MOLECULAR ETHOLOGY

An Immodest Proposal for Semantic Clarification

Heinz Von Foerster

Departments of Biophysics and Electrical Engineering
University of Illinois
Urbana, Illinois

I. INTRODUCTION

Molecular genetics is one example of a successful bridge that links a phenomenology of macroscopic things experienced directly (a taxonomy of species; intraspecies variations; etc.) with the structure and function of a few microscopic elementary units (in this case a specific set of organic macromolecules) whose properties are derived from other, independent observations. An important step in building this bridge is the recognition that these elementary units are not necessarily the sole constituents of the macroscopic properties observable in things, but are determiners for the synthesis of units that constitute the macroscopic entities. Equally helpful is the metaphor which considers these units as a "program," and the synthesized constituents in their macroscopic manifestation as the result of a "computation," controlled and initiated by the appropriate program. The genes for determining blue eyes are not blue eyes, but in blue eyes one will find replicas of genes that determine the development of blue eyes.

Stimulated by the success of molecular genetics, one is tempted to search again for a bridge that links another set of macroscopic phenomena, namely the behavior of living things, with the structure and function of a few microscopic elementary units, most likely the same ones that are responsible for shape and organization of the living organism. However, "molecular ethology" has so far not yet been blessed by success, and it may be worthwhile to investigate the causes.

One of these appears to be man's superior cognitive powers in dis-

criminating and identifying forms and shapes as compared to those powers which allow him to discriminate and identify change and movement. Indeed, there is a distinction between these two cognitive processes, and this distinction is reflected by a difference in semantic structure of the linguistic elements which represent the two kinds of apparitions, namely different nouns for things distinct in form and shape, and verbs for change and motion.

The structural distinction between nouns (cl_i^k) and verbs (v_i) becomes apparent when lexical definitions of these are established. Essentially, a noun signifies a class (cl^1) of objects. When defined, it is shown to be a member of a more inclusive class (cl^2), denoted also by a noun which, in turn, when defined is shown to be a member of a more inclusive class (cl^3), etc., [pheasant → bird → animal → organism → thing]. We have the following scheme for representing the definition paradigm for nouns:

$$cl^n = \{cl^{n-1}_{i_{n-1}}\{cl^{n-2}_{i_{n-2}}\{\cdots\{cl^m_{i_m}\}\}\}\} \tag{1}$$

where the notation $\{\epsilon_i\}$ stands for a class of elements ϵ_i ($i = 1, 2, \ldots, p$), and subscripted subscripts are used to associate these subscripts with the appropriate superscripts. The highest order n in this hierarchy of classes is always represented by a single undefined term "thing," "entity," "act," etc., which appeals to basic notions of being able to perceive at all. A graphic representation of the hierarchical order of nouns is given in Fig. 1 and a more detailed discussion of the properties of these (inverted) "noun-chain-trees" can be found elsewhere (Weston, 1964; Von Foerster, 1967a).

Essentially, a verb (v_i) signifies an action, and when defined is given by a set of synonyms $\{v_j\}$, by the union or by the intersection of the meaning of verbs denoting similar actions. [hit → {strike, blow, knock} →

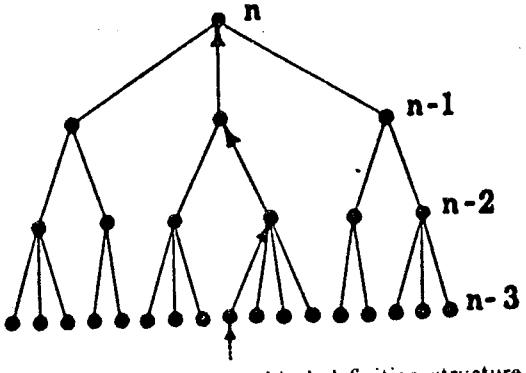


Fig. 1. Ascending hierarchical definition structure for nouns. (Nouns are at nodes; arrow heads: definiens; arrow tails: definiendum.)

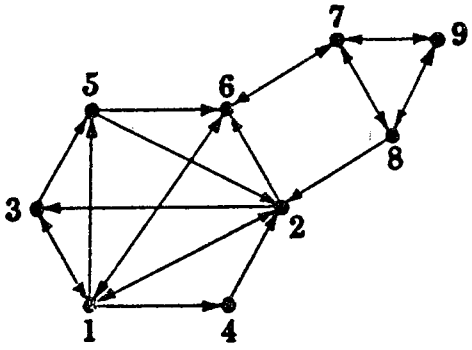


Fig. 2. Closed heterarchical definition structure for verbs. (Verbs are at nodes; arrow heads: definiens; arrow tails: definiendum.)

{(hit, blow, ...) (stir, move air, sound, soothe, lay eggs, ..., boast) (strike, blow, bump, collide ...)} → etc.]

$$v_i = \{v_j\} \vee \sum v_k \vee \prod v_e \qquad (2)$$

A graphic representation of this basically closed heterarchical structure is given in Fig. 2, and its corresponding representation in form of finite matrices is discussed elsewhere (Von Foerster, 1966).

The essential difference in the cognitive processes that allow for identification of forms and those of change of forms is not only reflected in the entirely different formalisms needed for representing the different definition structures of nouns [Eq. (1)] and of verbs [Eq. (2)], but also by the fact that the set of invariants that identify shape under various transformations can be computed by a single *deductive* algorithm (Pitts and McCulloch, 1947), while identification of even elementary notions of behavior requires *inductive* algorithms that can only be computed by perpetual comparison of present states with earlier states of the system under consideration (Von Foerster *et al.*, 1968).

These cognitive handicaps put the ethologist at a considerable disadvantage in developing a phenomenology for his subject matter when compared to his colleague the geneticist. Not only are the tools of expressing his phenomena devoid of the beautiful isomorphism which prevails between the hierarchical structures of all taxonomies and the definition of nouns that describe them, but, he may fall victim to a semantic trap which tempts him to associate with a conceptually isolable function a corresponding isolable mechanism that generates this function. This temptation seems to be particularly strong when our vocabulary suggests a variety of conceptually separable higher mental faculties as, for instance "to learn," "to remember," "to perceive," "to recall," "to predict," etc.,

and the attempt is made to identify and localize within the various parts of our brain the mechanisms that learn, remember, perceive, recall, predict, etc. The hopelessness of a search for mechanisms that represent these functions in isolation does not have a physiological basis as, for instance, "the great complexity of the brain," "the difficulty of measurement," etc. This hopelessness has a purely semantic basis. Memory, for instance, contemplated in isolation is reduced to "recording," learning to "change," perception to "input," and so on. In other words, in separating these functions from the totality of cognitive processes, one has abandoned the original problem and is now searching for mechanisms that implement entirely different functions which may or may not have any semblance to some processes that are subservient to the maintenance of the integrity of the organism as a functioning unit (Maturana, 1969).

Consider the two conceivable definitions for memory:

(a) An organism's potential awareness of past experiences.
(b) An observed change of an organism's response to like sequences of events.

While definition A postulates a faculty (memory$_A$) in an organism whose inner experience cannot be shared by an outside observer, definition B postulates the same faculty (memory$_A$) to be operative in the observer only—otherwise he could not have developed the concept of "change"—but ignores this faculty in the organism under observation, for an observer cannot "in principle" share the organism's inner experience. From this follows definition B.

It is definition B which is generally believed to be the one which obeys the ground rules of "the scientific method," as if it were impossible to cope scientifically with self-reference, self-description, and self-explanation, i.e., closed logical systems that include the referee in the reference, the descriptor in the description, and the axioms in the explanation.

This belief is unfounded. Not only are such logical systems extensively studied (e.g., Gunther, 1967; Löfgren, 1968), but also neurophysiologists (Maturana *et al.*, 1968), experimental psychologists (Konorski, 1962), and others (Pask, 1968; Von Foerster, 1969) have penetrated to such notions.

These preliminaries suggest that the explorer of mechanisms of mentation has to resolve two kinds of problems, only one of which belongs to physiology or, as it were, to physics; the other one is that of semantics. Consequently, it is proposed to reexamine some present notions of learning and memory as to the category to which they belong, and to sketch a conceptual framework in which these notions may find their proper place.

The next section, "Theory," reviews and defines concepts associated with learning and memory in the framework of a unifying mathematical

formalism. In the Section III various models of interaction of molecules with functional units of higher organization are discussed.

II. THEORY

A. General Remarks

Since we have as yet no comprehensive theory of behavior, we have no theory of learning and, consequently, no theory of memory. Nevertheless, there exists today a whole spectrum of conceptual frameworks ranging from the most naive interpretations of learning to the most sophisticated approaches to this phenomenon. On the naive side, "learning" is interpreted as a change of ratios of the occurrence of an organism's actions which are predetermined by an experimenter's ability to discriminate such actions and his value system, which classifies these actions into "hits" and "misses." Changes are induced by manipulating the organism through electric shocks, presentations of food, etc., or more drastically by mutilating, or even removing, some of the organism's organs. "Teaching" in this frame of mind is the administration of such "reinforcements" which induce the changes observed on other occasions.

On the sophisticated side, learning is seen as a process of evolving algorithms for solving categories of problems of ever-increasing complexity (Pask, 1968), or of evolving domains of relations between the organism and the outside world, of relations between these domains, etc. (Maturana, 1969). Teaching in this frame of mind is the facilitation of these evolutionary processes.

Almost directly related to the level of conceptual sophistication of these approaches is their mathematical naiveté, with the conceptually primitive theories obscuring their simplicity by a smoke screen of mathematical proficiency, and the sophisticated ones failing to communicate their depth by the lack of a rigorous formalism. Among the many causes for this unhappy state of affairs one seems to be most prominent, namely, the extraordinary difficulties that are quickly encountered when attempts are made to develop mathematical models that are commensurate with our epistemological insight. It may require the universal mind of a John von Neumann to give us the appropriate tools. In their absence, however, we may just browse around in the mathematical tool shop, and see what is available and what fits best for a particular purpose.

In this paper the theory of "finite state machines" has been chosen as a vehicle for demonstrating potentialities and limitations of some concepts in theories of memory, learning, and behavior mainly for two reasons. One is that it provides the most direct approach to linking a system's

external variables as, e.g., stimulus, response, input, output, cause, effect, etc., to states and operations that are internal to the system. Since the central issue of a book on "molecular mechanisms in memory and learning" must be the development of a link which connects these internal mechanisms with their manifestations in overt behavior, the "finite state machine" appears to be a useful model for this task.

The other reason for this choice is that the interpretations of its formalism are left completely open, and may as well be applied to the animal as a whole; to cell assemblies within the animal; to single cells and their operational modalities, for instance, to the single neuron; to subcellular constituents; and, finally, to the molecular building blocks of these constituents.

With due apologies to the reader who is used to a more extensive and rigorous treatment of this topic, the essential features of this theory will be briefly sketched to save those who may be unfamiliar with this formalism from having to consult other sources (Ashby, 1956; Ashby, 1962; Gill, 1962).

B. Finite State Machines

1. *Deterministic Machines*

Essentially, the theory of finite state machines is that of computation. It postulates two finite sets of external states called "input states" and "output states," one finite set of "internal states," and two explicitly defined operations (computations) which determine the instantaneous and temporal relations between these states.*

Let $\mathfrak{X}_i$ $(i = 1, 2, \ldots, n_x)$ be the n_x receptacles for inputs x_i each of which can assume a finite number, $v_i > 0$, of different values. The number

* Although the interpretation of states and operations with regard to observables is left completely open, some caution is advisable at this point if these are to serve as mathematical models, say, for the behavior of a living organism. A specific physical spatiotemporal configuration which is identifiable by the experimenter who wishes that this configuration be appreciated by the organism as a "stimulus" cannot *sui modo* be taken as "input state" for the machine. Such a stimulus may be a stimulant for the experimenter, but be ignored by the organism. An input state, on the other hand, cannot be ignored by the machine, except when explicitly instructed to do so. More appropriately, the distribution of the activity of the afferent fibers has to be taken as an input, and similarly, the distribution of activity of efferent fibers may be taken as the output of the system.

of distinguishable input states is then

$$X = \prod_{i=1}^{n_x} v_i \tag{3}$$

A particular input state $x(t)$ at time t (or x for short) is then the identification of the values x_i on all n_x input receptacles $\mathfrak{X}_i$ at that "moment":

$$x(t) \equiv x = \{x_i\}_t \tag{4}$$

Similarly, let $\mathfrak{Y}_j$ ($j = 1, 2, \ldots, n_y$) be the n_y outlets for outputs y_j, each of which can assume a finite number, $v_j > 0$, of different values. The number of distinguishable output states is then

$$Y = \prod_{j=1}^{n_y} v_j \tag{5}$$

A particular output state $y(t)$ at time t (or y for short) is then the identification of the values y_i on all n_y outlets $\mathfrak{Y}_j$ at that "moment":

$$y(t) \equiv y = \{y_i\} \tag{6}$$

Finally, let Z be the number of internal states z which, for this discussion (unless specified otherwise), may be considered as being not further analyzable. Consequently, the values of z may just be taken to be the natural numbers from 1 to Z, and a particular output state $z(t)$ at time t (or z for short) is the identification of z's value at that "moment":

$$z(t) \equiv z \tag{7}$$

Each of these "moments" is to last a finite interval of time, Δ, during which the values of all variables x, y, z are identifiable. After this period, i.e., at time $t + \Delta$, they assume values $x(t + \Delta)$, $y(t + \Delta)$, $z(t + \Delta)$ (or x', y', z' for short), while during the previous period $t - \Delta$ they had values $x(t - \Delta)$, $y(t - \Delta)$, $z(t - \Delta)$ (or x^*, y^*, z^* for short).

After having defined the variables that will be operative in the machine we are now prepared to define the operations on these variables. These are two kinds and may be specified in a variety of ways. The most popular procedure is first to define a "driving function" which determines at each instant the output state, given the input state and the internal state at that instant:

$$y = f_y(x, z) \tag{8}$$

Although the driving function f_y may be known and the time course of input states x may be controlled by the experimenter, the output states y as time goes on are unpredictable as long as the values of z, the internal

states of the machine, are not yet specified. A large variety of choices are open to specify the time course of z as depending on x, on y, or on other newly to be defined internal or external variables. The most profitable specification for the purposes at hand is to define z recursively as being dependent on previous states of affairs. Consequently, we define the "state function" f_z of the machine to be:

$$z = f_z(x^*, z^*) \tag{9a}$$

or alternately and equivalently

$$z' = f_z(x, z) \tag{9b}$$

that is, the present internal state of the machine is a function of its previous internal state and its previous input state; or alternately and equivalently, the next internal machine state is a function of both its present internal and input states.

With the three sets of states $\{x\}$, $\{y\}$, $\{z\}$ and the two functions f_y and f_z, the behavior of the machine, i.e., its output sequence, is completely determined if the input sequence is given.

Such a machine is called a sequential, state-determined, "nontrivial" machine and in Fig. 3a the relations of its various parts are schematically indicated.

Such a nontrivial machine reduces to a "trivial" machine if it is insensitive to changes of internal states, or if the internal states do not change (Fig. 3b):

$$z' = z = z_0 = \text{constant} \tag{10a}$$

$$y = f_y(x, \text{constant}) = f(x) \tag{10b}$$

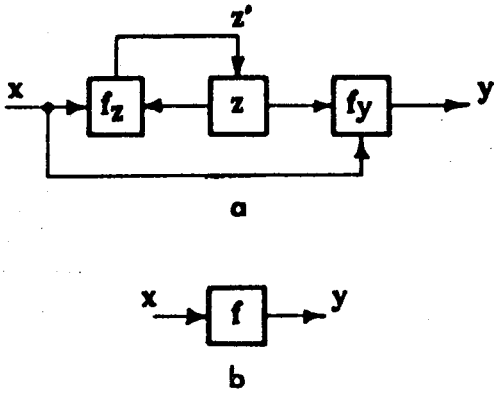


Fig. 3. Signal flow in a finite state machine (a); input–output relation in a trivial machine (b).

In other words, a trivial machine is one which couples deterministically a particular input state with a specific output state or, in the language of naive reflexologists, a particular stimulus with a specific response.

Since the concept of "internal states" is crucial in appreciating the difference between a trivial and a nontrivial machine, we shall now give various formal interpretations of these states to lift them from the limbo of "being not further analyzable."

First, it may appear that by an artifice one can get rid of these mysterious states by defining the driving function f_y in a recursive form. However, as we shall see shortly, these states reappear in just another form.

Consider the driving function [Eq. (8)] at time t and one step later $(t + \Delta)$:

$$y = f_y(x, z)$$
$$y' = f_y(x', z') \tag{8'}$$

and assume there exists an "inverse function" to f_y:

$$z = \phi_y(x, y) \tag{11}$$

We now enter the state function [Eq. (9b)] for z' into Eq. (8') and replace z by Eq. (11):

$$y' = f_y(x', f_z(x, \phi_y(x, y))) = F_y^{(1)}(x', x, y) \tag{12}$$

or alternately and equivalently

$$y = F_y^{(1)}(x, x^*, y^*) \tag{13}$$

However, y^* is given recursively through Eq. (13)

$$y^* = F_y^{(1)}(x^*, x^{**}, y^{**}) \tag{13*}$$

and inserting this into Eq. (13) we have

$$y = F_y^{(2)}(x, x^*, x^{**}, y^{**})$$

and for n recursive steps

$$y = F_y^{(n)}(x, x^*, x^{**}, x^{***} \cdots x^{(n)*}, y^{(n)*}) \tag{14}$$

This expression suggests that in a nontrivial machine the output is not merely a function of its present input, but may be dependent on the particular sequence of inputs reaching into the remote past, and an output state at this remote past. While this is only to a certain extent true—the "remoteness" is carried only over Z recursive steps and, moreover, Eq. (14) does not uniquely determine the properties of the machine—this dependence of the machine's behavior on its past history should not tempt one to project into this system a capacity for memory, for at best it may

look upon its present internal state which may well serve as *token* for the past, but without the powers to recapture for the system all that which has gone by.

This may be most easily seen when Eq. (13) is rewritten in its full recursive form for a linear machine (with x and y now real numbers)

$$y(t + \Delta) - ay(t) = bx(t) \tag{15a}$$

or in its differential analog expanding $y(t + \Delta) = y(t) + \Delta dy/dt$:

$$\frac{dy}{dt} - \alpha y = x(t) \tag{15b}$$

with the corresponding solutions

$$y(n\Delta) = a^n[y(0) + b \sum_{i=0}^{n} a^{-i}x(i\Delta)] \tag{16a}$$

and

$$y(t) = e^{\alpha t}\left[y(0) + \int_0^t e^{-\alpha\tau}x(\tau)d\tau\right] \tag{16b}$$

From these expressions it is clear that the course of events represented by $x(i\Delta)$ (or $x(\tau)$) is "integrated out," and is manifest only in an additive term which, nevertheless, changes as time goes on.

However, the failure of this simple machine to account for memory should not discourage one from contemplating it as a possible useful element in a system that remembers.

While in these examples the internal states z provided the machine with an appreciation—however small—of its past history, we shall now give an interpretation of the internal states z as being a selector for a specific function in a set of multivalued logical functions. This is most easily seen when writing the driving function f_y in form of a table.

Let $a, b, c \ldots X$ be the input values x; $\alpha, \beta, \gamma \ldots Y$ be the output values y; and $1, 2, 3 \ldots Z$ be the values of the internal states. A particular driving function f_y is defined if to all pairs $\{xz\}$ an appropriate value of y is associated. This is suggested in Table I.

Clearly, under $z = 1$ a particular logical function, $y = F_1(x)$, relating y with x is defined; under $z = 2$ another logical function, $y = F_2(x)$, is defined; and, in general, under each z a certain logical function $y = F_z(x)$ is defined.

Hence, the driving function f_y can be rewritten to read

$$y = F_z(x), \tag{17}$$

TABLE I

Computing Z Logical Function $F_z(x)$ on Inputs x

z	1	1	1	. . .	1	2	2	2	. . .	2	. . .	Z	Z	Z	. . .	Z
x	a	b	c	. . .	X	a	b	c	. . .	X	. . .	a	b	c	. . .	X
y	γ	α	β	. . .	δ	α	γ	β	. . .	ϵ	. . .	β	ϵ	γ	. . .	δ

which means that this machine computes another logical function $F_{z'}$ on its inputs x, whenever its internal state z changes according to the state function $z' = f_z(x, z)$.

Or, in other words, whenever z changes, the machine becomes a *different* trivial machine.

While this observation may be significant in grasping the fundamental difference between nontrivial and trivial machines, and in appreciating the significance of this difference in a theory of behavior, it permits us to calculate the number of internal states that can be effective in changing the *modus operandi* of this machine.

Following the paradigm of calculating the number $\mathfrak{N}$ of logical functions as the number of states of the dependent variable raised to the power of the number of states of the independent variables

$$\mathfrak{N} = (\text{no. of states of dep. variables})^{(\text{no. of states of indep. variables})} \tag{18}$$

we have for the number of possible trivial machines which connect y with x

$$\mathfrak{N}_T = Y^X \tag{19}$$

This, however, is the largest number of internal states which can effectively produce a change in the function $F_z(x)$, for any additional state has to be paired up with a function to which a state has been already assigned, hence such additional internal states are redundant or at least indistinguishable. Consequently

$$Z \leq Y^X$$

Since the total number of driving functions $f_y(x, z)$ is

$$\mathfrak{N}_D = Y^{XZ}, \tag{20}$$

its largest value is:

$$\bar{\mathfrak{N}}_D = Y^{XY^X} \tag{21}$$

Similarly, for the number of state functions $f_s(z, x)$ we have

$$\mathfrak{N}_S = Z^{X \cdot Z} \tag{22}$$

whose largest effective value is

$$\bar{\mathfrak{N}}_S = Y^{X \cdot X Y^X} = [\bar{\mathfrak{N}}_D]^X \tag{23}$$

These numbers grow very quickly into meta-astronomical magnitudes even for machines with most modest aspirations.

Let a machine have only one two-valued output ($n_y = 1$; $v_y = 2$; $y = \{0; 1\}$; $Y = 2$) and n two-valued inputs ($n_x = n$; $v_x = 2$; $x = \{0; 1\}$; $X = 2^n$). Table II gives the number of effective internal states, the number of possible driving functions, and the number of effective state functions for machines with from one to four "afferents" according to the equations

$$Z = 2^{2^n}$$

$$\mathfrak{N}_D = 2^{2^{2n}+n}$$

$$\mathfrak{N}_S = 2^{2^{2n}+2n}$$

These fast-rising numbers suggest that already on the molecular level without much ado a computational variety can be met which defies imagination. Apparently, the large variety of results of genetic computation, as manifest in the variety of living forms even within a single species, suggests such possibilities. However, the discussion of these possibilities will be reserved for the next section.

TABLE II

The Number of Effective Internal States Z, the Number of Possible Driving Functions $\mathfrak{N}_D$, and the Number of Effective State Functions $\mathfrak{N}_S$ for Machines with One Two-Valued Output and with from One to Four Two-Valued Inputs

n	Z	$\mathfrak{N}_D$	$\mathfrak{N}_S$
1	4	256	65536
2	16	2.10^{19}	6.10^{76}
3	256	10^{600}	$300.10^{4.10^3}$
4	65536	$300.10^{9.10^3}$	$1600.10^{7.10^4}$

2. *Interacting Machines*

We shall now discuss the more general case in which two or more such machines interact with each other. If some aspects of the behavior of an organism can be modeled by a finite state machine, then the interaction of the organism with its environment may be such a case in question, if the environment is likewise representable by a finite state machine. In fact, such two-machine interactions constitute a popular paradigm for interpreting the behavior of animals in experimental learning situations, with the usual relaxation of the general complexity of the situation, by chosing for the experimental environment a trivial machine. "Criterion" in these learning experiments is then said to have been reached by the animal when the experimenter succeeded in transforming the animal from a nontrivial machine into a trivial machine, the result of these experiments being the interaction of just two trivial machines.

We shall denote quantities pertaining to the environment (E) by Roman letters, and those to the organism (Ω) by the corresponding Greek letters. As long as E and Ω are independent, six equations determine their destiny. The four "machine equations," two for each system

$$E: \quad y = f_y(x, z) \tag{24a}$$

$$z' = f_z(x, z) \tag{24b}$$

$$\Omega: \quad \eta = f_\eta(\xi, \zeta) \tag{25a}$$

$$\zeta' = f_\zeta(\xi, \zeta) \tag{25b}$$

and the two equations that describe the course of events at the "receptacles" of the two systems

$$x = x(t); \qquad \xi = \xi(t) \tag{26a, b}$$

We now let these two systems interact with each other by connecting the (one step delayed) output of each machine with the input of the other. The delay is to represent a "reaction time" (time of computation) of each system to a given input (stimulus, cause) (see Fig. 4). With these connections the following relations between the external variables of the two systems are now established:

$$x' = \eta = u'; \quad \xi' = y = v' \tag{27a, b}$$

where the new variables u, v represent the "messages" transmitted from $\Omega \rightarrow E$ and $E \rightarrow \Omega$ respectively. Replacing x, y, η, ξ, in Eqs. (24) (25) by u, v according to Eq. (27) we have

$$\begin{aligned} v' &= f_y(u, z); \qquad u' = f_\eta(v, \zeta) \\ z' &= f_z(u, z); \qquad \zeta' = f_\zeta(v, \zeta) \end{aligned} \tag{28}$$

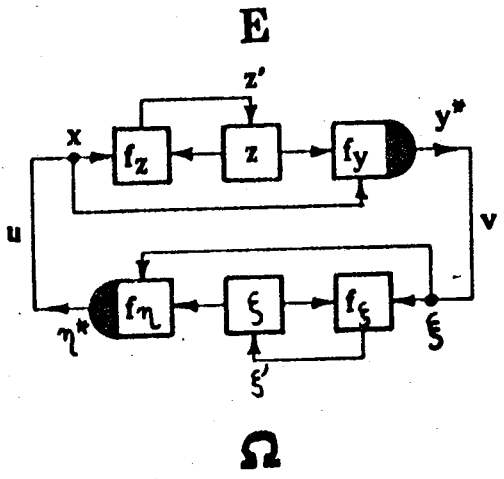


Fig. 4. Two finite state machines (E) (Ω) connected via delays (black semicircles).

These are four recursive equations for the four variables u, v, z, ζ, and if the four functions f_v, f_z, f_y, f_ζ are given, the problem of "solving" for $u(t)$, $v(t)$, $z(t)$, $\zeta(t)$, i.e., expressing these variables explicitly as functions of time, is purely mathematical. In other words, the "meta-system" ($E\Omega$) composed of the subsystems E and Ω, is physically as well as mathematically "closed," and its behavior is completely determined for all times. Moreover, if at a particular time, say $t = 0$ (initial condition), the values of all variables $u(0)$, $v(0)$, $z(0)$, $\zeta(0)$ are known, it is also completely predictable. Since this meta-system is without input, it churns away according to its own rules, coming ultimately to a static or dynamic equilibrium, depending on the rules and the initial conditions.

In the general case the behavior of such systems has been extensively studied by computer simulation (Walker, 1965; Ashby and Walker, 1966; Fitzhugh, 1963), while in the linear case the solutions for Eqs. (28) can be obtained in straight-forward manner, particularly if the recursions can be assumed to extend over infinitesimally small steps:

$$w' = w(t + \Delta) = w(t) + \Delta \frac{dw}{dt} \tag{29}$$

Under these conditions the four Eqs. (28) become

$$\dot{w}_i = \sum_{j=1}^{4} a_{ij} w_j \tag{30}$$

where the w_i $(i = 1, 2, 3, 4)$ are now the real numbers and replace the four variables in question, $\dot{w}$ represents the first derivative with respect to time, and the 16 coefficients a_{ij} $(i, j = 1, 2, 3, 4)$ define the four linear functions under consideration. This system of simultaneous, first-order, linear differential equations is solved by

$$w_i(t) = \sum_{j=1}^{4} A_{ij} e^{\lambda_j t} \tag{31}$$

in which λ_j are the roots of the determinant

$$| a_{ij} - \delta_{ij}\lambda | = 0 \tag{32}$$

$$\delta_{ij} = \begin{cases} 1 \ldots i = j \\ 0 \ldots i \neq j \end{cases}$$

and the A_{ij} depend on the initial conditions. Depending on whether the λ_j turn out to be complex, real negative or real positive, the system will ultimately oscillate, die out, or explode.*

While a discussion of the various modes of behavior of such systems goes beyond this summary, it should be noted that a common behavioral feature in all cases is an initial transitory phase that may move over a very large number of states until one is reached that initiates a stable cyclic trajectory, the dynamic equilibrium. Form and length of both the transitory and final equilibrial phases are dependent on the initial conditions, a fact which led Ashby (1956) to call such systems "multistable." Since usually a large set of initial conditions maps into a single equilibrium, this equilibrium may be taken as a *dynamic representation* of a set of events, and in a multistable system each cycle as an "abstract" for these events.

With these notions let us see what can be inferred from a typical learning experiment (e.g., John *et al.*, 1969) in which an experimental animal in a Y-maze is given a choice ($\xi_0 \equiv C$, for "choice") between two actions ($\eta_1 \equiv L$, for "left turn"; $\eta_2 \equiv R$, for "right turn"). To these the environment E, a trivial machine, responds with new inputs to the animal ($\eta_1 = x_1' \rightarrow y_1' = \xi_1'' \equiv S$, for "shock"; or $\eta_2 = x_2' \rightarrow y_2' = \xi_2'' \equiv F$, for "food"), which, in turn, elicit in the animal a pain ($\eta_3 \equiv$ "$-$") or pleasure ($\eta_4 \equiv$ "$+$") response. These responses cause E to return the animal to the original choice situation ($\xi_0 \equiv C$).

Consider the simple survival strategy built into the animal by which

* This result is, of course, impossible in a finite state machine. It is obtained here only because of the replacement of the discrete and finite variables u, v, z, ζ, by w_i which are continuous and unlimited quantities.

under neutral and pleasant conditions it maintains its internal state [$\zeta' = \zeta$, for $(C\zeta)$ and $(F\zeta)$], while under painful conditions it changes it [$\zeta' \neq \zeta$, for $(S\zeta)$]. We shall assume eight internal states ($\zeta = i$; $i = 1, 2, 3, \ldots, 8$).

With these rules the whole system (ΩE) is specified and its behavior completely determined. For convenience, the three functions, $f_y = f$ for the trivial machine E, f_η and f_ζ for Ω are tabulated in Tables IIIa, b, c.

With the aid of these tables the eight behavioral trajectories for the (ΩE) system, corresponding to the eight initial conditions, can be written. This has been done below, indicating only the values of the pairs $\xi\zeta$ as

TABLE IIIa

$y = f(x)$

$x\ (= \eta^*)$	$y\ (= \xi')$
L	S
R	F
$-$	C
$+$	C

TABLE IIIb

$\eta = f_\eta(\xi, \zeta)$

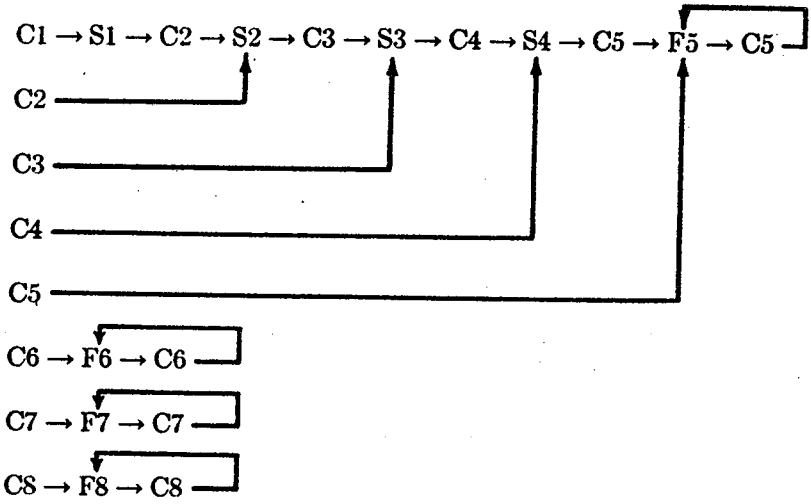
$\eta\ (= x')$		ζ 1	2	3	4	5	6	7	8
	C	L	L	L	L	R	R	R	R
$\xi\ (= y^*)$	S	−	−	−	−	−	−	−	−
	F	+	+	+	+	+	+	+	+

TABLE IIIc

$\zeta' = f_\zeta(\xi, \zeta)$

ζ'		ζ 1	2	3	4	5	6	7	8
	C	1	2	3	4	5	6	7	8
$\xi\ (= y^*)$	S	2	3	4	5	6	7	8	1
	F	1	2	3	4	5	6	7	8

they follow each other as consequences of the organism's responses and the environment's reactions.



These trajectories show indeed the behavior as suggested before, initial transients depending in length on the initial conditions, and ultimately a dynamic equilibrium flipping back and forth between two external states without internal state change. The whole system, and its parts, have become trivial machines. Since, even with maximum semantic tolerance, one cannot say a trivial machine has memory, one wonders what is intended to be measured when at this stage it is opened and the internal workings are examined. Does one wish to inspect its present workings? Or, to see how much it has changed since earlier examinations? At best, these are tests of the experimenter's memory, but whether the machine can appreciate any changes cannot, in principle, be inferred from experiments whose conceptual design eliminates the quality which they intend to measure.

3. *Probabilistic Machines*

This dilemma can be seen in still another light if we adopt for the moment the position of statistical learning theory (Skinner, 1959; Estes, 1959; Logan, 1959). Here either the concept of internal states is rejected or the existence of internal states is ignored. But whenever the laws which connect causes with effects are ignored, either through ignorance or else by choice, the theory becomes that of probabilities.

If we are ignorant of the initial state in the previous example, the chances are 50/50 that the animal will turn left or right on its first trial. After one run the chances are 5/8 for turning right, and so on, until the animal has turned from a "probabilistic (nontrivial) machine" to a "deterministic (trivial) machine," and henceforth always turns right. While a statistical learning theoretician will elevate the changing probabilities in each of the subsequent trials to a "first principle," for the finite state machinist this is an obvious consequence of the effect of certain inputs on the internal states of his machine: they become inaccessible when paired with "painful inputs." Indeed, the whole mathematical machinery of statistical learning theory can be reduced to the paradigm of drawing balls of different color from an urn while observing certain nonreplacement rules.

Let there be an urn with balls of m different colors labeled 0, 1, 2, . . . , $(m - 1)$. As yet unspecified rules permit or prohibit the return of a certain colored ball when drawn. Consider the outcomes of a sequence of n drawings, an "n-sequence," as being an n digit m-ary number (e.g., $m = 10$; $n = 12$):

$$\nu = 1\ 5\ 7\ 3\ 0\ 2\ 1\ 8\ 6\ 2\ 1\ 4$$

Last drawn — First drawn

From this it is clear that there are

$$\mathfrak{N}(n, m) = m^n$$

different n-sequences. A *particular* n-sequence will be called a ν-number, i.e.:

$$0 \leq \nu(m, n) = \sum_{i=1}^{n} j(i) m^{(i-1)} \leq m^n_{-1} \tag{33}$$

where $0 \leq j(i) \leq (m - 1)$ represents the outcome labeled j at the ith trial.

The probability of a *particular* n-sequence (represented by a ν-number) is then

$$p_n(\nu) = \prod_{i=1}^{n} p_i[\, j(i)\,] \tag{34}$$

where $p_i[\, j(i)\,]$ gives the probability of the color labeled j to occur at the ith trial in accordance with the specific ν-number as defined in Eq. (33).

Since after each trial with a "don't return" outcome all probabilities are changed, the probability of an event at the nth trial is said to depend

on the "path," i.e., on the past history of events, that led to this event. Since there are m^{n-1} possible paths that may precede the drawing of j at the nth trial, we have for the probability of this event:

$$p_n(j) = \sum_{\nu=0}^{m^{n-1}-1} p_n(j \cdot m^{n-1} + \nu(n-1, m))$$

where $j \cdot m^{n-1} + \nu(n-1, m)$ represent a $\nu(n, m)$-number which begins with j.

From this a useful recursion can be derived. Let j^* be the colors of balls which when drawn are *not* replaced, and j the others. Let n_{j^*} and n_j be the number of preceding trials on which j^* and j came up respectively $(\sum n_{j^*} + \sum n_j = n - 1)$, then the probability for drawing j (or j^*) at the nth trial with a path of $\sum n_{j^*}$ withdrawals is

$$p_n(j) = \frac{N_j}{N - \sum n_{j^*}} \cdot p_{n-1}(\sum n_{j^*}) \tag{35a}$$

and

$$p_n(j^*) = \frac{N_{j^*} - n_{j^*}}{N - \sum n_{j^*}} \cdot p_{n-1}(\sum n_{j^*}) \tag{35b}$$

where $N = \sum N_j + \sum N_{j^*}$ is the initial number of balls, and N_j and N_{j^*} the initial number of balls with colors j and j^* respectively.

Let there be N balls to begin with in an urn, N_w of which are white, and $(N - N_w)$ are black. When a white ball is drawn, it is returned; a black ball, however, is removed. With "white" $\equiv 0$, and "black" $\equiv 1$, a particular n-sequence $(n = 3)$ may be

$$\nu(3, 2) = 1\ 0\ 1$$

and its probability is:

$$p_3(1\ 0\ 1) = \frac{N - N_w - 1}{N - 1} \cdot \frac{N_w - 1}{N - 1} \cdot \frac{N - N_w}{N}$$

The probability of drawing a black ball at the third trial is then:

$$p_3(1) = p_3(1\ 0\ 0) + p_3(1\ 0\ 1) + p_3(1\ 1\ 0) + p_3(1\ 1\ 1)$$

We wish to know the probability of drawing a white ball at the nth trial. We shall denote this probability now by $p(n)$, and that of drawing a black ball $q(n) = 1 - p(n)$.

By iteratively approximating [through Eq. (35)] trial tails of length m as being path independent $[p_i(j) = p_1(j)]$ one obtains a first-order

approximation for a recursion in $p(n)$:

$$p(n) = p(n-m) + \frac{m}{N} q(n-m) \tag{36}$$

or for $m = n - 1$ (good for $p(1) \approx 1$, and $n/N \ll 1$):

$$p(n) = p(1) + \frac{n-1}{N} q(1) \tag{37}$$

and for $m = 1$ (good for $p(1) \approx 1$):

$$p(n) = p(n-1) + \frac{1}{N} q(n-1) \tag{38}$$

A second approximation changes the above expression to

$$p(n) = p(n-1) + \theta q(n-1) \tag{39}$$

where $\theta = \theta(N, N_w)$ is a constant for all trials. With this we have

$$p(n) - p(n-1) = \Delta p = \theta(1-p) \tag{40}$$

which, in the limit for

$$\lim_{\Delta n \to 0} \frac{\Delta p}{\Delta n} = \frac{dp}{dn}$$

gives

$$\frac{dp}{dn} = \theta(1 - p(n))$$

with the solution

$$p(n) = 1 - (1 - p_0) e^{-\theta n} \tag{41}$$

This, in turn, is an approximation for $p \approx 1$ of

$$p(n) = \frac{p_0}{p_0 + (1 - p_0) e^{-\theta n}} \tag{42}$$

which is the solution of

$$\frac{dp}{dn} = \theta p(1-p) \tag{43}$$

or, recursively expressed, of

$$p(n) = p(n-1) + \theta p(n-1) \cdot q(n-1) \tag{44}$$

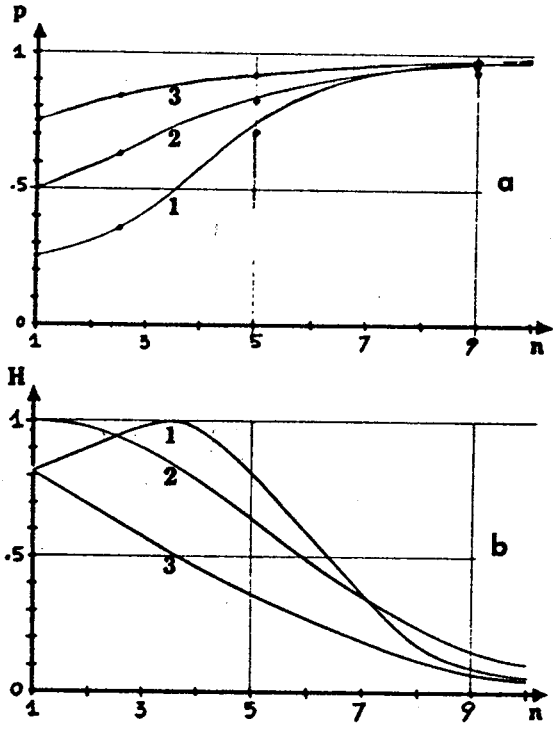


Fig. 5. Probability for drawing a white ball at the nth trial from an urn having initially four balls of which 1, 2, or 3 are white, the others black. White balls are replaced, black are not (a). Entropy at the nth trial (b).

Figure 5a compares the probabilities $p(n)$ for drawing a white ball at the n^{th} trial, as calculated through approximation [Eq. (42)] (solid curves), with the exact values computed by an IBM 360/50 system with a program kindly supplied by Mr. Atwood for an urn with initially four balls ($N = 4$) and for the three cases in which one, two, or three of these are white ($N_w = 1$; $N_w = 2$; $N_w = 3$). The entropy* $H(n)$ in bits per trial corresponding to these cases is shown in Fig. 5b, and one may note that while for some cases [$p(1) \leq 0.5$] it reaches a maximum in the course of this game, it vanishes in all cases when certainty of the outcome is approached [$p(n) \to 1$].

Although the sketch on probabilities dealt exclusively with urns, balls, and draws, students of statistical learning theory will have recognized in Eqs. (39), (41), and (42) the basic axioms of this theory [Estes, 1959;

* Or the "amount of uncertainty"; or the "amount of information" received by the outcome of each trial, defined by $-H(n) = p(n) \log_2 p(n) + q(n) \log_2 q(n)$.

Eqs. (5), (6), and (9)], and there is today no doubt that under the given experimental conditions animals will indeed trace out the learning curves derived for these conditions.

Since the formalism that applies to the behavior of these experimental animals applies as well to our urn, the question now arises: can we say an urn learns? If the answer is "yes," then apparently there is no need for *memory* in learning, for there is no trace of black balls left in our urn when it finally "responds" correctly with white balls when "stimulated" by each draw; if the answer is "no," then by analogy we must conclude it is not *learning* that is observed in these animal experiments.

To escape this dilemma it is only necessary to recall that an urn is just an urn, and it is animals that learn. Indeed, in these experiments learning takes place on two levels. First, the experimental animals learned to behave "urnlike," or better, to behave in a way which allows the experimenter to apply urnlike criteria. Second, the experimenter learned something about the animals by turning them from nontrivial (probabilistic) machines into trivial (deterministic) machines. Hence, it is from studying the experimenter whence we get the clues for memory and learning.

C. Finite Function Machines

1. *Deterministic Machines*

With this observation the question of where to look for memory and learning is turned into the opposite direction. Instead of searching for mechanisms in the environment that turn organisms into trivial machines, we have to find the mechanisms within the organisms that enable them to turn their environment into a trivial machine.

In this formulation of the problem it seems to be clear that in order to manipulate its environment an organism has to construct—somehow—an internal representation of whatever environmental regularities it can get hold of. Neurophysiologists have long since been aware of these abstracting computations performed by neural nets from right at the receptor level up to higher nuclei (Lettvin *et al.*, 1959; Maturana *et al.*, 1968; Eccles *et al.*, 1967). In other words, the question here is how to compute functions rather than states, or how to build a machine that computes programs rather than numerical results. This means that we have to look for a formalism that handles "finite *function* machines." Such a formalism is, of course, one level higher up than the one discussed before, but by maintaining some pertinent analogies its essential features may become apparent.

Our variables are now functions, and since relations between functions are usually referred to as "functionals," the essential features of a calculus of recursive functionals will be briefly sketched.

Consider a system like the one suggested in Fig. 3a, with the only difference that it operates on a finite set of functions of two kinds, $\{f_{yi}\}$ and $\{f_{zj}\}$. These functions, in turn, operate on their appropriate set of states $\{y_i\}$ and $\{z_j\}$. The rules of operation for such a finite function machine are modeled exactly according to the rules of finite state machines. Hence:

$$f_y = F_y[x, f_z] \tag{45a}$$

$$f_z' = F_z[x, f_z] \tag{45b}$$

where F_y and F_z are the functionals which generate the driving functions f_y and the subsequent internal function f_z' from the present internal function f_z and an input x. One should note, however, that the input here is still a state. This indicates an important feature of this formalism, namely, the provision of a link between the domain of states with the entirely different domain of functions. In other words, this formalism takes notice of the distinction between entities and their representations and establishes a relation between these two domains.

Following a procedure similar to that carried out in Eqs. (10) through (14), the functions of type f_z can be eliminated by expressing the present driving function as result of earlier states of affairs. However, due to some properties that distinguish functionals from functions, these earlier states of affairs include both input states as well as output functions. We have for n recursive steps:

$$f_y = \Phi_y^{(n)}[x, x^*, x^{**}, x^{***}, \ldots, x^{(n)*}; f_y^*, f_y^{**}, \ldots, f_y^{(n)*}] \tag{46}$$

Comparing this expression with its analog for finite state machines [Eq. (14)], it is clear that here the reference to past events is not only to those events that were the system's history of inputs $\{x^{(i)*}\}$, but also to its history of potential actions $\{f_y^{(i)*}\}$. Moreover, when this recursive functional is solved explicitly for time ($t = k\Delta$; $k = 0, 1, 2, 3, \ldots$;) [compare with Eq. (16)], it is again the history of inputs that is "integrated out"; however, the history of potential actions remains intact, because of a set of n "eigenfunctions" which satisfy Eq. (46). We have explicitly for $(k\Delta)$, and for the ith eigenfunction:

$$f_y^i(k\Delta) = K_i(k\Delta) \cdot [\pi_i(f_y^{(i)*}) + G_i(x, x^*, x^{**}, \ldots, x^{(n)*})] \tag{47}$$

$$i = 1, 2, 3, \ldots, n$$

with K_i and G_i being functions of $(k\Delta)$, the latter one giving a value that depends on a tail of values in $x^{(i)*}$ which is n steps long. π_i is again a

functional, representing the output function f_y of i steps in the past in terms of another function.

Although this formalism does not specify any mechanism capable of performing the required computations, it provides us, at least, with an adequate description of the functional organization of memory. Access to "past experience" is given here by the availability of the system's own *modus operandi* at earlier occasions, and it is comfortable to see from expression (47) that the subtle distinction between an experience in the past ($f_y^{(i)*}$), and the present experience of an experience in the past [$\pi_i(f_y^{(i)*})$] —i.e., the distinction between "experience" and "memory"—is indeed properly taken care of in this formalism. Moreover, by the system's access to its earlier states of functioning, rather than to a recorded collection of accidental pairs $\{x_i, y_i\}$ that manifest this functioning, it can compute a stream of "data" which are consistent with the system's past experience. These data, however, may or may not contain the output values $\{y_i\}$ of those accidental pairs. This is the price one has to pay for switching domains, from states to functions and back again to states. But this is a small price indeed for the gain of an infinitely more powerful "storage system" which computes the answer to a question, rather than stores all answers together with all possible questions in order to respond with the answer when it can find the question (Von Foerster, 1965).

These examples may suffice to interpret without difficulty another property of the finite function machine that is in strict analogy to the finite state machine. As with the finite state machine, a finite function machine will, when interacting with another system, go through initial

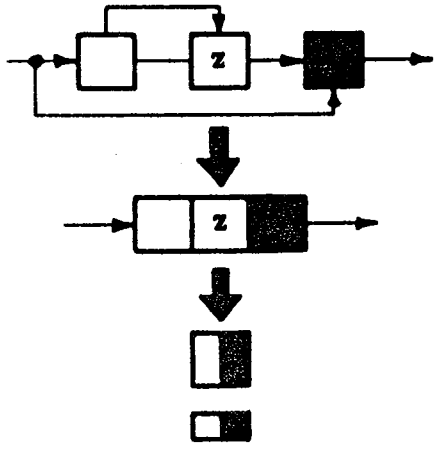


Fig. 6. Symbolization of a finite state machine by a computational tile. Input region white; output region black.

transients depending on initial conditions and settle in a dynamic equilibrium. Again, if there is no internal function change ($f_z' = f_z = f_0$) we have a "trivial finite function machine" with its "goal function" f_0. It is easy to see that a trivial finite function machine is equivalent to a nontrivial finite state machine.*

Instead of citing further properties of the functional organization of finite function machines, it may be profitable to have a glance at various possibilities of their structural organization. Clearly, here we have to deal with aggregates of large numbers of finite state machines, and a more efficient system of notation is required to keep track of the operations that are performed by such aggregates.

2. *Tesselations*

Although a finite state machine consists of three distinct parts, the two computers, f_y and f_z, and the store for z, (see Fig. 3a), we shall represent the entire machine by a single square (or rectangle); its input region denoted white, the output region black (Fig. 6). We shall now treat this unit as an elementary computer—a "computational tile," T_i—which, when combined with other tiles, T_j, may form a mosaic of tiles—a "computational tesselation," $\mathfrak{I}$. The operations performed by the ith tile shall be those of a finite state machine, but different letters, rather than subscripts, will be used to distinguish the two characteristic functions. Subscripts shall refer to tiles.

$$y_i = f_i(x_i, z_i)$$

$$z_i = g_i(x_i, z_i) \tag{48}$$

Figure 7/I sketches the eight possible ways (four each for the parallel and the antiparallel case) in which two tiles can be connected. This results in three classes of elementary tesselations whose structures are suggested in Fig. 7/II. Cases I/1 and I/3, and I/2 and I/4 are equivalent in the parallel case, and are represented in II/1 ("chain") and II/2 ("stack") respectively. In the antiparallel case the two configurations I/1 and I/3 are ineffective, for outputs cannot act on outputs, nor inputs on inputs; cases I/2 and I/4 produce two autonomous elementary tesselations $A = [a^+, a^-\}$, distinct only by the sense of rotation in which the signals are processed.

Iterations of the same concatenations result in tesselations with the

* In the case of several equilibria $\{f_{oi}\}$, we have, of course, a *set* of nontrivial finite state machines that are the outcomes of various initial conditions.

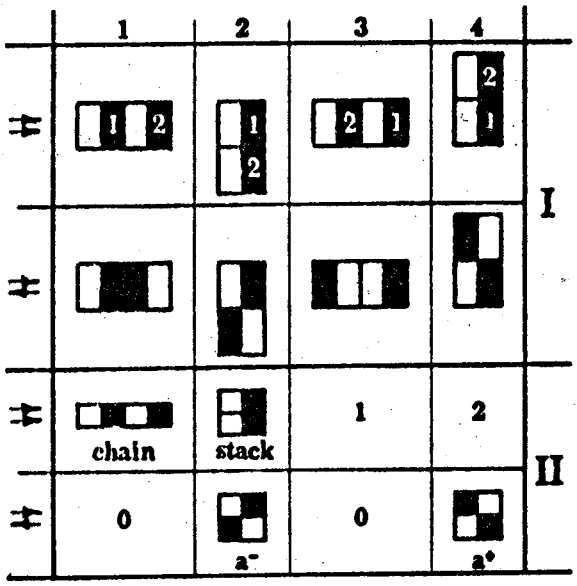


Fig. 7. Elementary tesselations.

following functional properties (for n iterations):

1. *Stack*

$$nT: \qquad y = \sum_{1}^{n} f_i(x_i, z_i) \tag{49}$$

2. *Chain*

$$T^n: \qquad y = f_n(f_{n-1}(f_{n-2}\ldots(x^{(n)*}, z^{(n)*})\ldots z^{**}_{n-2})z^{*}_{n-1})z_n \tag{50}$$

3. $A = \{a^+, a^-\}$

$$\left.\begin{matrix} a^+a^- \\ a^-a^+ \end{matrix}\right\} = 0 \qquad \left.\begin{matrix} a^+a^+ \\ a^-a^- \end{matrix}\right\} \neq 0$$

(i) *Stack* nA^n (51)

(ii) *Chain* A^n (52)

Introducing a fourth elementary tesselation by connecting horizontally

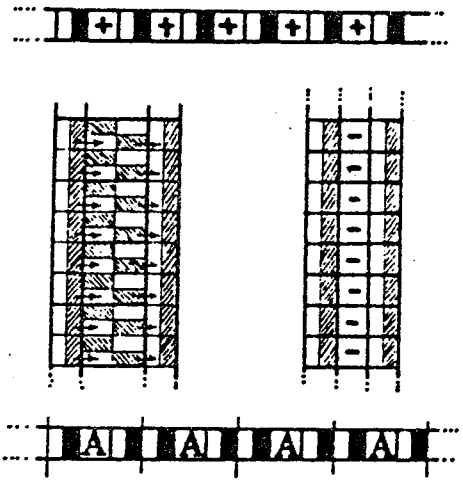


Fig. 8. Some examples of simple tesselations.

$T \rightarrow A \rightarrow T$, or TAT, we have

4. TAT

(i) *Stack* $n(TA^nT)$ (53)

(ii) *Chain* $(TAT)^n$ (54)

Figure 8 suggests further compositions of elementary tesselations. All of these contain autonomous elements, for it is the presence of at least two such elements as, e.g., in $(TAT)^2$, which constitute a finite function machine. If none of these elements happens to be "dead"—i.e., are locked into a single state static equilibrium—they will by their interaction force each other from one dynamic equilibrium into another one. In other words, under certain circumstances they will turn each other from one trivial finite function machine into another one, but this is exactly the criterion for being a nontrivial finite function machine.

It should be pointed out that this concept of formal mathematical entities interacting with each other is not new. John von Neumann (1966) developed this concept for self-reproducing "automata" which have many properties in common with our tiles. Lars Löfgren (1962) expanded this concept to include self-repair of certain computational elements which are either stationary or freely moving in their tesselations, and Gordon Pask (1962) developed similar ideas for discussing the social self-organization of aggregates of such automata.

It may be noted that in all these studies ensembles of elements are contemplated in order to achieve logical closure in discussing the proprietory concept and autonomous property regarding the elements in question as, e.g., *self*-replication, *self*-repair, *self*-organization, *self*-explanation, etc. This is no accident, as Löfgren (1968) observed, for the prefix "self-" can be replaced by the term to which it is a prefix to generate a second-order concept, a concept of a concept. Self-explanation is the explanation of an explanation; self-organization is the organization of an organization (Selfridge, 1962), etc. Since cognition is essentially a self-referring process (Von Foerster, 1969), it is to be expected that in discussing its underlying mechanisms we have to contemplate function of functions and structure of structures.

Since with the build-up of these structures their functional complexity grows rapidly, a detailed discussion of their properties would go beyond the scope of this article. However, one feature of these computational tesselations can be easily recognized, and this is that their operational modalities are closely linked to their structural organization. Here function and structure go hand in hand, and one should not overlook that perhaps the lion's share of computing has been already achieved when the system's topology is established (Werner, 1969). In organisms this is, of course, done mainly by genetic computations.

This observation leads us directly to the physiology and physics of organic tesselations.

III. BIOPHYSICS

A. General Remarks

The question now arises whether or not one can identify structural or functional units in living organisms which can be interpreted in terms of the purely mathematical objects mentioned previously, the "tiles," the "automata," the "finite function machines," etc. This method of approach, first making an interpretation and then looking for confirming entities, seems to run counter to "the scientific method" in which the "facts" are supposed to precede their interpretation. However, what is reported as "fact" has gone through the observer's cognitive system which provides him, so to say, with a priori interpretations. Since our business here is to identify the mechanisms that observe observers (i.e., becoming "self-observers"), we are justified in postulating first the necessary functional structure of these mechanisms. Moreover, this is indeed a popular approach, as seen by the frequent use of terms like "trace," "engram," "store," "read-in," "read-out," etc., when mechanisms of memory are discussed.

Clearly, here too the metaphor precedes the observations. But metaphors have in common with interpretations the quality of being neither true nor false; they are only useful, useless, or misleading.

When a functional unit is conceptually isolated—an *animal*, a *brain*, the *cerebellum*, *neural nuclei*, a *single neuron*, a *synapse*, a *cell*, the *organelles*, the *genomes*, and other molecular building blocks—in its abstract sense the concept of "machine" applied to these units is useful, if it were only to discipline the user of this concept to identify properly the structural and functional components of his "machine." Indeed, the notions of the finite state machine, or all its methodological relatives, have contributed—explicitly or implicitly—much to the understanding of a large variety of such functional units. For instance, the utility of the concepts "transcript," "en-coding," "de-coding," "computation," etc., in molecular genetics cannot be denied.

Let the n-sequence of the four bases ($b = 4$) of a particular DNA molecule be represented by a ν-number $\nu(n, b)$ [see Eq. (33)]; let $Tr(\nu) = \bar{\nu}$ be an operation which transforms the symbols (0, 1, 2, 3) $\rightarrow$ (3, 2, 1, $\emptyset$), in that order, with 0 $\equiv$ thymine, 1 $\equiv$ cytosine, 2 $\equiv$ guanine, 3 $\equiv$ adenine, and $\emptyset \equiv$ uracil, and I be the identity operation $I(\nu) = \nu$; finally, let $\Phi[\bar{\nu}(n, b)] = \nu(n/3, a) = \mu(m, a)$, with $a = 20$, and $j = 0, 1, \ldots, 19$, representing the 20 amino acids of the polypeptide chain. Then

$$\text{(i) DNA replication: } \nu = I(\nu) \tag{55a}$$

$$\text{(ii) DNA/RNA transcript: } \bar{\nu} = Tr(\nu) \tag{55b}$$

$$\text{(iii) Protein synthesis: } \mu = \Phi(\bar{\nu}) \tag{55c}$$

While the operations I and Tr require only trivial machines for the process of transcription, Φ is a recursive computation of the form

$$j(i) \equiv y(i) = y(i-1) + a^i f(x) \tag{56}$$

Using the suggested recursion [compare with Eq. (14)]:

$$y(i) = a^i f(x) + a^{i-1} f(x^*) + a^{i-2} f(x^{**}) \ldots$$

or

$$y(i) = \sum_{k=0}^{i} a^{i-k} f(x^{(k)*}) \tag{57}$$

and

$$y(m) \equiv \mu(m, a)$$

The function f is, of course, computed by the ribosome which reads the codon x, and synthesizes the amino acids which, in turn, are linked together by the recursion to a connected polypeptide chain.

Visualizing the whole process as the operations of a sequential finite state machine was probably more than just a clue in "breaking the genetic code" and identifying as the input state to this machine the triplet (u, v, w) of adjacent symbols in the ν-number representation of the messenger RNA.

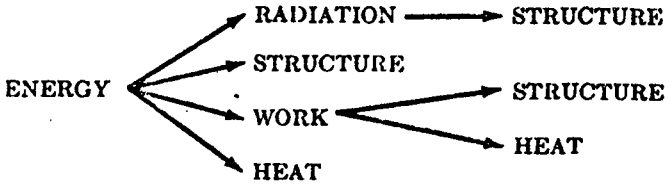
A method for computing ν-numbers of molecular sequences directly from properties of the generated structure was suggested by Pattee (1961). He used the concept of a sequential "shift register," i.e., in principle that of an autonomous tile. For computing periodic sequences in growing helical molecules, the computation for the next element to be attached to the helix is solely determined by the present and some earlier building block. No extraneous computing system is required.

If on a higher level of the hierarchical organization the neuron is taken as a functional unit, the examples are numerous in which it is seen as a recursive function computer. Depending on what is taken to be the "signal," a single pulse, an average frequency code, a latency code, a probability code (Bullock, 1968), etc., the neuron becomes an "all or nothing" device for computing logical functions (McCulloch and Pitts, 1943), a linear element (Sherrington, 1906), a logarithmic element, etc., by changing in essence only a single parameter characteristic for that neuron (Von Foerster, 1967b). The same is true for neural nets in which the recursion is achieved by loops or sometimes directly through recurring fibers. The "reverberating" neural net is a typical example of a finite state machine in its dynamic equilibrium.

In the face of perhaps a whole library filled with recorded instances in which the concept of the finite state machine proved useful, it may come as a surprise that on purely physical grounds these systems are absurd. In order to keep going they must be nothing less than perpetual motion machines. While this is easily accomplished by a mathematical object, it is impossible for an object of reality. Of course, from a heuristical point of view it is irrelevant whether or not a model is physically realizable, as long as it is self-consistent and an intellectual stimulus for further investigations.

However, when the flow of energy between various levels of organization is neglected, and the mechanisms of energy conversion and transfer are ignored, difficulties arise in matching descriptive parameters of functional units on one level to those of higher or lower levels. For instance, a relation between the code of a particular nuclear RNA molecule and, say, the pulse frequency code at the same neuron cannot be established, unless mechanisms of energy transfer are considered. As long as the question as to what keeps the organism going and how this is done is not asked, the gap between functional units on different levels of organization remains open. Can it be closed by thermodynamics?

Three different kinds of molecular mechanisms that offer themselves readily for this purpose will be briefly discussed. All of them make use of various forms of energy as radiation (νh), potential energy (V, structure), work ($p\Delta v$), and heat ($k\Delta T$), and its various conversions from one form to another.



We remain in the terminology of finite state machines and classify the three kinds of mechanism according to their inputs and their outputs, dropping, however, for the moment all distinctions of forms of energy, except that of potential energy (structure) as distinct from all other forms (energy).

(i) Molecular store: Energy in,
Energy out.
(ii) Molecular computer: Energy in,
Structure out.
(iii) Molecular carrier: Energy and structure in,
Energy out.

These three cases will now be briefly reviewed.

B. Molecular Store

Probably the most obvious, and hence perhaps the oldest, approach to link macroscopic behavior, as for example, the forgetting of nonsense syllables (Ebbinghaus, 1885), with the quantum mechanical decay of the available large number of excited metastable states in macromolecules, assumes no further analyzable "elementary impressions" that are associated with a molecule's meta-stable state (Von Foerster, 1948; Von Foerster, 1949). By a nondestructive read-out they can be transferred to another molecule, and a record of these elementary impressions may either decay or else grow, depending on whether the product of the quantum decay time constant with the scanning rate of the read-out is either smaller or else larger than unity. While this model gives good agreement between macroscopic variables such as forgetting rates, temperature dependence of conceived

lapse of time (Hoagland, 1951; Hoagland, 1954), and such microscopic variables as binding energies, electron orbital frequencies, it suffers the malaise of all recording schemes, namely, it is unable to infer anything from the accumulated records. Only if an inductive inference machine which computes the appropriate behavior functions is attached to this record can an organism survive (Von Foerster *et al.*, 1968). Hence, one may abandon speculations about systems that just record specifics, and contemplate those that compute generalizations.

C. Molecular Computer

The good match between macroscopic and microscopic variables of the previous model suggests that this relation should be pursued further. Indeed, it can be shown (Von Foerster, 1969) that the energy intervals between excited meta-stable states are so organized that the decay times in the lattice vibration band correspond to neuronal pulse intervals, and their energy levels to a polarization potential of from 60 mV to 150 mV. Consequently, a pulse train of various pulse intervals will "pump" such a molecule up into higher states of excitation, depending on its initial condition. However, if the excitation level reaches about 1.2 eV, the molecule undergoes configurational changes with life spans of 1 day or longer. In this "structurally charged" state it may now participate in various ways in altering the transfer function of a neuron, either transmitting its energy to other molecules or facilitating their reaction. Since in this model undirected electrical potential energy is used to cause specific structural change, it is referred to as "energy in—structure out." This, however, gives rise to a concept of molecular computation, the result of which is deposition of energy on a specific site of utilization. This is the content of the next and last model.

D. Molecular Carrier

One of the most widely used principles of energy dissemination in a living organism is that of separation of sites of synthesis and utilization. The general method employed in this transfer is a cyclic operation that involves one or many molecular carriers which are "charged" at the site where environmental energy can be absorbed, and are "discharged" where this energy must be used. Charging and discharging is usually accomplished by chemical modifications of the basic carrier molecules. One obvious example of the directional flow of energy and the cyclic flow of matter is, of course, the complementarity of the processes of photosynthesis and respiration (Fig. 9). Light energy, νh, breaks the stable bonds of inorganic oxides and transforms them into energetically charged organic molecules.

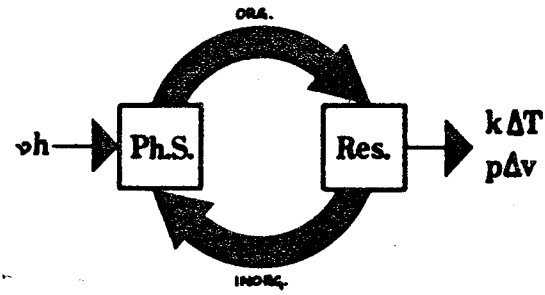


Fig. 9. Directional flow of energy and cyclic flow of matter in photosynthesis coupled with respiration.

These, in turn, are burned up in the respiratory process, releasing the energy in the form of work, $p\Delta v$, or heat, $k\Delta T$, at the site of utilization and return again as inorganic oxides to the site of synthesis.

Another example is the extremely involved way in which in the mitochondria the uphill reaction is accomplished. This reaction not only synthesizes adenosine triphosphate (ATP) by coupling a phosphate group to adenosine diphosphate (ADP), but also charges the ATP molecule with considerable energy which is effectively released during muscular contraction; the contraction process converts ATP back again into ADP by losing the previously attached phosphate group.

Finally, the messenger RNA may be cited as an example of separate sites for synthesis and utilization, although in this case the energetics are as yet not so well established as in the other cases. Here, apparently it is structure which is to be transferred from one place to another, rather than energy.

Common in all these processes is the fact that during synthesis not only a releasable package of energy, ΔE, is put on this molecular carrier but also an address label saying where to deliver the package. This address requires an additional amount of organization, $-\Delta H$, (negentropy), in order to locate its destination. Hence we have the crucial condition

$$\frac{\Delta E}{\Delta H} < 0 \tag{58}$$

which says "for high energy have a low entropy, and for low energy have a high entropy." This is, of course, contrary to the usual course of events in which these two quantities are coupled with each other in a positive relationship.

It can be shown, however, that if a system is composed of constituents which in the ground state are separated, but when "excited" hang together

by "reasonably stable" metastable states, it fulfills the crucial condition above (Von Foerster, 1964).

Let

$$V = \pm\left(Ae^{-x/\kappa} + B \sin \frac{2\pi x}{p}\right) \tag{59}$$

with

$$A/B \gg 1 \quad \text{and} \quad \kappa/p \gg 1$$

be the potential distribution in two one-dimensional linear "periodic crystals," C^+ and C^-, where the $\pm$ refer to corresponding cases. The essential difference between these two linear structures which can be envisioned as linear distributions of electric charges changing their sign (almost) periodically is that energy is required to put "crystal" C^+ together, while for "crystal" C^- about the same energy is required to decompose it into its constituents. These linear lattices have metastable equilibria at

$$C^+ \rightarrow x_1, x_3, x_5 \ldots$$

$$C^- \rightarrow x_0, x_2, x_4 \ldots$$

which are solutions of

$$e^{x/\kappa} \cos \frac{2\pi x}{p} = \frac{1}{2\pi} \frac{Ap}{B\kappa} \approx 1$$

These states are protected by an energy threshold which lets them stay in this state on the average of amount an time

$$\tau = \tau_0 e^{\Delta v/kT} \tag{60}$$

where τ_0^{-1} is an electron orbital frequency, and ΔV is the difference between the energies at the valley and the crest of the potential wall $[\pm \Delta V_n = V(x_n) - V(x_{n+1})]$.

In order to find the entropy of this configuration, we solve the Schrödinger equation (given in normalized form)·

$$\psi'' + \psi[\lambda - V(x)] = 0 \tag{61}$$

for its eigenvalues λ_i and eigenfunctions ψ_i, ψ_i^*, which, in turn, give the probability distribution for the molecule being in the ith eigenstate:

$$\left(\frac{dp}{dx}\right)_i = \psi_i \cdot \psi_i^* \tag{62}$$

with, of course,

$$\int_{-\infty}^{+\infty} \psi_i \cdot \psi_i^* dx = 1 \tag{63}$$

whence we obtain the entropy

$$H_i = -\int_{-\infty}^{+\infty} \psi_i \cdot \psi_i^* \ln \psi_i \cdot \psi_i^* \tag{64}$$

for the ith eigenstate.

It is significant that indeed for the two crystals C^+ and C^- the change in the ratio of energy to entropy for charging ($\Delta E = e(V(x_n) - V(x_{n+2}))$ goes into opposite directions:

$$C^- \rightarrow \left(\frac{\Delta E}{\Delta H}\right)^- > 0$$

$$C^+ \rightarrow \left(\frac{\Delta E}{\Delta H}\right)^+ < 0$$

This shows that the two crystals are quite different animals: one is dead (C^-), the other is alive (C^+).

IV. SUMMARY

In essence this paper is a proposal to restore the original meaning of concepts like memory, learning, behavior, etc. by seeing them as various manifestations of a more inclusive phenomenon, namely, cognition. An attempt is made to justify this proposition and to sketch a conceptual machinery of apparently sufficient richness to describe these phenomena in their proper extension. In its most concise form the proposal was presented as a search for mechanisms within living organisms that enable them to turn their environment into a trivial machine, rather than a search for mechanisms in the environment that turn the organisms into trivial machines.

This posture is justified by realizing that the latter approach—when it succeeds—fails to account for the mechanisms it wishes to discover, for a trivial machine does not exhibit the desired properties; and when it fails does not reveal the properties that made it fail.

Within the conceptual framework of finite state machines, the calculus of recursive functionals was suggested as a descriptive (phenomenological) formalism to account for memory as potential awareness of previous interpretations of experiences, hence for the origin of the *concept* of

"change," and to account for transitions in domains that occur when going from "facts" to "description of facts" and—since these in turn are facts too—to "descriptions of descriptions of facts" and so on.

Elementary finite function machines can be strung together to form linear or two-dimensional tesselations of considerable computational flexibility and complexity. Such tesselations are useful models for aggregates of interacting functional units at various levels in the hierarchical organization of organisms. On the molecular level, for instance, a stringlike tesselation coiled to a helix may compute itself (self-replication) or, in conjunction with other elements, compute other molecular functional units (synthesis).

While in the discussion of descriptive formalisms the concept of recursive functionals provides the bridge for passing through various descriptive domains, it is the concept of energy transfer connected with entropic change that links operationally the functional units on various organizational levels. It is these links, conceptual or operational, which are the prerequisites for interpreting structures and function of a living organism seen as an autonomous self-referring organism. When these links are ignored, the concept of "organism" is void, and its unrelated pieces becomes trivialities or remain mysteries

ACKNOWLEDGMENT

Some of the ideas and results presented in this article grew out of work jointly sponsored by the Air Force Office of Scientific Research under Grant AF-OSR 7-67, by the Office of Education under Grant OEC-1-7-071213-4557, and by the Air Force Office of Scientific Research under Grant AF 49(638)-1680.

REFERENCES

Ashby, W. R., 1956, "An Introduction to Cybernetics," Chapman and Hall, London.

Ashby, W. R., 1962, The Set Theory of Mechanisms and Homeostasis, Technical Report 7, NSF Grant 17414, Biological Computer Laboratory, Electrical Engineering Department, University of Illinois, Urbana, 44 pp.

Ashby, W. R., and Walker, C., 1966, On Temporal Characteristics of Behavior in Certain Complex Systems, Kybernetik 3:100.

Bullock, T. H., 1968, Biological Sensors, in "Vistas in Science" (D. L. Arm, ed.) pp. 176-206, University of New Mexico, Albuquerque.

Ebbinghaus, H., 1885, "Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie," Drucker & Humbold, Leipzig.

Eccles, J. C., Ito, M., and Szentagothai, J., 1967, "The Cerebellum as a Neuronal Machine," Springer-Verlag, New York.

Estes, W. K., 1959, The Statistical Approach to Learning Theory, in "Psychology: A Study of a Science, I/2" (S. Koch, ed.) pp. 380-491, McGraw-Hill, New York.

Fitzhugh, H. S. II, 1963, Some considerations of polystable systems, IEEE Transactions 7:1.

Gill, A., 1962, "Introduction to the Theory of Finite State Machines," McGraw-Hill, New York.

Gunther, G., 1967, Time, timeless logic and self-referential systems, in "Interdisciplinary Perspectives of Time" (R. Fischer, ed.) pp. 396-406, New York Academy of Sciences, New York.

Hoagland, H., 1951, Consciousness and the chemistry of time, in "Problems of Consciousness Tr. First Conf." (H.A. Abramson, ed.) pp. 164-198, Josiah Macy Jr. Foundation, New York.

Hoagland, H., 1954, (A remark), in "Problems of Consciousness Tr. Fourth Conf." (H. A. Abramson, ed.) pp. 106-109, Josiah Macy Jr. Foundation, New York.

John, E. R., Shimkochi, M., and Bartlett, F., 1969, Neural readout from memory during generalization, Science 164:1534.

Konorski, J., 1962, The role of central factors in differentiation, in "Information Processing in the Nervous System" (R. W. Gerard and J. W. Duyff, eds.) Vol. 3, pp. 318-329, Excerpta Medica Foundation, Amsterdam.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W., 1959, What the frog's eye tells the frog's brain, Proc. I.R.E. 47:1940.

Löfgren, L., 1962, Kinematic and tesselation models of self-repair, in "Biological Prototypes and Synthetic Systems" (E. E. Bernard and M. R. Kare, eds.) pp. 342-369, Plenum Press, New York.

Löfgren, L., 1968, An axiomatic explanation of complete self-reproduction, Bull. of Math. Biophysics 30(3):415.

Logan, F. A., 1959, The Hull-Spence approach, in "Psychology: A Study of a Science, I/2" (S. Koch, ed.) pp. 293-358, McGraw-Hill, New York.

McCulloch, W. S., and Pitts, W., 1943, A logical calculus of the ideas immanent in nervous activity, Bull. of Math. Biophysics 5:115.

Maturana, H. R., 1969, Neurophysiology of cognition, in "Cognition — A Multiple View" (P. L. Garvin, ed.) in press, Spartan Books, New York.

Maturana, H. R., Uribe, G., and Frenk, S., 1968, A Biological Theory of Relativistic Colour Coding in the Primate Retina, Supplemento No. 1, Arch. Biologia y Med. Exp., University of Chile, Santiago, 30 pp.

Pask, G., 1962, A proposed evolutionary model, in "Principles of Self-Organization" (H. Von Foerster and G. W. Zopf, Jr., eds.) pp. 229-254, Pergamon Press, New York.

Pask, G., 1968, A cybernetic model for some types of learning and mentation, in "Cybernetic Problems in Bionics" (H. L. Oestreicher and D. R. MOore, eds.) pp. 531-586, Gordon & Breach, New York.

Pattee, H. H., 1961, On the origin of macro-molecular sequences, Biophys. J. 1:683.

Pitts, W., and McCulloch, W. S., 1947, How we know universals; the perception of auditory and visual forms, Bull. of Math. Biophysics 9:127.

Selfridge, O. G., 1962, The organization of organization, in "Self-Organizing Systems" (M. C. Yovits, G. T. Jacoby and G. D. Goldstein, eds.) pp. 1-8, Spartan Books, New York.

Sherrington, C. S., 1906, "Integrative Action of the Nervous System," Yale University Press, New Haven.

Skinner, B. F., 1959, A case history in scientific method, in "Psychology: A Study of a Science, I/2" (S. Koch, ed.) pp. 359-379, McGraw-Hill, New York.

Ungar, G., 1969, Chemical transfer of learning, in "The Future of the Brain Sciences" (S. Bogoch, ed.) pp. 373-374, Plenum Press, New York.

Von Foerster, H., 1948, "Das Gedächtnis; Eine quantenmechanische Untersuchung," F. Deuticke, Vienna.

Von Foerster, H., 1949, Quantum mechanical theory of memory, in "Cybernetics, Transactions of the Sixth Conference" (H. Von Foerster, ed.) pp. 112-145, Josiah Macy Jr. Foundation, New York.

Von Foerster, H., 1964, Molecular bionics, in "Information Processing by Living Organisms and Machines" (H. L. Oestreicher, ed.) pp. 161-190, Aerospace Medical Division, Dayton.

Von Foerster, H., 1965, Memory without record, in "The Anatomy of Memory" (D. P. Kimble, ed.) pp. 388-433, Science and Behavior Books, Palo Alto.

Von Foerster, H., 1966, From stimulus to symbol, in "Sign, Image, Symbol" (G. Kepes, ed.) pp. 42-61, George Braziller, New York.

Von Foerster, H., 1967a, Biological principles of information storage and retrieval, in "Electronic Handling of Information: Testing and Evaluation" (A. Kent et al., eds.) pp. 123-147, Academic Press, London.

Von Foerster, H., 1967b, Computation in neural nets, Currents Mod. Biol. 1:47.

Von Foerster, H., 1969, What is memory that it may have hindsight and foresight as well?, in "The Future of the Brain Sciences" (S. Bogoch, ed.) pp. 19-64, Plenum Press, New York.

Von Foerster, H., Inselberg, A., and Weston, P., 1968, Memory and inductive inference, in "Cybernetic Problems in Bionics" (H. L. Oestreicher and D. R. MOore, eds.) pp. 31-68, Gordon & Breach, New York.

von Neumann, J., 1966, "The Theory of Self-Reproducing Automata," (A. Burks, ed.) University of Illinois Press, Urbana.

Walker, C., 1965, A Study of a Family of Complex Systems, An Approach to the Investigation of Organism's Behavior, Technical Report 5, AF-OSR Grant 7-65, Biological Computer Laboratory, Electrical Engineering Department, University of Illinois, Urbana, 251 pp.

Werner, G., 1969, The topology of the body representation in the somatic afferent pathways, in "The Neurosciences, II" Rockefeller University Press, New York.

Weston, P., 1964, Noun chain tress, unpublished manuscript.