BEHAVIORISM TO COGNITION: A SYSTEM-THEORETIC INQUIRY INTO BRAINS, MINDS AND MECHANISM

John L. Casti

July 1985
WP-85-49

Paper prepared for presentation at the workshop on Brain Research, Cognition and Artificial Intelligence, to be held in Abisko, Sweden, May 1986.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

ABSTRACT

Arguments from mathematical system theory are used to show that the behaviorist-cognitivist debate in psychology is actually a non-issue: ***abstractly***, the two are equivalent; but from the standpoint of a predictive, scientific theory of brains and behavior only the cognitivist program holds any promise.

After a brief summary of the algebraic theory of systems, the paper employs these algebraic tools to propose a functional means by which a brain (human or artificial) may compactly store and retrieve information. This scheme is then extended to provide a means for the generation of thoughts and emotions, as well.

Finally, the paper concludes with a discussion of the interconnections between the brain model suggested here and a number of other models proposed in the literature.

1. Introduction

In the early 1970s, there was a brief flurry of activity directed toward the transcription of classical system theory into the terminology of category theory. One of the consequences of these efforts was a particularly clear and explicit clarification of the relationship between an input/output and a state-variable description of a dynamical process. In category-theoretic terms, they are adjoints. Thus, with each input/output description there is automatically associated a *natural* state-variable description, and conversely. In this sense, the two descriptions are abstractly equivalent.

Having been sensitized by a certain amount of reading and a strong personal interest in problems of mind and human psychology, when I first encountered the duality between external and internal system descriptions, my immediate thought was that such a result was a systems version of the behaviorist-cognitivist split in psychology, and that perhaps the system concepts would provide a framework for consideration of this dichotomy in more formal and precise terms. During the past decade I have had occasion to periodically re-consider this duality, each time armed with somewhat more powerful system-theoretic tools provided by the substantial advances in mathematical system theory over this period. The occasion of this meeting at the systems interface between brain research, cognitive psychology and artificial intelligence provides the opportunity to put forward what amounts to a model for *abstract* thought processes. The details of the framework presented here are almost sure to be wrong; nonetheless, I would be greatly (but not unhappily) surprised if when the final word is written on the structure of the brain, the general concepts presented here do not prove to be the foundation upon which a working theory of any brain, real or artificial, is constructed.

The basic questions that the paper addresses are:

i) do internal cognitive states exist;

ii) if they do, in what way do they store experiences as memory;

iii) how do such cognitive states interact to produce thoughts;

iv) is it possible for artificial devices like computers and non-neuronal intelligences to have mental states, or are such states uniquely characteristic of human brains?

The mathematical structure presented here provides a framework for the *abstract* consideration of these matters. Their interpretation for real physical brains remains a topic for future research.

Before moving on, I want to emphasize that this paper is not an attack on behaviorism; on the one hand, it's fruitless to beat a dead (or, at least, dying) horse, while on the other hand the system-theoretic arguments given here *strengthen* the behaviorist school, at least to the degree that they show that abstractly behaviorism and abstract cognition are two sides of the same coin. One side contains mental states; the other doesn't. But the coin cannot be split apart and the two halves separated. The best we can do is to view it one side at a time. Our principal argument is that one view is more *useful* than the other, not more "correct".

2. Behaviorism, Structuralism and System Models

Stimulated by the general philosophical idea of logical positivism which was in vogue at the time, in the early-1920s John Watson made the radical suggestion that behavior does not have mental causes. This thesis, further developed and modified by Hull, Skinner and others, has come to be termed *psychological behaviorism*. A principal motivation for adoption of the behaviorist view was to rid psychology of the dualist attitude that mind is a non-physical entity, somehow disjoint from the physical brain. The behaviorist solution is to eliminate all notions of mind, mental

states and mental representation from psychological investigation, concentrating solely upon externally observable *stimulus-response* behavior patterns.

By the early-1960s, it was recognized that both the dualist and the behaviorist approach to human behavior were unattractive, and effort was focused upon developing a materialist theory of mind that allowed for mental causes. One such theory, termed *logical behaviorism*, was quite similar to classical behaviorism and is really just classical behaviorism in a semantic form. Another theory, *central-state identity*, postulates that mental events, states and processes are identical with neurophysiological events in the brain. Thus, under the central-state identity theory, a behavioral effect is the result of a causal pattern of physical events in the brain. The problem with the central-state identity notion is that in either its weak or its strong form, *token* and *type physicalism*, resp., it asserts that all mental particulars that exist or could ever exist are neurophysiological. Thus, the logical possibility of machines and other disembodied spirits having mental properties is ruled out because they are not composed of neurons.

During the last decade or so, a way out of these dilemmas has been provided by the theory of *functionalism*, an outgrowth of that amalgam of physics, neurophysiology, computer science and psychology loosely labeled, "cognitive science." Functionalism is based upon the idea that a mental state can be defined by its causal relations to other mental states and that such mental states can, both in principle and in deed, be realized by many systems. In essence, behavior is driven by software, not hardware. A very readable account of these various notions is given in the popular article by Fodor [1] or the books [2,3,22]. Since it will not be necessary for us to distinguish between the central-state identity theory and functionalism, we adopt the generic term *structuralism* to represent any theory of the mind that involves physical mental states, be they manifested in a human brain, a disembodied cloud from space or a collection of silicon wafers in a machine.

The principal aim of this paper is to provide a precise, system-theoretic argument for asserting the *abstract* equivalence of behaviorism and structuralism, while at the same time showing that *operationally* only the structuralist view offers the basis for a predictive, causal view of human behavior. Such a conclusion is a natural consequence of the so-called Realization Theorem of mathematical system theory. Following the path laid out by the structuralist framework, we then provide a fairly detailed mathematical description of the way in which a "brain" would process and store external stimuli in order to generate observed behavioral responses. The paper then concludes with some speculations based upon the theory of system invariants for how thoughts are generated as consequences of internal system dynamics.

3. Stimulus–Response Patterns and External System Models

Let us imagine our information-processing object O (human being, machine, cloud, ...) as consisting of the proverbial "black-box" connected to its environment by certain input and output channels (Fig. 1). Assume that at any given moment t, the stimulus $u(t)$ is selected from some set of symbols U, while the observed response at that moment, $y(t)$, belongs to another set of symbols Y. To simplify the exposition, assume that t takes on only the discrete values $t = 0,1,2, \cdots$.

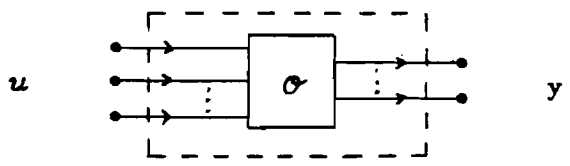


Figure 1. Information processing object.

Then a given stimulus-response pattern of O, B_O, is represented by the sequence

$$B_O = (u(t), y(t)) \ , \ t = 0,1,2, \cdots \ .$$

If we let Ω denote the set of all possible stimuli sequences, with Γ representing the set of all response sequences, then the overall *external behavior* of the object O can be denoted by a stimulus-response map

$$\begin{aligned} f: \ \Omega &\to \Gamma \\ \omega &\mapsto \gamma \ , \end{aligned}$$

where

$$\omega = \{u(0), u(1), u(2), \cdots\} \ , \ \omega \in \Omega \ , \ u(\cdot) \in U \ ,$$

while

$$\gamma = \{y(0), y(1), y(2), \cdots\} \ , \ \gamma \in \Gamma \ , \ y(\cdot) \in Y \ .$$

According to the behaviorists, all that can ever be known about O are the sets Ω and Γ, together with the map f. The entire content of the behaviorist program is to determine f, given Ω and Γ, without postulating any internal mechanisms inside the box. Or, put another way, a behaviorist would claim that to be given f would be to be given everything that could be known about the disposition of the object to behave in a certain way, and that it would be nonscientific to assert the existence of any unobservable internal mechanism generating f. Mathematical system theory provides an honest, true, clear and direct refutation of this claim.

4. Cognitive States and Internal Models

An internal model Σ of the behavioral pattern f involves postulating the existence of a set X of internal *state variables*, and a dynamic relationship g linking the stimuli u and the states, as well as a rule h specifying how internal states combine to generate the response y. More compactly, we have

$$\begin{aligned} x(t+1) &= g(x(t), u(t)), \ x(0) = x_0, \\ y(t) &= h(x(t)), \end{aligned} \qquad (\Sigma)$$

$x(t) \in X$, $u(t) \in U$, $y(t) \in Y$. We would then say that Σ is an internal model of the observed behavior f if the stimulus-response pattern $B_\Sigma = B_O$, i.e., if the

observed input-output behavior of Σ agrees with that of O. Note that in order for $B_\Sigma = B_O$, it is necessary to construct an appropriate set X, together with appropriate maps

$$g : X \times U \to X \ ,$$
$$h : X \to Y \ .$$

From an abstract point of view, the first step in the structuralist program is to ensure that for any given external model $O = (\Omega, \Gamma, f)$, a corresponding internal model $\Sigma = (X, g, h)$ exists. If this is the case, then it would be natural to associate the abstract states X with the postulated physical states of the brain in some fashion, while at the same time interpreting the maps g and h as means for encoding and decoding external stimuli and mental states, respectively. It is one of the great triumphs of mathematical system theory to have been able to provide a rather definitive resolution of this question, happily in the affirmative. The remainder of the paper is devoted to an account of this solution in the above context, together with a detailed exposition of how the encoding/decoding operations are explicitly carried out, followed by some semi-speculative discussion of the process of cognition from a systems perspective.

5. Realizations and Canonical Models

Loosely speaking, we can phrase the behaviorist–structuralist problem as follows:

> *Given a stimulus–response pattern B_O, find a "good" internal model Σ such that $B_O = B_\Sigma$.*

The catch in the above statement is the qualifying condition that the model Σ be "good." It turns out that without imposition of this condition the solution to the problem is trivially easy: there are an infinite number of models $\Sigma = (X, g, h)$ such

that $B_O = B_\Sigma$. But how can we identify a *good* model from this infinitude? The answer hinges upon invoking a system-theoretic translation of Occam's Razor, i.e., a "good" model is one which is "compact" or "minimal" in some well-defined sense. Now let us make this idea more precise.

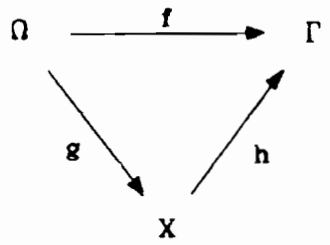
Assume we are given *any* model $\Sigma = (X, g, h)$. Then we say that Σ is *completely reachable* if for any state $x^* \in X$, there exists an input sequence $\omega \in \Omega$ and a time T, such that $x(T; x_0, 0, \omega) = x^*$, i.e., the input ω transfers the system state from x_0 at time $t = 0$ to x^* at time T. Notice that the property of complete reachability depends upon Ω, T and g, but is independent of the output function h.

Now let us focus upon the output of Σ. We call Σ *completely observable* if there exists an input $\omega \in \Omega$ and a time $T > 0$, such that the initial state x_0 can be uniquely determined from observation of the system output $y(t)$, $0 < t \leq T$. Note that observability depends upon Ω, T, and g, as well as h.

Putting the two concepts together, we call Σ *canonical* if it is both completely reachable and completely observable. The minimality criterion is now clear: the state space X of a canonical model is minimal in the sense that there are no elements in X that cannot be accessed using some input, and no two distinct initial states give rise to the same output sequence. Thus, a canonical model is characterized by a state space containing no elements "extraneous" to its input–output behavior B_Σ.

Now we return to the problem of modeling the external behavior B_O by a canonical model Σ. A compact way of viewing the situation is that we seek to construct a space X and maps g and h such that

i) the diagram



commutes and

ii) the map g is *onto*, while the map h is *one-to-one*.

Requirement (i) is just the condition that $B_0 = B_\Sigma$, while (ii) insures that the model $\Sigma = (X,g,h)$ is canonical, i.e. reachable and observable.

The main result of mathematical system theory is the following

Realization Theorem. Given an input–output map $f : \Omega \to \Gamma$, there always exists a canonical model $\Sigma = (X,g,h)$ such that $B_0 = B_\Sigma$. Furthermore, the model Σ is unique, up to a change of coordinates in X.

The proof of this assertion can be found, for example, in [4–5].

Returning now to the behaviorist–structuralist debate, we can re-state the Realization Theorem in psychological terms.

Cognitive Theorem: Given any stimulus–response pattern B_0, there always exists a structuralist model whose behavior is identical to B_0. Furthermore, this structuralist model is essentially unique.

Remarks

1) The Cognitive Theorem only states that associated with any ***physically observable*** stimulus–response pattern B_0, there is an ***abstract*** set X and ***abstract***

maps g and h, such that $\Sigma = (X,g,h)$ forms a canonical model with $B_\Sigma = B_O$. In order for the Cognitive Theorem to form the basis for a structuralist (or materialist) theory of behavior, it is necessary for these abstract objects to somehow be related to *actual* mental states. The latter part of the paper examines just how this might be done.

2) At one level, the Cognitive Theorem says that there is no essential difference between the behaviorist and structuralist schools of thought: they are *abstractly equivalent*. On another level, the two theories are worlds apart; it all depends upon your point of view. The structuralist model provides an *explanatory mechanism* (the states X and the maps g and h) for behavior that also has a built-in *predictive capacity*, as well (the dynamics $x(t+1) = g(x(t),u(t))$). The behaviorist model provides neither; it offers only a catalogue of experimental observations; the raw data, so to speak. From this point of view, a behaviorist can only *report* the data, while a structuralist can actually *generate* it.

6. Algebraic System Theory – A Brief Review*

The diagrammatic representation of the realization problem given in the last section makes it transparently clear that the role of the state space X is to somehow "mediate" between the external inputs from Ω and the observed outputs from Γ. But what kind of psychological interpretation can we attach to this process of mediation? Or, more precisely, what functional interpretation can we attach to the maps g and h? The only consistent answer to this question is to assert that the role of g is to "encode" an external stimuli ω and represent ω internally as a state, while the role of h is to "decode" a state and thereby produce an observable output γ. At this level of abstraction, these observations are

* This section makes heavy mathematical demands upon the reader. While the ideas are a standard part of any introductory course in abstract algebra, many readers may wish to skip this section upon first reading.

fairly obvious. But in order to make any meaningful progress toward tieing together the abstract elements X, g and h and the actual physical structures in a brain, it is necessary to be much more specific about the precise nature of these encoding and decoding operations. The hope, of course, is that the specific mathematical structures involved in coding an input ω as a state will offer a clue as to the physical structures to look for in a real brain. Similarly, the way in which an abstract state is decoded to produce an output γ provides a suggestion as to what pathways to look for in the brain to explain actual behavior.

The simplest and most instructive way to proceed on this question is to examine the case when the input–output behavior f is *linear*. The past decade has seen the development of an elegant algebraic theory of such processes, spearheaded by the original work of Kalman [5]. A good treatment of the algebraic theory of linear systems is found in the works [6–7]. Here we shall sketch the bare essentials of this theory as needed for our subsequent arguments.

The first step in the theory is to show that the input and output spaces Ω and Γ can be given the structure of $k[z]$-modules where k is an arbitrary commutative ring, and $k[z]$ is the usual set of polynomials with coefficients in k. (*Remark*: for most purposes, we usually let k be some number *field* but the added generality of k = ring comes in handy in certain applications). The module structure on Ω and Γ is imposed by noting that we can formally associate any finite input ω with a polynomial $\pi(x)$ having coefficients taken from k, i.e.,

$$\omega \approx \alpha_t z^t + \alpha_{t-1} z^{t-1} + \cdots + \alpha_1 z + \alpha_0 \ ,$$

$\alpha_t \in k^m$, where m is the number of input terminals. The time marker t just indicates the time units prior to $t = 0$ at which the signal α_t was applied.* Thus, under the above identification $\omega \in k^m[z]$, the set of m vectors each of whose components

* For technical reason, it is convenient to assume that the input starts at time $\tau = -t$ and stops at $\tau = 0$. The output then begins at time $\tau = 1$. Such a convention insures causality and can be made without loss of generality.

is a polynomial in $k[z]$. It is easy to see that under scalar multiplication by $k[z]$, the set Ω admits the structure of a finitely-generated free module. For us the importance of this result is that it enables every input to be associated with a polynomial.

A similar construction associating outputs $\gamma \in \Gamma$ with formal power series $k^p[[z^{-1}]]$ can also be carried out. This construction makes Γ a $k[z]$ module, as well. Since our focus is upon inputs, we omit details of this construction, at least for the moment.

The linearity of f, together with the above module structures on Ω and Γ, means that *algebraically* the input–output behavior f is a $k[z]$-*homomorphism*. Thus, by standard algebraic theory there is a module that is naturally induced by f: the quotient $k[z]$-module, $\Omega/\ker f$. Since this module is essential for our development, let us examine the system-theoretic content of this quotient module construction.

We can define a natural notion of equivalence between two inputs ω, $\hat{\omega}$ by

$$\omega \approx_f \hat{\omega} \quad \text{iff} \quad f(\omega) = f(\hat{\omega}) \ .$$

Let us call this *module equivalence*. It is easy to prove that the equivalence classes X_f under $\approx_f$ admit the structure of a $k[z]$-module which we denote $X_f = \Omega/\ker f$. On the other hand, from a systems perspective a somewhat more natural type of equivalence on Ω is the so-called *Nerode equivalence*, in which two inputs ω, $\hat{\omega}$ are Nerode equivalent if and only if the output sequences $f(\omega)$, $f(\hat{\omega})$ are the same and remain identical whenever both ω and $\hat{\omega}$ are followed by an arbitrary $\nu \in \Omega$, i.e.,

$$\omega \sim_N \hat{\omega} \quad \text{iff} \quad f(\omega \circ \nu) = f(\hat{\omega} \circ \nu)$$

for all $\nu \in \Omega$. If we denote the Nerode equivalence classes by $(\omega)_N$ and the module equivalence classes by $[\omega]_f$, the natural question is to ask whether there is any relationship between the two sets of equivalence classes. The answer is *they are*

identical, $(\omega)_N \equiv [\omega]_f$ for all $\omega \in \Omega$.

For a variety of reasons, both substantive and technical, the Nerode classes constitute the natural definition of the state set associated with an input–output map f. Thus, we are led to what Kalman [5] terms

> ***The Fundamental Theorem of Linear System Theory.*** *The natural state set $X_f = \Omega/\ker f$ associated with a linear input–output map f over k admits the structure of a $k[z]$-module.*

In addition, we may define two maps G_f and H_f using the state space X_f. They are

$$\begin{aligned} G_f \;:\; \Omega &\to X_f \\ \omega &\mapsto [\omega]_f \end{aligned}$$

and

$$\begin{aligned} H_f \;:\; X_f &\to \Gamma \\ [\omega]_f &\mapsto f(\omega) \end{aligned}$$

It is easy to verify that both of these maps are $k[z]$-homomorphisms. The triple (X_f, G_f, H_f) is called the ***module of*** f.

Now let us turn our attention to a linear system Σ given in internal form, i.e., we want to associate a $k[z]$-module structure with the system

$$\begin{aligned} x(t+1) &= Fx(t) + Gu(t), \\ y(t) &= Hx(t) \end{aligned} \qquad (\Sigma)$$

$x \in X_\Sigma = k^n$, $u \in k^m$, $y \in k^p$, where F, G, H are k-homomorphisms.

Our first order of business is to impose a $k[z]$-module structure on $X_\Sigma = k^n$. Using standard arguments, since F is a k-endomorphism of the k-vector space X_Σ, we can impose a $k[z]$-module structure on X_Σ given by the rule

$$(\pi, x) \mapsto \pi \cdot x = \pi(F)x \quad ,$$

$\pi \in k[z]$, $x \in X_\Sigma$, i.e., we evaluate the polynomial π on the matrix F, then apply the result to the state x.

The construction of X_Σ as a $k[z]$-module shows us that dynamical action can be regarded as a type of multiplication. More precisely, if $\omega \in \Omega$ is some input, while $x \in X_\Sigma$ is the initial state, then after one time unit, the input ω will have shifted x to the new state

$$x \circ \omega = F^{1+\deg\omega}x + \sum_{t=-\deg\omega}^{0} F^{-t}G\omega(t)$$

Thus, we see that dynamical action can be expressed as module action, i.e., the map $x \mapsto Fx$ in vector-space notation is equivalent to $x \mapsto z \cdot x$ in module notation. Using this fact, we can also write the above expression in module notation as

$$x \circ \omega = z^{1+\deg\omega} \cdot x + \sum_{k=1}^{m} \omega_k \cdot g_k \quad ,$$

where $g_k = Ge_k$, e_k being the unit vector with a "1" in the k-th position, "0" elsewhere, and ω_k is the input that must be applied to the k-th input terminal so that $x = 0 \circ \omega$.

Just as we could associate maps G_f and H_f with the state module X_f, we can also associate maps G_Σ and H_Σ with X_Σ. These maps are

$$\begin{aligned} G_\Sigma : \Omega &\to X_\Sigma \\ \omega &\mapsto 0 \circ \omega = \sum_{k=1}^{m} \omega_k \cdot g_k \quad , \end{aligned}$$

$$\begin{aligned} H_\Sigma : X_\Sigma &\to \Gamma \\ x &\mapsto (Hx, H(z \cdot x), H(z^2 \cdot x), \cdots) \end{aligned}$$

It is straightforward to verify that both G_Σ and H_Σ are $k[z]$-homomorphisms. We call $(X_\Sigma, G_\Sigma, H_\Sigma)$ the ***module of*** Σ.

Now let us turn again to the realization problem, algebraic style. We first of all note that realizations certainly exist for any f. For example, take $X_\Sigma = \Omega$, $F = \sigma_\Omega$, the left-shift operator on Ω, G = identity and $H = f$. This trivial realization is highly non-canonical and useless. The module machinery given above provides the basis for another realization which is natural, non-trivial and useful.

The underlying idea is to associate with f its module (X_f, G_f, H_f) and to then view this triple as a dynamical system Σ. The first problem is to show that, in some sense, such a realization is the only one that need be considered. To this end, we have the

> ***Fundamental Theorem of Realization Theory.*** *Any two canonical realizations of f are isomorphic as $k[z]$-modules, and therefore as dynamical systems.*

Recall, we say a realization $\Sigma = (F, G, H)$ is ***canonical*** if it is both completely reachable and completely observable. Algebraically, this means that the following matrices have full rank

$$(G, FG, F^2G, \cdots, F^{n-1}G) \quad \text{and} \quad (H', H'F', H'(F')^2, \cdots, H'(F')^{n-1}) \ .$$

In order to actually construct a canonical realization Σ from f, we need the idea of a ***torsion module***. Let R be an arbitrary ring and X an arbitrary R-module. Then if there exists an $r_x \in R$, $r_x \neq 0$, such that $r_x \cdot x = 0$ for each $x \in X$, we say X is a ***torsion*** module. Now let Y be any subset of X. Then the ***annihilator*** A_Y of Y is

$$A_Y = \{r \in R : r \cdot y = 0 \ , \ y \in Y\} \ .$$

The case of interest for us is when $R = k[z]$. Then if X is an arbitrary torsion $k[z]$-module,

$$A_X = k[z]\Psi_X \ , \ \Psi_X \neq 0 \ .$$

We call Ψ_X the ***minimal polynomial*** of X.

Now let us ***assume*** that X_f is a torsion module with minimal polynomial Ψ_f. (***Remark***: this is equivalent to assuming that f has a finite-dimensional realization, i.e., $\dim X_\Sigma < \infty$, when we regard X_Σ as a k-vector space). Under this condition we can represent f by a ***transfer matrix***.

Transfer Function Theorem. Let $f: \Omega \to \Gamma$ *be any* $k[z]$*-homomorphism with the minimal polynomial of* X_f *being* Ψ_f*. Then* f *is uniquely determined by its* $p \times m$ *transfer matrix* W_f *whose columns are the* p*-vector rational functions* $w_k = f(e_k)$, $k = 1,2,\cdots,m$.

Since the transfer matrix is entirely equivalent to the module (X_f, G_f, H_f), we can construct a canonical realization of f by realizing W_f. To do this, we need to make use of the Invariant Factor Theorem for matrices over a principal ideal domain R (such as $k[z]$). Firstly, let us consider the polynomial matrix ΨW (here we drop the unnecessary subscript f). By the Invariant Factor Theorem, we can write

$$\Psi W = A \, \Pi \, B \ ,$$

where A and B are $p \times p$ and $m \times m$ matrices (not necessarily unique) over R, respectively, with $\det A$, $\det B$ = units in R, while

$$\Pi = \text{diag}\,(\lambda_1, \cdots, \lambda_q, 0, \cdots, 0) \ , \ \lambda_i \in R \ .$$

Π is unique (up to units in R), and $\lambda_i \mid \lambda_{i+1}$, $i = 1,2,\cdots,q-1$. Here $q = \text{rank}\ \Psi W$. The λ_i are called the *invariant factors* of ΨW. But what we need are the invariant factors $\{\Psi_i\}$ of W. These are easily obtained from the $\{\lambda_i\}$ by the following procedure. Let $\Theta_i = (\lambda_i, \Psi)$ (= greatest common factor between λ_i and Ψ). Then

$$\begin{aligned} \Psi_1 &= \Psi \ , \\ \Psi_2 &= \Psi / \Theta_2 \ , \\ &\ \ \vdots \\ \Phi_r &= \Psi / \Theta_r \ , \end{aligned}$$

where r is the smallest integer such that $\Psi \mid \lambda_i$ for $i = r+1, \cdots, q = \text{rank}\ \Psi W$. In other words, the Ψ_i are the denominators of the scaler transfer functions λ_i / Ψ after cancellation of all common factors.

In order to carry-out our realization algorithm, it is somewhat more

convenient to factor ΨW over the ring $k[z]/k[z]\Psi$ rather than over $k[z]$. This gives

$$\Psi W = PLQ \bmod \Psi \ ,$$

with det P, det Q units in $k[z]/k[z]\Psi$ with L a diagonal matrix unique up to units in the same ring.

Next, let F_i be a cyclic matrix with characteristic polynomial Ψ_i, e.g., if $\Psi_i = z^r + \alpha_1 z^{r-1} + \cdots + \alpha_r$, then

$$F_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_r & -\alpha_{r-1} & -\alpha_{r-2} & \cdots & -\alpha_1 \end{bmatrix}$$

For any such F_i, we have

$$\Psi_i(z)(zI - F_i)^{-1} = v_i(z)w'_i(z) \bmod \Psi_i \ ,$$

where the components of v and w are linearly independent over k and if $\hat{v}_i\hat{w}'_i = v_i w'_i \bmod \Psi_i$, then $\hat{v}_i = \varepsilon_i v_i$, $\hat{w}_i = \varepsilon_i^{-1} w_i$, where ε_i = unit in $k[z]/k[z]\Psi_i$. Further, define $L = \text{diag}\,(l_1, l_2, \cdots)$, $\mu_i = \Psi/\Psi_i$ and let $p_i = i$-th column of P, $q'_i = i$-th row of Q in the factorization of ΨW. Lastly, let G_i and H_i be the solution of the equations

$$\begin{aligned} H_i v_i &= (l_i/\mu_i)p_i \bmod \Psi_i \ , \\ w'_i G_i &= q'_i \bmod \Psi_i \ , \ i = 1,2,\cdots,q \end{aligned} \qquad (*)$$

This system has a unique solution as long as we agree to let the polynomial vectors on the right sides have degree $<$ deg Ψ_i, which can always be done in view of the "mod Ψ_i" operation.

Finally, we can state the

Canonical Realization Theorem. Every proper rational transfer matrix W may be canonically realized as the direct sum of the sys-

tems

$$\Sigma_i = (F_i, G_i, H_i) \ ,$$

where F_i *is a cyclic matrix with characteristic polynomial* Ψ_i *and* G_i *and* H_i *are the solutions to the system (*),* $i = 1,2, \cdots, q$.

Remarks: (1) The components Σ_i are the elementary "atoms" of the linear system Σ, which canonically realizes W (or f). Thus, knowledge of the invariant factors of the module (X_f, G_f, H_f) essentially determines the realization Σ. Schematically, the canonical realization $\Sigma = \oplus \Sigma_i$ is as shown in Fig. 2. The high level of internal connectivity of this realization should be contrasted with the conventional (and highly non-canonical) realization of W depicted in Fig. 3. This observation is clearly of some significance in the context of brain models, as we shall see later.

(2) While it is not necessary for our development here, it is of interest to ask whether f can be realized *without* knowing the invariant factors of W. Perhaps surprisingly, the answer is *yes*, as first shown in the algorithm of B.L. Ho [8], which is described in detail in [4–5].

The preceding set-up shows that the canonical state space X_f of the object $O_f = (\Omega, \Gamma, f)$ is isomorphic to an equivalence class of polynomials in $k[z]$ mod Ψ_f, where Ψ_f is the characteristic polynomial of the $k[z]$-module X_f. In other words, the system Σ_f is a *pattern recognition* device: the input pattern ω is "remembered" as the state $[\omega]_f$, which is represented by any polynomial ω^* such that $\omega - \omega^* = 0 \bmod \Psi_f$. The simplest such polynomial ω^* is obtained by dividing ω by Ψ_f and designating the remainder as ω^*.

It is well to keep in mind that the abstract property: "Σ_f remembers ω as ω^*" is a coordinate-free property of the system; however, if $\omega(z) = \alpha_1 + \alpha_2 z + \cdots + \alpha_n z^{n-1} \bmod \Psi_f$ is an actual input with $\alpha_i \in k^m$, the

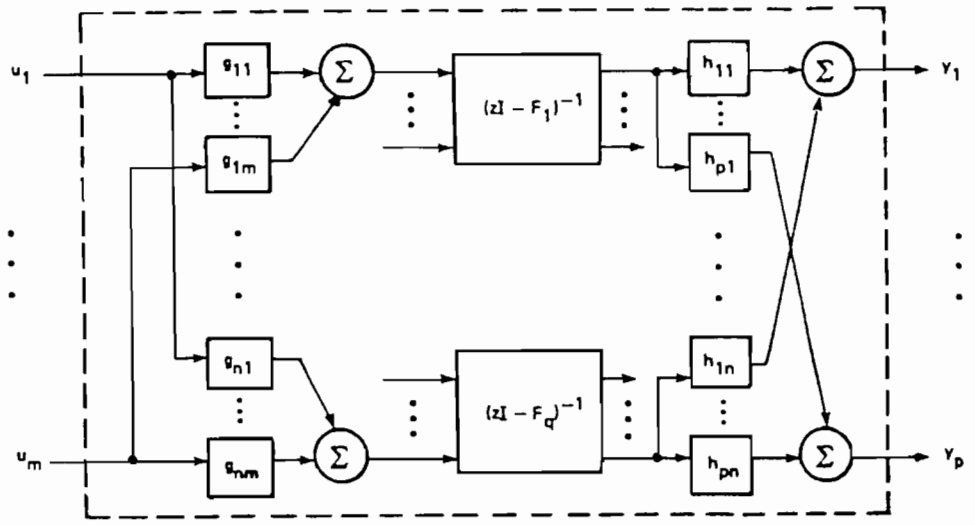


Figure 2. Canonical realization of W.

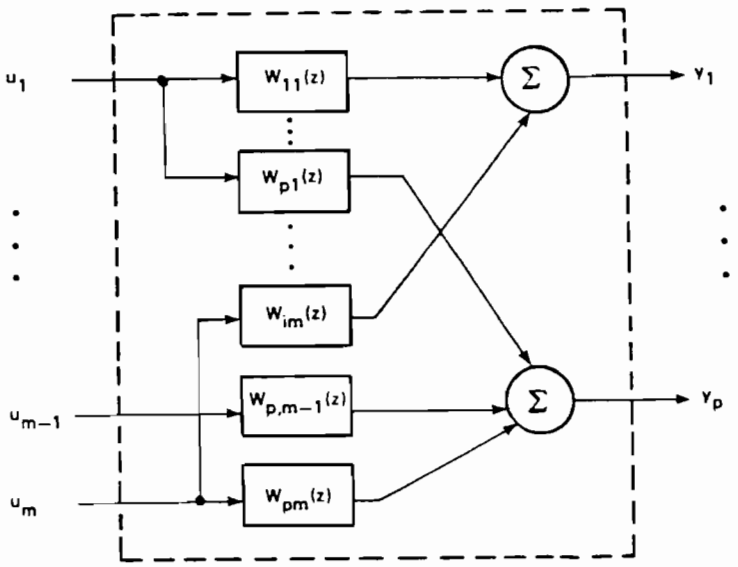


Figure 3. Conventional realization of W.

$\{\alpha_t\}$ *do* depend upon the choice of basis in X_f. Thus, there are many coordinate maps $\omega \mapsto \alpha_t(\omega)$, and the practical, or operational, realization problem really amounts to finding good ways to compute these coordinate maps. This crucial fact was most concisely expressed by Kalman in [9]:

> "In any dynamical system, linear or not, the coordinate maps correspond to abstracting certain characteristic features of the input relative to the basis chosen for the state space. The whole problem of realization may be viewed as the problem of effective computation of coordinate maps. Even if the internal structure of a system is known, its operation in response to inputs cannot be fully understood until it is possible to say exactly what each state says about the corresponding equivalence class $[\omega]_f$ of inputs."

While the algebraic development thus far has been confined to linear input–output maps f, the foregoing statement makes it clear that the underlying principle has broader currency applicable to *any* nonlinear behavior. The trick is to capture the behavior in a usable mathematical form. For linear systems the relevant form is polynomials; for general nonlinear systems the relevant structures are unknown. What is known, however, is that for important classes of nonlinear behaviors the structure of the state space X_f can be algebraically characterized as a manifold or, even more specifically, a variety. For example, in [9] it was shown that the natural state set of a multilinear map f is an n-dimensional algebraic variety V_f embedded in k^N, where N is the number of coordinates necessary to parameterize the Nerode equivalence classes. Thus, the state of such a system is given by N coordinates, of which only n need be stored. The remaining $N-n$ coordinates can then be computed as algebraic functions of the first n coordinates. However, even for simple bilinear systems, examples in [9] show that the state coordinates in k^N are rather complicated functions of the inputs. All of this is very suggestive when interpreted in terms of the way a brain may actually store information. We examine this question in the next section.

7. Pattern Recognition, Brains, and Codes

The algebraic framework outlined above may seem needlessly elaborate, or even pretentiously eccentric, to those schooled in a more traditional approach to control and system analysis via conventional tools of real vector spaces, matrices, Laplace transforms and the like. In this section we attempt to dispel these

prejudices by displaying the power of the module-theoretic machinery to painlessly deal with a number of basic questions that arise in attempts to model brain functions as dynamical systems.

To keep things simple, for the moment let us assume that the stimulus–response map f is linear and that the associated state module X_f is cyclic, i.e., X_f is generated by a single element $g \in k[z]$. Further, let us suppose that $\Omega = k[z]$, i.e., $m = 1$ (a single-input system). Then Ω is generated by 1 (the polynomial which is identically 1). Thus, $X_f = \Omega/\ker f$ is also cyclic and is generated by $g = [1]_f$, and $\ker f\ (= k[z]\Psi_f)$ consists of all polynomials ω such that $f(\omega) = 0$. Let Ψ_f be the polynomial of least degree in this set. Putting all these remarks together, we have the

Representation Theorem. If X_f is a cyclic $k[z]$-module with annihilating polynomial Ψ_f, then X_f is isomorphic to $\{$all polynomials $\pi \in k[z]$: deg $\pi < \Psi_f\}$.

Remark: The importance of cyclic modules in the overall scheme of things comes from the fact that all finite modules over a principal ideal domain (like $k[z]$) are isomorphic to a direct sum of such cyclic modules. The components Σ_i of the Canonical Realization Theorem of the last section are the concrete embodiments of these abstract cyclic modules. Thus, the subsystems corresponding to the cyclic components in the decomposition of X_f form the elementary "building blocks" of the linear system f. The same type of result can also be shown to hold for broad classes of nonlinear systems at the expense of a more elaborate mathematical machinery.

By the Representation Theorem, we may view X_f as essentially a pattern recognition device. The input ω is *coded* as the state $\tilde{\omega}$, where $\tilde{\omega}$ is the polynomial

of least degree in the class $[\omega]_f$. It is potentially of considerable significance that the coded pattern $\tilde{\omega}$, the remainder after dividing ω by Ψ_f, may bear no obvious relationship whatsoever to the original stimulus ω. The following examples from [5] illustrate this point. Letting $\omega = \sum_{i=1}^{r} \omega_i z^i$, if

(1) $\Psi_f(z) = z$, then $\tilde{\omega} = \omega_0$, the last received input value. Such a system "forgets" everything in the stimulus sequence except the last element.

(2) $\Psi_f(z) = z^k$, then $\tilde{\omega} = \sum_{i=0}^{k-1} \omega_i z^i$. This system remembers the last k elements of the input sequence.

(3) $\Psi_f(z) = z - 1$, then, $\tilde{\omega} = \omega_0 + \omega_1 + \cdots + \omega_r$, $r = \deg \omega$. Thus, this system codes the input ω by *adding* all of the stimuli received, i.e., it is an integrator.

(4) $\Psi_f = z^k - \alpha$, then $z^l \omega = \alpha^l \omega \bmod \Psi_f$ for any $l \geq 0$. Thus, the system is sensitive to inputs of period k, while other non-periodic inputs tend to be averaged out. The factor α enables past inputs to be either enhanced ($\alpha > 1$) or diminished ($\alpha < 1$). To see explicitly how this system works, assume $k = 2$, so that the system is sensitive to inputs of period 2. Let $\omega = \omega_0 + \omega_1 z + \omega_2 z^2 + \omega_3 z^3 + \omega_4 z^4$. Then a simple calculation yields,

$$\tilde{\omega} = (\omega_1 + \alpha\omega_3)z + (\omega_0 + \alpha\omega_2 + \alpha^2\omega_4) \ .$$

If ω has period 2, then $\omega_0 = \omega_2 = \omega_4$ and $\omega_1 = \omega_3$. Such a structure will uniformly increase each of the coefficients in $\tilde{\omega}$, while any other structure in ω will tend to affect each coefficient differently. Furthermore, the form of $\tilde{\omega}$ shows that $\alpha > 1$ will enhance past inputs, while $\alpha < 1$ will tend to deemphasize earlier signals.

These examples show part of the range of coding possibilities that each of the cyclic atoms of a linear system can possess. The Canonical Realization Theorem

shows that however complicated the behavior of f may be, it is composed of a combination of elementary behavioral "atoms' of the above sort, interconnected as in Fig. 2. It is the high level of interconnectivity seen in Figure 2 that enables an arbitrarily complicated behavior f to be composed of such elementary behaviors as those in the preceding examples. From an evolutionary point of view, we would not expect to see a non-canonical realization of f as in Fig. 3 because it is just too large and unwieldy (too many complicated components substituting for the high level of connectivity). In short, it is more efficient to interconnect many simple behavior modes than to rely on fewer, more complicated types. And this is exactly what we see when we look at a real brain: the elementary components (neurons) have extremely simple behavior, but the density of interconnections is overwhelmingly large. Such experimental observations strongly argue for the view of a brain as a canonical realization of observed behavior. Now we exploit the module-theoretic structure in order to address two important issues related to the relationship of brains and machines: simulation and pattern recognition.

The *simulation problem* is of direct significance to the question of whether or not machines can emulate the behavior of brains. Imagine that W_1 is the transfer matrix for the stimulus-response behavior of a given organism. We now ask under what circumstances a "machine" with transfer matrix W_2 can simulate W_1. The answer to this question involves the notion of one transfer matrix dividing another.

Definition 1. Let W_1 and W_2 be transfer matrices. Then $W_1 | W_2$ (W_1 *divides* W_2) if and only if there exist matrices V,U over $k[z]$ such that $W_1 = VW_2U$.

If Σ_1 and Σ_2 are the canonical realizations of W_1, then we have

Definition 2. $\Sigma_1 | \Sigma_2$ (Σ_1 *simulates* Σ_2) if and only if $X_{\Sigma_1} | X_{\Sigma_2}$, i.e., if and only

if X_{Σ_1} is isomorphic to a submodule of X_{Σ_2} (or isomorphic to a quotient module of X_{Σ_2}.

From a dynamical point of view, the condition $W_1 | W_2$ means that the inputs and outputs of the machine having transfer matrix W_2 are re-coded by replacing the original input ω_2 by $\omega_1 = U(z)\omega_2$, while the output γ_2 is replaced by $\gamma_1 = V(z)\gamma_2$. Such a change involves delay, but no feedback. If Ψ_{W_2} is the characteristic polynomial of W_2, $d = \deg \Psi_{W_2}$ then the re-coding can be physically carried out with d time units of delay. Under such a re-coding, we clearly have $W_1 | W_2$

$$W_1 = VW_2U \quad ,$$

which is also satisfied if V and U are reduced mod Ψ_{W_2}. Putting all these definitions and observations together, we can easily prove the

Simulation Theorem. Σ_1 *can be simulated by* Σ_2 *if and only if* $\Psi_i(W_1)$ *divides* $\Psi_i(W_2)$, *for all* i.

Thus, a computer with transfer matrix W_2 can simulate a brain with transfer matrix W_1 if and only if each invariant factor $\Psi_i(W_1)$ of the brain is a divisor of the corresponding invariant factor $\Psi_i(W_2)$ of the computer.

Now let us turn to the *pattern recognition problem.* It is clear that one of the characteristic features of human intelligence is the ability to learn and respond to a wide array of external stimuli (patterns). Once a pattern is learned, in some fashion the brain must be able to recognize the pattern again among the myriad patterns presented by the external world. The module setup provides a simple criterion for how this can be accomplished.

Imagine the pattern we want to recognize is represented by the input $\Phi \in \Omega$ and we want to build a machine that fails to react to any other input $\pi \neq \Phi$. By

what has gone before, we know that the machine is totally characterized by its minimal polynomial Ψ, so our problem is: find a machine Ψ that recognizes Φ, but fails to respond to an input $\pi \neq \Phi$. In module language the solution is trivial: we need to find a Ψ such that $\Psi | \pi$ but $\Psi \not| \Phi$. A solution is possible if and only if $\pi \not| \Phi$.

Interpretation. The pattern discrimination problem can be solved as long as the pattern Φ to be recognized is not a multiple of the recognition circuit ***and*** every pattern we want to reject is such a multiple. In order to physically carry out such a discrimination, a brain must clearly have many such elementary "circuits" wired-up in various series-parallel combinations. Referring to Fig. 2, we could imagine each of the blocks in the canonical realization as being one such elementary circuit, the entire circuit being devoted to recognition of a ***single*** such pattern Φ. Then a brain would consist of an unimaginably large number of copies of Fig. 2.

8. Thoughts and Group Invariants

So far, we have created a plausible mechanism at the functional level whereby external behavioral modes of an organism can be coded and decoded via the internal mental states of some kind of "brain". Now we wish to explore the manner in which this mechanism might give rise to what we ordinarily regard as internal "thoughts," as distinct from external behavioral activities. Any decent model for a brain must account for subjective emotional experiences like pain, love, jealousy, pleasure, etc. and not just externally observed actions like motion, sleep, talking and so forth. Such considerations edge dangerously close to the classical mind–body conundrum that has been explored for centuries by armies of philosophers, neurophysiologists, computer scientists and other armchair speculators. The position argued here is probably closest to a central-state identity view on this issue, although it is not incompatible with other materialist views, including

the dual interactionism of Popper and Eccles [10]. In any case, in this section we consider how the system-theoretic machinery can be interpreted in a manner that allows us to use it as a mathematical metaphor for the way thoughts are generated and put together inside a brain. The ideas underlying our development involve a combination of the concepts of a change of state and the notion of system invariants.

According to recent work in brain physiology, the central cortex of the brain consists of around 4 million neuronal modules, each composed of a few thousand nerve cells. Each module is a column that is vertically oriented across the cerebral cortex, and is about 0.25mm across and 2 to 3mm long. These modules are now recognized to be the functional units of communication throughout the association cortex which forms about 95% of the human neocortex. Thus, in the analogy of Eccles[11], a human brain can be thought of as something like a piano with 4 million keys. Carrying the musical analogy a step further, Eccles also postulates four parameters that the cortical modules utilize in generating the virtually infinite number of spatio-temporal patterns that constitute the conscious experiences that can be derived from the brain. These parameters are ***intensity*** (the integral of the impulse firing in the particular module's output lines), the ***duration*** of the impulse firing from the module, the ***rhythm***, or temporal pattern of modular firings and the ***simultaneity*** of activation of several modules.

As a working hypothesis, we will associate each neuronal module with an internal model $\Sigma_f = (F,G,H)$ of a particular behavioral pattern f. The first point to observe here is that even though a module Σ_f is ***originally*** needed to account for the pattern f, once the mechanism (wiring diagram) corresponding to Fig. 2 is physically implemented in the brain, the module Σ_f may generate many other behavioral responses, as well. It is trivial to verify that the output from Σ_f is

$$y(t) = \sum_{s=0}^{t-1} HF^{t-s-1}Gu(s)$$

Thus, Σ_f will reproduce the input–output behavior f as long as the input sequence $\{u(\cdot)\}$ is that originally given as part of f; however, if a different input sequence $\{u^*(\cdot)\}$ is given, then Σ_f will, in general, produce an output sequence $\gamma^* = \{y^*(\cdot)\} \neq \{y(\cdot)\} = \gamma$. As a consequence of this observation, each of our neuronal modules Σ_f will correspond to a particular "learned" behavioral mode f, but it could also produce an infinite variety of other modes $f^* \neq f$, once the neuronal pathways (essentially the connective structure of F, together with the connections G and H linking the inputs and outputs to the states) have been laid down. This type of ambiguity (or lack of one-to-one correspondence) between f and Σ_f can be eliminated by employing the tacit assumption that a standard input sequence is used, generally $u(0) = 1$, $u(t) = 0$, $t \neq 0$. It is tempting to conjecture that much of the processing of stimuli carried out by the body's receptor organs is arranged to implement such a normalization prior to the input reaching the neuronal module. We shall assume that this is the case and that there is a one-to-one match between cortical modules and behaviors.

Under the foregoing hypotheses, there are on the order of 4 million or so "elementary" behaviors, one for each cortical module. These elements of life correspond to the keys on the piano. The intermodular connections coupled with the 4 parameters of intensity, duration, rhythm and simultaneity, then generate all behavioral modes. Let us now take a look at how these elementary behavioral modules could be stored in the brain.

First of all, each module consists of about 2500 neurons capable, therefore, of storing 2500 bits of information. If we assume that a single real number requires 25 bits, then this means that a given module can store around 100 real numbers. If the system $\Sigma_f = (F,G,H)$ corresponding to the module has a state-space of dimension n, and the number of input channels m and output channels p are such that $p,m \leq n$, then to store Σ_f requires $O(n^2)$ numbers. With a brute force storage

arrangement of this type, each module can only have a system Σ_f such that dim $\Sigma_f \leq 10$. But this seems much too small to be able to account for even reasonably complex "elementary" behaviors.

A way out of the foregoing dilemma is to recall that the canonical model Σ_f is determined only up to a change of coordinates in the state-space. Thus, any other model $\hat{\Sigma}_f = (\hat{F},\hat{G},\hat{H}) = (TFT^{-1},TG,HT^{-1})$, det $T \neq 0$, will display exactly the same elementary behavior. By standard arguments, it can be shown [4,12] that as T ranges through the group of non-singular matrices $GL(n)$, there exists a representative of the behavior class of f, call it $\tilde{\Sigma}_f = (\tilde{F},\tilde{G},\tilde{H})$, such that the number of non-fixed elements in $\tilde{\Sigma}_f$ is $O(n)$, i.e., by viewing the states in an appropriate basis, it is possible to represent the behavior f by storing only $O(n)$ numbers. These numbers form ***invariants*** of the group action and completely determine Σ_f. A reasonable conjecture is that evolutionary adaptation has arranged matters so that the "hard-wired" neuronal connections in the cortex are such that the brain represents each learned behavior in something close to this optimal coordinate system in the state-space. Thus, with the same 2500 neurons in each module, it is possible to accommodate elementary behavioral modes f requiring canonical realizations Σ_f such that dim Σ_f can be on the order of 100 or so, an order of magnitude increase over the brute-force storage scheme.

Up to now we have considered each cortical module Σ_f as a means for representing a given ***observed*** behavioral pattern f. But what about internal ***thoughts***? How can we account for aspects of consciousness involving notions like hope, fear, pain, jealousy, hunger, thirst, and other such non-behavioral, but nonetheless real, mental phenomena? Is there any way to accommodate these aspects of consciousness within the systems framework developed above? We contend that there is.

To make progress on the problems of emotional states and thoughts, let us reconsider the diagram of Σ given in Fig.2 and examine the meaning of the blocks denoted there as g_{ij}, h_{ij} and $(zI-F_i)^{-1}$. Our contention is that the elements g_{ij} and h_{ij} are just pre- and post-processors linking the module to sensory effectors/effectors, *as well as to the parts of the brain and to other modules*, while the elements $(zI-F_i)^{-1}$ and the lines into and out of the blocks $(zI-F_i)^{-1}$ represent the internal workings of the cortical module itself. With this picture in mind, let us consider the question of emotional states and cognitive states separately.

There is now a great deal of experimental evidence suggesting that most emotional states (hunger, pain, taste, etc.) have their origin in the *limbic system*, that collection of nuclei and connecting pathways at the base of the brain. If this is indeed the case, then as far as cortical modules are concerned it makes little difference whether the inputs come from external sensory stimuli or from another part of the brain, like the limbic system. From the perspective of the cerebral cortex where our modules Σ_i "live", inputs from the sense organs and inputs from the limbic system are treated equally, and appropriate cortical modules are developed early-on to handle each. In terms of Fig. 2, some of the input channels to the g_{ij} come from sensory receptors, and others from the limbic system. The emotional states arising in the limbic system may or may not evoke observable outputs depending upon the post-processors h_{ij} since, as we know, sometimes emotional states generate observable responses (crying, hunger pangs, violent movements) and sometimes not. In any case, in our set-up there is no need to distinguish emotional states from sensory stimuli, other than that one comes from the outside world, while the other comes only from outside the neo-cortex.

Accounting for cognitive thoughts poses a somewhat more delicate task, since such thoughts are assumed to be self-generated within the cortex itself, quite

independently of stimuli from either the sense organs or other parts of the brain. Our somewhat speculative approach to this problem is to contend that such thoughts are by-products of primary cortical stimulation through the external input channels g_{ij}. We have already asserted that each module Σ_f is established by a particular behavioral mode f, with the g_{ij} conditioned to pre-process the appropriate stimuli, transforming it into a standard form. But it is also the case that each such cortical module shares connections with 10 or so neighboring modules, which may generate stimuli that feed *directly* into the internal blocks $(zI-F_i)^{-1}$, by-passing the preprocessors. Such inputs would, in general, cause the module Σ_f to emit outputs to the h_{ij} that may even result in a behavioral output different from f if the threshold of the h_{ij} is attained.

In general, we may assume that such direct stimuli from the other modules is weak compared to that from the pre-processors so when the "real" input signal for f is present, the "noise" from the other modules is too feeble to influence Σ_f. Note also that in order for Σ_f to be ready to function properly when the right stimuli for f are applied, it must be the case that the matrix F is stable with rather quick damping back to the zero state. Otherwise Σ_f would not be in a position to respond properly to rapid repitition of the same stimuli.

Thus, we conclude that thoughts are generated only when the module Σ_f is in its quiescent state waiting to perform its main function, and such thoughts are generated by the noise present in Σ_f from other modules.

To summarize: the brain's cortical modules correspond to elementary behaviors f, which are represented internally by the objects Σ_f. For compactness and efficiency, we further contend that nature has arranged things so that the objects Σ_f are stored by the invariants of Σ_f under the group of state coordinate changes $GL(n)$. Each such collection of numbers characterizes an entire *class* of systems Σ_f, all of which canonically represent the same external behavior

f. The simplest such element Σ_f of each class contains $O(n)$ parameters, enabling the brain to efficiently reproduce elementary behaviors involving state spaces of dimension on the order of 100. Since there are around 4 million such cortical modules in the human brain, various series-parallel connections of those elementary behavioral/cognitive "atoms" provide ample material for the almost unlimited variety of thoughts, emotions and experiences of human life.

9. Discussion

The work presented here represents a materialistic account of how behavioral and cognitive phenomena can be represented and/or generated in a brain-like object. The basic arguments are schematically displayed in Fig. 4. The central element in this entire scheme is the Realization (or Cognitive) Theorem, which asserts that Behaviorism and Structuralism are formally equivalent. The balance of our argument is then devoted to supporting the thesis that the structuralist view, while equivalent to behaviorism at the formal level, is vastly more useful as a means for investigating the brain-mind problem. Let us now examine the compatibility of the scheme of Fig. 4 with other mathematical models of the brain.

A. *The Hoffmann Model* – in a series of papers [13–15] over the past two decades, Hoffman has developed a model of brain function that is based upon the premise that the neuron is an infinitesimal generator of our perceptions, cognitions and emotions. This model makes extensive use of the correspondence between a Lie group germ and neuron morphology to give a very stimulating account of many aspects of form memory and vision. In this theory, memory consists of invariant recognition under time changes. Hoffman uses the usual mathematical structure governing invariance in the presence of an infinitesimal generator, namely Lie transformation groups together with their prolongations, to establish higher order differential invariants. These structures then show how the memory engram is stored within the brain.

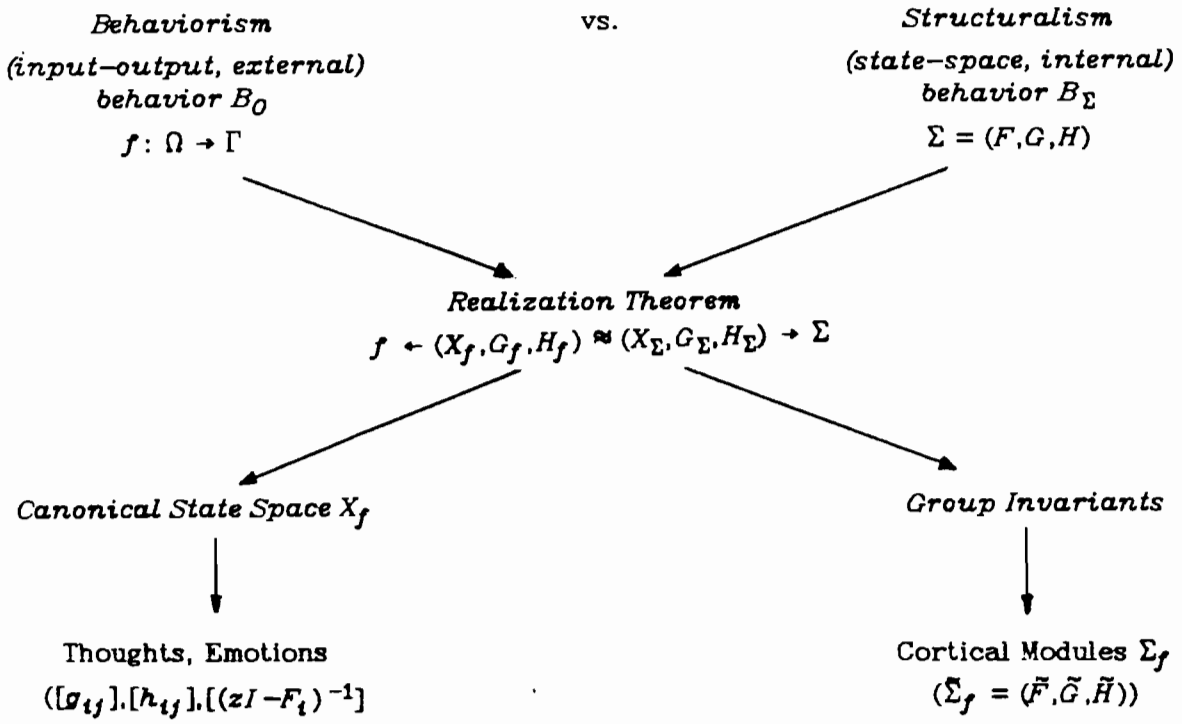


Figure 4. System-theoretic view of a brain.

The extensive development expressed by Hoffman is in no essential conflict with the view presented here. Our treatment of pattern storage (the states X_f) has remained deliberately at the functional level, omitting any conjecture as to exactly *how* these residue classes are *physically* stored in the brain. The treatment by Hoffman provides a quite plausible micro-level means by which the actual brain "hardware" implements the coding scheme that we have proposed.

B. *The Pribram Model* – the neurophysiologist Karl Pribram and the physicist David Bohm have jointly proposed a model of the brain as a *hologram* [16–18]. Roughly speaking, their idea is that the physical brain somehow "tunes-in" or "reads" a holographic universe that exists on some plane transcending space and time. Philosophically, such a concept is a close relative of the dual-interactionism hypothesis of Popper and Eccles [10], whereby mind and brain are two entirely

different entities interacting through some sort of "liaison" modules in the physical brain. We neither accept nor reject this notion, although our sympathies tend toward the central-place identity position outlined earlier. Be that as it may, there is an interesting physically observed aspect of the holographic model that bears upon the ideas presented here. This is the issue of ***distributed memory***.

Beginning with the experiments of Lashley, it has been observed that the memory of a specific event seems not to be localized in the brain. Patients suffering extensive surgical removal of brain tissue due to illness or accident report that memory of specific events is still retained, albeit in a somewhat degraded form. This is exactly the type of result one would expect from the coding of an event as a hologram, in which each part of the hologram contains the *entire* event. At first hearing, such a phenomena would seem to deal a serious blow to our idea of each elementary behavioral mode (event) f being associated with a *specific* cortical module Σ_f. How can it be that the memory of f would be retained if Σ_f were somehow damaged or destroyed? The key to understanding how this can come about is to recall the way in which Σ_f codes the behavior f.

Essentially, the input ω associated with the event f is stored as $\tilde{\omega} = \omega \bmod \Psi_f$, where Ψ_f is the minimal polynomial of the module Σ_f. The problem then becomes that of determining how a different module Σ_g could also code ω as $\tilde{\omega}$. But, this is easy. Σ_g will also code ω as $\tilde{\omega}$ if and only if Ψ_f divides Ψ_g. So, all cortical modules whose minimal polynomial contains Ψ_f as a factor will also code ω as $\tilde{\omega}$. It is not unreasonable to suppose that with 4 million such modules available, each having a minimal polynomial of degree around 10^2, there will be many such modules able to act as a "back-up" for storage of any given event f. Furthermore, since we can expect that for such a back-up module Σ_g, we would have deg $\Psi_f <$ deg Ψ_g, implying that while Σ_g will remember f, it will only do so as a secondary function, its primary function being to recall g. Thus, if Σ_g must be employed due

to the removal of Σ_f, we might expect some corruption or noise in the recall process from the interference by the primary pattern g that Σ_g recognizes.

C. *The Grossberg Model* – in a long series of papers summarized in the book [19], S. Grossberg has presented a theory of learning, perception, cognition, development and motor control that involves a rather elaborate theory of nonlinear processes, emphasizing the role of "adaptive resonances" in neural circuitry to explain behavioral phenomena. While the details differ considerably from those presented here, the work of Grossberg is very much within the same spirit as that given here, and a reasonable conjecture would be that extensions and elaborations of our scheme for nonlinear behaviors, coupled with a closer adherence to experimental results, ultimately may lead to a convergence of Grossberg's work with ours.

The real difference between the Grossberg theory and ours is that we begin with data (as does Grossberg) and then deduce a canonical model using certain global system principles (like reachability and observability) to insure the "goodness" of the model. On the other hand, Grossberg's procedures follow along somewhat different lines. He uses the data to *postulate* certain dynamics containing various tunable parameters which are then adjusted to demonstrate the system's capacity to adapt, learn, remember, etc. But nowhere in the model itself is there any global property that *a priori* insures that this class of models possesses the relevant properties; they must be deduced after the fact. Nonetheless, a closer inspection of the Grossberg models from the systems perspective offered here would undoubtedly pay considerable dividends.

Given the high degree of compatibility of our proposed structure with these much more well-developed models, it is well to inquire as to whether the world needs another brain model such as ours. In what way does our framework hold-out the promise of being an improvement over existing models of the above sort? Our

justification for putting forth the ideas in this paper are primarily aesthetic. The main modeling paradigm set forth here is conceptually simpler than any of the models discussed above and, more importantly, our framework rests on a far less *ad hoc* foundation than any of the models presented earlier. This means that the actual mathematical structure will *emerge* from the experimental data, rather than trying to "tune" parameters in a pre-defined structure to fit the observations.

We noted at the beginning of the paper that the approach presented here is completely speculative, without a shred of experimental existence to directly support it. Validation of the claims made here would involve a research program of many years, involving many neurophysiologists, system theorists and computer scientists. The principal task of such a program is to make meaningful associations between the abstract elements of our paradigm (the sets Ω, Γ, X_f and the elements F,G,H as well as the transducers g_{ij}, h_{ij}, etc.) and actual physical structures in the sensory organs and the brain. Once the associations are made, laboratory experiments can be devised to investigate the claims made here about the way the brain stores behavioral modes in memory, the organization of memory, emotional states, cognitive thoughts and so on. It is easy to make armchair speculations and develop any number of models of the type considered here: only experiments of the type suggested can separate the modeling wheat from the chaff.

At the level of mathematical abstraction, probably the most important work needed is to explore the precise linkages between the formalism suggested here and that put forth in the Hoffman and Grossberg models cited earlier. We have already noted that the same dynamical "spirit" as displayed here pervades their approach as well, but the details are quite different in each case. The advantage of developing the comparisons and interrelationships is that both Hoffman and Grossberg have extensively pursued the experimental/physical interpretations of their models and have correlated many of the abstract mathematical properties

with actual laboratory results. Thus, both the points of overlap and separation of our scheme with theirs would provide an indirect means for linking our framework with experiment.

Let me conclude with a few remarks concerning the emphasis upon *linear* structures in this work. After all, given the enormous complexity of the functions the brain clearly performs, on what grounds can we justify the arguments given here which seem to be highly dependent upon linearity? There are several answers to this objection depending upon the level at which the question is considered. Let us consider some of them in turn.

• *Realization Level* – our main tool has been the Canonical Realization Theorem and, as already noted, the equivalence between a behavior f and a canonical internal state model Σ_f is in no way dependent upon the linearity of f. The theorem is true under very weak hypotheses on f, Ω and Γ. So, at this level, there is no objection.

We have focused most of our specifics upon the case of linear f because it is the situation in which the algebraic ideas can most easily be made explicit, and accessible to non-algebraists. An important aspect of our development was the description of the canonical state space as an equivalence class of inputs. For linear processes, this space can be described explicitly by simple mathematical gadgets – polynomials. For more general f, an explicit characterization is either impossible, or at least algebraically much more complicated as, for instance, when f is bilinear, in which case the state space is an algebraic variety.

Thus, we don't necessarily claim that the intrinsic brain modules are actually linear, only that they are based upon the same *concepts* as given here *explicitly* for the linear case.

• *Dynamical Level* – at another level, one might object that the dynamical processes of observed neural phenomena are so complicated that there *must* be

complicated nonlinearities at work. Perhaps so, but much recent work by Wolfram with cellular automata [20] and a whole array of researchers in chaos theory [21] seems to indicate otherwise. The main message of all of this work is that very complicated dynamics can and do emerge from simple (even linear) local interactions, when the number of interconnected sub-systems is great enough. And with at least 4 million or so cortical modules, it is not unreasonable to suppose that almost arbitrarily complicated patterns might arise in the brain from linear or almost linear building blocks.

• *Approximation Level* – as discussed, we do not claim that the cortical modules are necessarily linear; however, if they are truly nonlinear we have the comforting system-theoretic fact that *any* reasonably smooth behavior f can be arbitrarily closely approximated by a *bi*linear process of the type discussed briefly in Section 5. And, as we have noted, such processes are amenable to the same sort of algebraic treatment we have presented for linear processes.

So, in summary it is not the linearity of f that is important; it is the concept of a canonical realization and the algebraic structure of its associated state space. These are the ingredients that make our magic work.

REFERENCES

1. Fodor, J., "The Mind-Body Problem", *Scientific American*, 244(1981), 114–124.

2. Flanagan, O., *The Science of the Mind*, MIT Press, Cambridge, MA, 1984.

3. Miller, G. ed., *Mathematics and Psychology*, Wiley, New York, 1964.

4. Casti, J., *Dynamical Systems and their Applications: Linear Theory*. Academic Press, New York, 1977.

5. Kalman, R., P. Falb and M. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

6. Casti, J., *Nonlinear System Theory*, Academic Press, New York, 1985.

7. Kalman, R., *Lectures on Controllability and Observability*, Centro Internationale Matematico Estivo Summer Course, Cremonese, Rome, 1968.

8. Ho., B.L. and R.E. Kalman, 'Effective Construction of Linear State-Variable Models from Input–Output Data", *Regelungstechnik*, 14(1966), 545–548.

9. Kalman, R., "Pattern Recognition Properties of Multilinear Machines", IFAC Symposium, Yerevan, Armenia, USSR, 1968.

10. Popper, K. and J. Eccles, *The Self and Its Brain*, Springer, Berlin, 1977.

11. Eccles, J., *The Human Psyche*, Springer, Berlin, 1980.

12. Tannenbaum, A., *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Mathematics, vol. 845, Springer, Berlin, 1981.

13. Hoffman, W., "The Neuron as a Lie Group Germ and a Lie Product", *Q. Applied Math.*, 25(1968), 423–441.

14. Hoffman, W., "Memory Grows", *Kybernetik*, 8(1971), 151–157.

15. Hoffman, W., "Subjective Geometry and Geometric Psychology", *Math. Modelling*, 3(1981) 349–367.

16. Wilber, K. ed., *The Holographic Paradigm*, Shambhala, Boulder, CO, 1982.

17. Bohm, D., *Wholeness and the Implicate Order*, Routledge and Kegan Paul, London, 1980.

18. Pribram, K., "Toward a Holonomic Theory of Perception", in *Gestaltheorie in der Modernen Psycologie*, S. Ertel, ed., Steinkopff, Durnstadt, 1975.

19. Grossberg, S., *Studies of Mind and Brain*, Reidel, Dordrecht, 1982.

20. *Cellular Automata*, D. Farmer, T. Toffoli and S. Wolfram, eds., North-Holland, Amsterdam, 1984.

21. *Order in Chaos*, N. Campbell and H. Rose, eds., North-Holland, Amsterdam, 1983.

22. Churchland, P., *Matter and Consciousness*, MIT Press, Cambridge, 1984.