sentence or definition a "surface constructed as envelope of planes".)

The second difference from Fourier analysis (the first being non-linearity) is that now the set of fundamental modes, or axes in $\mathcal{F}$, is not *constant*, but *changes in time* because of the M.E. Thus, at a given t a pattern is represented by a point in a given frame S in $\mathcal{F}$ (N.E.); this frame, however, is not fixed, but changes slowly (A.L.H.) in time (M.E.): $S = S(t)$.

This scheme, if correct, has profound implications: for instance, the efficiency of the machine will depend tremendously on the method followed in its education, because to the same external stimulus machines which have been educated differently may offer very quick and simple, or very slow and involved responses.

3. An interesting consequence of our theory is the fact that a machine

thus constructed necessitates periods of rest, or "sleep". Indeed, even giving it a tremendously large number of elements, the long duration of reverberations (and induced motions in general), which goes far beyond that of the actual stimulations, will cause, if stimuli are unceasingly offered to it, an ever increasing cumulation of activity; the possibility of co-existence of disjoint reverberations will become less and less, interference will cause "confusion of ideas" and inability to give correct answers even to familiar questions.

A period of "sleep", i.e. cessation of activity through the suave interaction of the reticular system, will permit the gradual extinction first of the less facilitated, then of all other reverberations, and the fading of semi-permanent memories of relatively short duration as well. A more quick and drastic treatment, such as "electroshock", followed by total quiescence for only the period of time which is required for such semi-permanent memories to disappear, will produce the same effect. This reminds one of some Yogi techniques which are said to achieve, in a relatively short time, the same state of rest which follows a full night's sleep.

4. As a final remark, we wish to point out that it might be interesting, in the light of the present considerations, to attempt an analysis of E.E.G.'s in terms of subharmonics instead of the customary one in terms of harmonics of a suitably chosen frequency. The *amplitude* of the E.E.G. recordings should depend strongly on the mechanism described in 5, A, 1

B. SELF-ORGANIZATION INTO RELIABLE OPERATION

1. In the study of any system containing a very large number of interacting elements built with realistic tolerances—let this be the machine which is being discussed here, or a bee-hive, or the whole socio-political and administrative framework of a nation—the central question is certainly whether this system obeys instantaneous and evolutionary laws which guarantee its spontaneous *convergence in time towards more efficient operation*, or whether the system may rather show *erratic* or *divergent* performance. This question was formulated with full clarity by N. Wiener, who, besides emphasizing the vital importance of it, gave also powerful mathematical tools for its investigation; in his treatment the non-linearity of the interaction laws was rightly stressed as the key to the whole problem.

We can do nothing better than refer the reader to his work for a deeper elaboration of these ideas; it is important, however, here, to show that there is very satisfactory qualitative evidence that a machine obeying *suitable* N.E. and M.E. will satisfy Wiener's principle of self-organization. We shall keep this discussion at a qualitative level by resorting to physical more than to mathematical arguments, and by examining later in detail,

instead of the actual machine proposed here, a simplified functional model of it.

That non-linearity of some sort is necessary for self-organization is physically obvious. A linear system (with *frozen* coefficients, or else non-linearity intervenes: cf. A, 2 above) cannot perform a spontaneous transition from one of its states to another (we use the language of quantum mechanics only because of its greater appeal to intuition; of course, the same is true classically); such transitions—without which the system could not choose spontaneously, from our point of view, states with a better (or worse) organization—are only possible if there are *perturbations* to cause them (e.g. the interaction with the electromagnetic field causes the quantum jumps in atoms); furthermore, *the system itself* must originate these perturbations (the electrons of the atom are the source of the electromagnetic field in spontaneous emission), or else its changes of state would be "induced" rather than "spontaneous". The equations of the system must therefore be non-linear in an essential way; such, for instance, as we have in electrodynamics when the electromagnetic field is expressed in terms of electron sources.

Only a non-linear system of this sort can change its state spontaneously; it will not be generally true, however, that its changes are necessarily "for the better". Again, simple examples suffice to prove this statement; leaving aside those, plentiful indeed, which come from societies or civilizations that go bankrupt, we observe that the behavior of a non-linear physical system the energy of which is not restricted by a finite lower bound is certainly catastrophic. We expect therefore that such a system, in order to satisfy Wiener's principle, shall obey additional *necessary* "convergence" conditions—of the type, for instance, that in quantum mechanics secures the existence of a ground state, which we may identify with the "best state for efficient operation". It is our belief that these conditions will amount to satisfying the third criterion of stability (Section 4, C, 4) and to choosing N.E. with the qualitative behavior exhibited by (5).

The study of the restrictions that this criterion would impose on N.E. (2) and M.E. (5) appears to be a rather straightforward problem, which we add to the list of those which we formulated and set aside in Section 4 for future investigation. A dynamical interpretation of (2) and (5) would be indeed quite natural, as the M.E. (5) just express a non-linear coupling of the N.E. (2) with themselves; we deem it more meritorious, at this stage, to resist the temptation of adapting the available quantum-field-theoretical knowledge to these problems, than to yield to it.

Clearly, the A.L.H. imposes the distinction between two types of non-linearity: that which is expressed by the N.E. (2) and that which is

expressed by the M.E. (5). The first is useful, but only the second is necessary for the type of self-organization we are discussing; this statement will appear obvious after the discussion in 2 below, which is dedicated, as was announced, to a simplified version of our problem.

2. We wish to discuss here a *model* of our machine (which we may also regard as another, though less sophisticated, model of the brain; as such it was briefly discussed at an early stage of our work) which consists in a system of N linearly coupled harmonic oscillators, the constants of which obey M.E. of type (5) ($r = o$), under the A.L.H. This was mentioned in Section **4**, B, 3, and is a model of our machine in the sense that, as was said there, "reverberations" correspond in it to normal modes, "subharmonics" to ordinary harmonics, etc. The wealth of solutions of N.E. of type (2) is now lost, because of the linearity of the N.E. of this model; for it the central problem reduces to the determination of the solutions of a secular equation of degree N:

$$f(a_{hk}; \lambda) = 0 \tag{6}$$

This digression is useful both because it readily provides a qualitative insight into the behavior of a learning machine with respect to self-organization, and because it shows that many different mechanisms may be built, with various degrees of convenience, to produce "thought": the only essential thing is that they obey *some* N.E., *some* M.E. (*suitable*, but not necessarily of type (2) and (5)), and closely enough the A.L.H.

We wish to compute the average change $< \delta\bar{\lambda} >$ of a solution $\bar{\lambda}$ of (6), when the coefficients a_{hk} of the equations of the system undergo infinitesimal random variations (from a "macroscopic" point of view, small variations due to learning will appear as "random"):

$$a_{hk} \to a_{hk}(1 + \rho_{hk}) = a_{hk} + \delta a_{hk} \tag{7}$$

such that $< \rho_{hk} > = \rho$.

From:

$$f(a_{hk} + \delta a_{hk}; \bar{\lambda} + \delta\bar{\lambda})$$

$$= \sum_{hk} a_{hk}\rho_{hk} \frac{\partial f}{\partial a_{hk}} + \left(\frac{\partial f}{\partial \lambda}\right)_{\lambda = \bar{\lambda}} \delta\bar{\lambda} = 0 \tag{8}$$

we find, on taking averages (since (6) is homogeneous of degree zero in the term λ^N and of degree one in all other terms, with respect to the variables a_{hk}):

$$< \delta\bar{\lambda} > = \frac{(\bar{\lambda})^N}{f'(\bar{\lambda})} \rho \tag{9}$$

which tells us several interesting things. It requires, first of all, that $f'(\bar{\lambda}) \neq 0$, or else $\bar{\lambda}$ would be a degenerate eigenvalue, and even with

$\rho = 0$ the degeneracy might be removed just the same by (7). We suppose therefore $f'(\bar{\lambda}) \neq 0$; to gain some insight into the behavior of $< \delta\bar{\lambda} >$ we just suppose here that the roots of (6) are all simple and equally spaced, so that

$$\bar{\lambda}_k = k\lambda_0, \quad (k = 1, 2, 3, \ldots N) \tag{10}$$

then (9) gives:

$$\left| \frac{< \delta\bar{\lambda}_k >}{\bar{\lambda}_k} \right| = \frac{1}{(N-1)!} \binom{N-1}{k-1} k^{N-1} \rho \tag{11}$$

whence, for $k = 1$ and $k = N$ $(N \gg 0)$:

$$\left| \frac{< \delta\bar{\lambda}_1 >}{\bar{\lambda}_1} \right| = \frac{1}{(N-1)!} \rho \tag{12}$$

$$\left| \frac{< \delta\bar{\lambda}_N >}{\bar{\lambda}_N} \right| \sim e^N \rho \tag{13}$$

For our qualitative purposes, (12) and (13) suffice amply to show that random variations, due to learning as expressed by M.E. of type (5), may alter by a vanishing amount the smaller, by increasingly relevant amounts the larger eigenfrequencies of the normal modes of this model.

If a frequency which is kept long enough unchanged becomes permanent, this model will evolve therefore with learning so as to preserve the modes with small frequency; the modes with higher frequency, as well as those that correspond to degenerate solutions, will change at random without staying long enough at any given value to become permanent. After a "long" time, the system will have shifted spontaneously, because of its learning ability, toward a "ground state", in which there is no degeneracy, but the allowed frequencies stay as close to one another as the maximum and minimum values conceded by the M.E. to the constants will permit; this is the "senile age" of the system, in which no learning is possible because all available memories are already engrammed permanently (cf. (7): all $\delta a_{hk} \equiv 0$). The "infancy" of the system is characterized instead by $\rho > 0$, because for $t > 0$ (when all the $a_{hk}(0)$ have minimum absolute values) they can only increase monotonically: learning is somewhat slower, degeneracies are removed faster. The "adult" age corresponds to the period in which it is mostly $\delta a_{hk} \neq 0$, $\rho = 0$.

This cursory and incomplete glance at the properties of this model is intended only to show in which sense we should expect Wiener's principle to be verified by it: the "reliable" information is that carried by the small eigenfrequencies; when the erratic behavior of the higher frequencies pushes one of them down enough, it may be permanently "engrammed", increasing thus the reliability of the system.

3. The main differences between the machine we study and the model of it discussed in 2 above lie, as regards the validity of Wiener's principle, in the non-linearity of the N.E. (2) and in the 2nd criterion of stability (Section **4**, C, 4). The situation is made worse to some extent than in (2) by this non-linearity; in any case, though, by assuming the number of elements to be large enough, one ought to obtain a preferential decrease of the effect of random learning variations on the faster reverberations (to higher frequencies of the model described in 2 slower reverberations of the machine now correspond), so as to reproduce, although perhaps less dramatically, a situation like that represented in 2 by (12) and (13).

The existence of a thalamus and of a reticular system (2nd criterion of stability) makes instead a tremendous improvement upon the model discussed in 2, which could still satisfy Wiener's principle even though it is not endowed with these homeostatic devices. Conditioning of the first kind and arousal of attention by the thalamus, prevention of too diffuse (and therefore not very meaningful) reverberations by the reticular system, are only examples, clearly, of what controlling devices of such effectiveness may do in the way of forcing the machine into learning important information in a reliable manner. These devices are especially important if the memory is prevalently of *negative*, rather than of *positive* type as is assumed here (Section **3**, C, 1), because then the machine would have, in its infancy, a tendency towards epilepsy.

Of importance in a discussion of this principle are also, clearly, all the additional special-purpose devices that Nature uses. The machine also is better with these devices, as intermediate links between external inputs and outputs, and inputs and outputs to the machine proper or cortex; their consideration, however, belongs to engineering more than to physics, and would not be relevant at this place. In conclusion, we should like to stress once more our firm conviction that the speediest way to progress in all the problems connected with the actual construction of machines of this sort is humble resort to Nature's own doings, through neuroanatomy and physiology.

C. FURTHER OUTLOOKS

As a final comment, we think it appropriate to remark that the general formalism of N.E., M.E. and A.L.H. which is expounded here seems to us to admit of a far wider range of applicability than that to which it has been restricted in this work. The N.E. in fact, for instance, serve only to express, in a more or less schematic manner, the fact that a decision is taken, after a weighted evaluation of the information which lasts a finite time, by a member of a set; and that such a decision is bound to affect other decisions, etc. We may change their name into that of *decision equations*, call the

M.E. *evolution equations*, and take all our considerations over to the study of social or economical or other collective phenomena.

We have pursued this line of thought in several directions for personal amusement, and have soon found, to our surprise, that the qualitative analysis given here for thought-processes applies as well, *mutatis mutandis* (that is, names), to a great many other instances. We believe that, as soon as it becomes possible to agree on a concrete choice of schemes and numbers, quite reasonable predictions may be made in this way about, say, the operation of a stock-exchange, the variation in time of a parameter in feminine fashion, the type of national government that would best obey Wiener's principle, and so on. This we say with at least the same degree of assurance that we have found in the economists who apply the Schrödinger equation to the study of their problems.

Although we have refrained here from a quantitative analysis of the several mathematical problems formulated in the course of this work, the results we have already obtained in this direction seem to justify some optimism; if these expectations are not illusory, then the present formalism might help us to gain a finer knowledge of some physical phenomena that can now be treated only with statistical methods.

This research would not have been possible without the generous and enthusiastic collaboration of Dr. V. Braitenberg, neuroanatomist, among whose merits were certainly not least patience toward this writer's initial ignorance and presumption, and success in eliminating the latter; of Dr. F. Lauria, a young mathematician who dared to place mathematics not too high above common brains; and of many others, to all of whom it is our duty and pleasure to extend our sincerest thanks.

REFERENCES

BEURLE, R. L. (1956). *Phil. Trans.* **B669,** 55.

BRAITENBERG, V. & LAURIA, F. *Nuovo Cim.* Suppl. (In press).

BRAITENBERG, V., CAIANIELLO, E. R., LAURIA, F. & ONESTO, N. (1959). *Nuovo Cim.* **11,** 278.

MCCULLOCH, W. S. & PITTS, W. (1943). *Bull. math. Biophys.* **5**.

MORUZZI, G. & MAGOUN, H. W. (1949). *Electroenceph. clin. Neurophysiol.* **1,** 455.

WIENER, N. (1958). "Nonlinear Problems in Random Theory". Massachusetts Institute of Technology.