*Perceptual and Motor Skills*, 1960, 11, 71-88. © Southern Universities Press 1960
Monograph Supplement 1-V11

# A GENERAL FEEDBACK THEORY OF HUMAN BEHAVIOR: PART I

W. T. POWERS, R. K. CLARK, R. L. MC FARLAND
*Veterans Administration Research Hospital, Chicago*

## INTRODUCTION

In this paper we introduce a conceptual model of human behavior, based on some of the fundamental considerations of feedback theory and leading to a generalized theory of behavior. About six years of development lie behind what is presented here, so obviously we cannot explore in this one paper all the ramifications and applications of this theoretical structure which have occurred to us during this period. What we intend to do here is simply to present the theory as concisely as possible, so as to provide a basic paper in the literature to which we can refer when discussing experiments and further theoretical considerations in other papers.

The concepts presented in this paper represent a synthesis of many ideas, some of which have been in print for many years. Indeed, the literature of psychology alone, if interpreted in the light of what is known about feedback control systems, could be used to form the basis for our theory. Our approach did not begin from a psychological orientation but from the physical and mathematical, because the first two authors are physicists, who only after several years of work on this model, began to acquire a more thorough acquaintance with the work of psychologists. Thus, we find it most natural to develop the theoretical model first, before attempting to outline the applications of this model in language appropriate to psychology.

At present we will present in Table 1 just 12 of the references in the literature which have given us key ideas and which have provided us with the necessary conceptual techniques. In later papers we will discuss the contributions of the psychological works mentioned here as well as many others, treating the major theorists and experimentalists in what we hope will be a thorough and orderly manner.

We strongly advise the reader who has something more than passing acquaintance with feedback *not* to skip over the initial parts of this paper in which we develop some of the basic feedback concepts. We have split up the generalized feedback system somewhat differently than is customary, and in our discussion of the operation of this type of system we will be introducing terminology to be used extensively later on in the paper. Furthermore, we have often found that some of our hearers have previously developed misconceptions about how feedback systems operate, which circumstance has led to pointless arguments about the properties of control systems. Before challenging our statements about how the generalized feedback control system operates, consult a servomechanisms engineer!

TABLE I

REFERENCES CITED

1. ASHBY, W. R. *Design for a brain.* New York: Wiley, 1952.

See particularly paragraphs 1/1 through 1/6; note defects in 2/3 and 2/4; note that 2/7 implies strictly a transient-response study. Compare his "primary operation" with our "test of significant variable." Also see 3/11 for lucid discussion of feedback loops and lack thereof in most psychological experiments (Skinner's conditionally excepted).

2. FRANK, L. K., HUTCHINSON, G. E., LIVINGSTON, W. K., MCCULLOUCH, W. S., & WIENER, N. Teleological mechanisms. *Ann. N. Y. Acad. Sci.,* 1948, 50, 187-278.

3. FULTON, J. F. *Physiology of the nervous system.* New York: Oxford Univer. Press, 1949.

Compare "Cerebral cortex: architecture, intracortical connections, motor projections," by Lorente de No. Pp. 288-330. See especially the diagram on page 307 for connections suggestive of upgoing feedback signals (a and a'), outgoing output signals, and imagination connections ($S_9$, $S_7$, $S_5$). Of course far too few connections are shown to perform any complex functions. This geometry is typical of most of the cortex.

4. HEBB, D. O. *Brain mechanisms and consciousness.* Springfield: Thomas, 1954.

5. HEBB, D. O. *The organization of behavior: a neuropsychological theory.* New York: Wiley, 1949.

"Phase sequence" and "cell assembly" are primitive feedback concepts. Many good examples of various orders of feedback control actions.

6. HICK, W. E., & BATES, J. A. V. *The human operator of control mechanisms.* (Monogr. No. 17-204) London: Ministry of Supply, 1950.

7. KORN, G. A., & KORN, T. M. *Electronic analogue computers.* New York: McGraw-Hill, 1952.

See pp. 4-10 for discussion of signal function, block diagrams. Note that the fact that the variables are *identified* as voltages has no bearing on the relationships discussed concerning their *magnitudes.*

8. KRENDEL, E. S., & GEORGE, H. B. *Interim report on human frequency response studies.* Wright-Patterson AFB, Ohio: Wright Air Development Center, Air Research and Development Command, USAF, 1954. (WADC Tech. Rep: 54-370)

A good example of what we are *not* trying to accomplish.

9. SHANNON, C., & WEAVER, W. *The mathematical theory of communication.* Urbana: Univer. of Illinois Press, 1949.

The start of present-day "information theory."

10. SOROKA, W. W. *Analogue methods in computation and simulation.* New York: McGraw-Hill, 1954.

See Preface: rest of book is useful as demonstration that physical form of analogue is completely irrelevant to "behavior;" only relationships among magnitudes of variables are of interest for functional analysis of a system.

11. TRUXALL, J. G. *Control system synthesis.* New York: McGraw-Hill, 1955.

See particularly Ch. 2, "Signal Flow Diagrams and Feedback Theory." Note that roles of arrows in diagrams correspond to boxes in this paper, and nodes correspond to our arrows. Both representations are commonly used.

12. WIENER, N. *Cybernetics.* New York: Wiley, 1948.

See diagram on p. 121: the arrow labeled "input" is our *reference level*: this is thus conceived of as a system with *internal* loops. If X is taken to be our *R,* and "Multiplies Operator," the environment, the equations following describe our system for any one order of control. See also diagrams on p. 132.

## FUNDAMENTAL DEFINITIONS

We will often employ the term "system" in this paper. Much work has been done on general systems theory, but we have found that for our purposes we have needed to formulate our own concepts, for convenience in discussing later ideas.

A system, as we use the term, is a collection of functions (not, as is often proposed, a collection of variables). A function is a relationship among several variables, and a variable is a combination of two classes of percept. Thus, to define "system," we start by defining "percept."

A *percept* is the basic unit of experience. It is that "bit" of perception which is self-evident to us, like the intensity of a light, or the taste of salt. In Part II of this paper we will give another definition which relies less on the subjective sympathy of the reader.

A *variable* is always a combination of two classes of percept. One class contains percepts which *do not vary;* by these percepts we keep track of the "identity" of the variable. The other class contains percepts which do change; these percepts carry the information about the "magnitude" of the variable. "Magnitude" is used here in its most general sense, including the meanings of "intensity," "size," or any other word for the general class of variable attributes.

A *function* is the direct relationship between any two or more variables. We shall uniformly imply by this term a *stable* relationship, which does not alter its form over reasonable periods of time. Since the variables we shall be talking about are assumed to correspond to physical events, we will always assume that whatever functional relationship is seen among variables is imposed by the operation of some physical "device," such as a neural network or a muscle or a chemical reaction. We shall sometimes represent these functions as mathematical expressions, in which case they are to be taken as idealized representations of some physically-occurring relationship.

A *system* is a set of functions interrelated in a special way. Given a set of variables and the physical devices which relate them in pairs or larger groups, we can define the environment of the system as all those variables and functions not included within the set chosen as our system. Within the defined system, in order for just one system to be under discussion, one must be able to trace relationships through the system (variable, function, variable, function, variable. . .) in a connected way such that no chain of relationships is independent of all the others within the system except for effects transmitted through an environmental loop. If the only relationship between two such chains of functions is through an environmental intermediary, then we would count two systems, not one.

The *input boundary* of a system we will define for the present purposes as the set of all functions which relate environmental variables to system vari-

ables *in a unidirectional fashion;* environmental variables affect, through some physical device, a system variable, but the device does not work backward.

The *output boundary* of the system will consist of all system functions which relate system variables to environmental variables, operating unidirectionally in the outward direction.

If any bi-directional function exists at the boundary, we would represent it twice, once as a unidirectional input function and again as a unidirectional output function.

All functions within the system will be treated as above; thus, we will be dealing strictly with unidirectional functions which may be *described* mathematically as working in either direction, but which in actuality operate in one direction only. Thus, for any function in the system we can define a variable or set of variables as the input to the function and a second set as the output from the function. We will often refer to such sets of variables as a single variable.

Finally, when we speak of variables we will be referring exclusively to the magnitude of the variable; its identity is incidental. In other words we are concerned only with information flow, and not with the means by which the information is transmitted nor the physical form in which it is transmitted. Thus, we conceive of the whole system as basically an analogue, not a digital device. Digital functions can, of course, be constructed of such analogue functions. These considerations are not basic to our theory, but might explain some of our biases.

## The Basic Feedback Control System

There are two major classes of feedback in common knowledge. One is the type which is wholly internal to a system, involving closed loops which do not cross the input or output boundaries of the system, and the other is the type in which the feedback path exits through the output boundary, passes through the environment (with attendant modification of the information) and reenters at the input boundary, the rest of the loop being completed within the system. Both types of feedback can exist simultaneously, but only the external type is unequivocally perceivable as a feedback loop by an external observer. The behavior of any system with internal feedback could be simulated exactly by another system with no internal loops, so such internal loops cannot be firmly identified by external observations.

We will be primarily concerned with externally connected feedback loops. Since we will be attempting to build a model of human behavior, we will regularly assume, unless special circumstances dictate otherwise, that the sense of the feedback is *negative;* this is, indeed, necessary if a feedback *control-system* is to exist. The meaning of the term "negative feedback" will become apparent as we discuss the operation of the general control system.

The general control system consists of three functions plus an environment function, and five variables. We will discuss these in order from the input boundary, through the system to the output boundary, and through the environment back to the input boundary.

The input boundary consists of a function we call the Feedback Function, abbreviated $F$ in equations. The environmental variable which is the input to this function we call $v_e$ (which may represent, remember, many variables). The output variables of this function we call the *feedback signal*, "$f$," reserving, as we shall do consistently, the term "signal" for variables inside the system. The feedback signal is some function of $v_e$, the form of the function being determined by the properties of the input device. Mathematically, the relationship would be written

$$f = F(v_e). \qquad [1]$$

The next function is the Comparator Function $(C)$, which receives both the feedback signal $f$ and a *reference-signal*, symbolized as "$r$." The Comparator Function subtracts $f$ from $r$ and its output signal is called the *error-signal*, "$e$," representing the discrepancy between $f$ and $r$.

The function at the output boundary we call the Output Function, $(O)$, which receives the error signal as its input signal and produces the *output-signal* (or variable), "$o$." This would be written

$$o = O(e) = O(r-f). \qquad [2]$$

The Comparator function is often only implicit in the operation of the output function, some devices being capable of responding directly to the difference between two input signals. For clarity we shall usually speak of the Comparator as a separate function and the error signal, $e$, as a real signal inside the system.

The ouput variable $o$ is the input variable to the Environment Function, $(E)$, which in turn produces as an output variable (or set of variables) $v_e$, the input to the system. Thus, the loop is completed: see Fig. 1. We would write

$$v_e = E(o). \qquad [3]$$

For this system to be a control-system, it is necessary that for any error signal, the operation of all the various functions be such as to tend to bring $f$ closer to $r$ (in other words, to reduce the magnitude of the error signal). This is exactly what is meant by "negative feedback." If the environment offers no resistance at all to the output, so that $o$ is capable of altering $v_e$ to any desired extent, then the system will come to equilibrium with the feedback signal equal to the reference-signal. If the reference-signal is altered by some (unnamed) agency, the system will automatically respond to the ensuing error-signal by

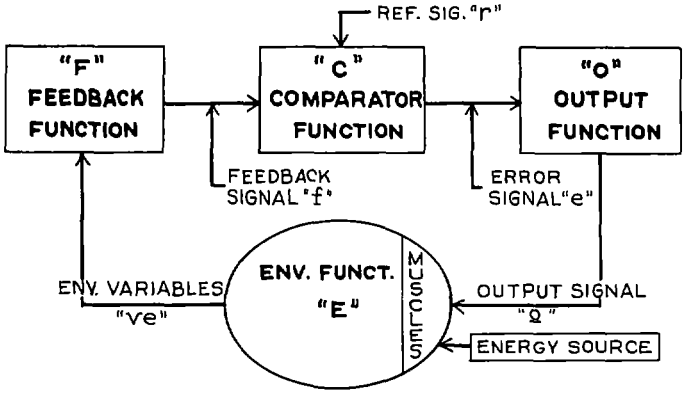


FIG. 1. Feedback control system, general form

bringing $f$ to the same (new) magnitude as $r$, thus erasing the error-signal and simultaneously reducing the output of the system to zero. For a system in this kind of environment, it can be shown that under all conditions within the operating range of the various functions, the feedback signal will be caused by the actions of the system to "track" a slowly changing reference-signal. Thus, the reference-signal is the obvious means by which the system can be controlled.

In an environment which resists the output efforts of the system, or which introduces arbitrary disturbances into $v_e$, the system will still come to equilibrium, but an error-signal of non-zero magnitude will exist at equilibrium; this error-signal (or the discrepancy between $f$ and $r$) will be just sufficient to maintain the output function at the right level of activity to keep equilibrium. In a reasonably efficient feedback control system, the error will be only a small fraction of the total magnitude of the reference-signal; the feedback signal will still be maintained to a reasonable approximation "at the reference-level." Only when environmental disturbances cause some signal in the system to exceed the level its associated devices can handle would we expect to find any appreciable discrepancy between $f$ and $r$.

For the benefit of the reader familiar with transfer-function studies presently being conducted by many workers, we should mention that we are concerned here only with the steady-state relationships in these control systems. We view any such system, therefore, on a time-scale on which transient disturbances occupy so little time that we can neglect them. For some human systems, this may mean that we pay no attention to intervals smaller than 0.1 sec., and for others, that we ignore all events lasting less than several seconds, minutes, or even days. By limiting ourselves to consideration of quasi-static equilibrium, we have found that the over-all organization of a complex system is much easier to conceive. This does not imply that the system is motionless,

but only that all error-signals remain small, the feedback signals normally being maintained at whatever value the reference signal may have for the time being. A system in which all error-signals are comparatively minute could still be engaged in violent activity, as various reference-signals are altered to cope with a changing environment.

A final word on this basic feedback unit. We are going to use it as the building-block (with some modifications) of a complex many-leveled system. If we were faced with the task of designing such a system that would actually be overall-stable, not oscillating wildly or locking itself up in internal conflicts, we would give up right here. Fortunately, we are not concerned with design criteria, for the human system we deal with is normally very stable, with no crippling conflicts and no obvious uncontrolled oscillations going on. Thus, questions of stability criteria, non-linearities, limits, and the like do not concern us in our basic attempt to construct a man-like system. We assume that the various functions have forms, including transient response terms, which result in stability, so that by leaving the details of the functions unspecified, we have by definition a stable system. Later on, when the model is completed, we can consider a few of the pathological conditions that might correspond to conflict among feedback systems and various forms of instability.

### Aggregates of Feedback Control Systems

Let us consider a collection of functions in an extensive system (which may in some cases prove to be more than one system). As we have already noted, some of these functions will be members of the input boundary, others of the output boundary, imposing relationships between system and environmental variables, in one direction or the other.

Some of the boundary functions will be found to form feedback control-systems (in pairs, one input system and one output system) with perhaps some intermediate function within the total system. All such boundary feedback systems will classify as first-order systems. In the human being, these boundary systems correspond largely to what have been unfortunately labelled as the "spinal reflexes." The spinal reflex systems are fairly efficient control-systems having proprioceptive inputs and motor outputs and receiving reference-signals both in the output function (muscle-bundle) and in a comparator function (ventral horn cells). Indeed, these first-order systems almost monopolize the output facilities of the organism. There are input functions, however, which are not part of these control-loops.

Idealizing from this neurological hint, we will restrict our model so that *all* its ouput boundary functions belong to first-order control systems, and none are controlled directly and exclusively by "higher" systems. We allow some input functions to generate signals within the system which are not part of first-order control-loops.

In the human systems, it is the rule that many first-order systems affect the same variables in the local environment and thus affect each others' input variables $v_o$. It will be common, then, that many first-order systems will act as environmental disturbances on the inputs of other first-order systems. These disturbances will be corrected, or at least resisted, by each local system, and chaos will obviously result if reference-signals are not properly coordinated.

We can now select out of all the remaining functions in our system those which form *second-order* control systems to perform this coordination. These control-systems will receive not only the output signals from some of the "unused" first-order input functions, but will also receive as inputs the same variables which serve as feedback signals in the first-order systems [in the human system, it is well-known that the proprioception feedback signals in the first-order spinal loops (and peripheral nerves in the cranium) divide, one branch going to more central systems].

Thus, if we wished we could now define a second-order input and output boundary; crossing the input boundary will be all or most of the signals generated by first-order feedback functions, whether involved in the first-order loops or not, and crossing the output boundary will be a set of output signals which enter the first-order systems. These signals cannot be considered as adding to the outputs of the first-order systems, because feedback systems tend to go into violent conflict if their outputs are tied together, thus inactivating those systems (the theory of conflict will be discussed later). The only feasible control-point is the reference-signals of the lower-order systems; therefore, in our model we identify (for the time being) the ouput signals of second-order systems with the reference-signals of first-order systems. To put it graphically, the output of a second order system is not a muscular force, but a goal toward which first-order systems automatically adjust their input signals (proprioceptive sensations). Thus, the second-order system acts, so to speak, by specifying for the first-order system the kind of sensation it is to seek; the first-order system adjusts its output until its input signals match as closely as possible, in the given environment, the "example" given by the reference-signal, thus (quite incidentally) producing environmental effects which an external observer could see.

This viewpoint is extremely important to understand: in all the feedback systems we will discuss, it is of no concern at all to the feedback system what actual effects are produced in the environment. The system reacts only to the signals injected into it by its feedback function, and for any one system nothing else exists. Even when we speak of systems which deal in human interrelationships, these complex systems not only do not "care" about what is actually going on in the "real" environment, they cannot even know what is going on "out there." They perform the sole function of bringing their feedback signals, the

only reality they can perceive, to some reference-level, the only goal they know. If we were discussing servomechanisms, such anthropomorphisms would be unnecessary, but when we are talking of the very systems in which we live, now and always, which we must employ even to think, anthropomorphism is an essential ingredient of understanding.

It is evident now that we could go on defining successively higher orders of control until we had exhausted our collection of functions. We would then find all the sub-systems, each a feedback control system, arranged in a hierarchy (or many overlapping hierarchies) in which a system of any one order perceives an environment made up of the feedback signals of the systems in the next lower order, and which acts to change that environment by producing output signals which are the reference-signals of the same lower-order systems. This structure is exactly the basic organization of our model. A model of this type could be constructed (ignoring practical difficulties) which would reproduce any kind of human behavior that did not involve changing the form of any functions or adding new systems to the structure: the model thus far is intended as a model of those human systems which produce learned behavior, *after* learning has taken place. This model, being built entirely of feedback control systems, is inherently capable of maintaining dynamic equilibrium (error-signals small, but not necessarily a physically static system) in the presence of a wide variety of environments, both familiar and strange. It is "adaptive" to the extent that it can cope with a large variety of new environmental *configurations,* but it cannot do a thing about an environment which changes its *properties* (summed up as the *E*-function in Fig. 1). We still lack something to account for non-rote learning, for that requires altering the *structure* of the system, not merely its information content.

## The Negentropy System

We borrow the term "negentropy" from information theorists to refer to the process of decreasing entropy in a local system (at the expense, of course, of increasing entropy elsewhere), which process has been identified by some with an increase of organization within a system. We conceive of the central nervous system as being a collection of neurones forming a complex and largely random network, which can have its effective structure altered by activating and inactivating connections within the net to produce networks with semi-permanent and well-defined functions, which to human beings would appear less random.

The processes which alter the connections within the basic bed of "uncommitted neurones" (McCullouch's term) to form the various orders of feedback control must themselves represent the working of a system which is *not* the result of learning, but which is present and active from birth or before. This system may be physically indistinguishable from the resulting learned systems

(perhaps it is implicit in the "random" connections in the "unorganized" neurones), but it is functionally quite different. Its output must be complicated and must extend throughout the CNS, because systems which have been learned are apparently subject to further modifications or additions. Rather than attempt to postulate what the nature of this output must be, we will define it simply in terms of what it must do.

The output of the *N*-system, we hypothesize, results in the following kinds of events. (1) Uncommitted neurones in physically suitable regions become tentatively organized to process a number of feedback signals from the highest existing order of control (which in the beginning may be first order). (2) Other uncommitted neurones likewise undergo tentative organizations which generate signals serving as reference-signals for the next lower order of system. (3) These tentative organizations of input and output can occur at a variable rate. (4) When a particular organization has occurred often enough[1] within a collection of uncommitted neurones, the organization tends to persist, and the input and output functions of a new order of control system have been formed (as Hebb and others have suggested).

Thus, we have *identified* the output variable of the *N*-system as "the processes which alter organization in uncommitted neurones" (as well as in existing systems). The *magnitude* of this variable we postulate to be measured by the rate at which new organizations are formed one after the other.

The changing organizations occurring in potential output functions will result in a continuous alteration of the reference-signals in the momentary highest-order systems; this results in observable trial-and-error behavior, which shows some organization owing to the existing hierarchy. The continuing reorganization occurring in the new input function does not have such externally-observable results, but is subjectively recognized as a kind of trial-and-error effort to perceive new patterns, a common experience in a learning situation which includes what we experience as tentative formulation of hypotheses. The "hypotheses" here should be thought of as tentative definitions of new variables, which may or may not prove to repeat themselves in experience, depending on the organization of lower-order perceptual functions and the properties and nature of the environment.

The input variables which affect the input boundary of the *N*-system we call "intrinsic signals;" we suppose these to be a set of sensory signals which are measures of a set of physiological states, including but not necessarily limited to the ones commonly associated with the "drives." When these variables are each at some certain critical level, the organism is operating optimally, as far

[1] "Often enough" means one or more times.

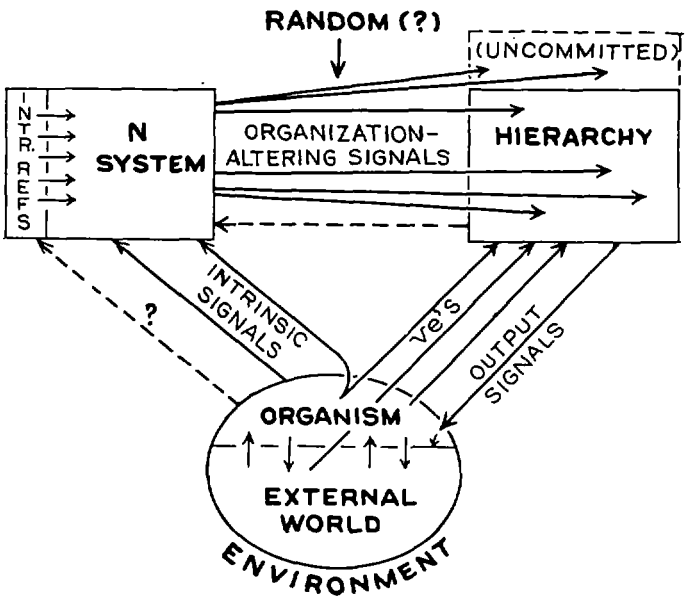


FIG. 2. Overall organization in model

as the *N*-system is concerned. There may be many effects, such as those due to radiation damage, which are deleterious to the organism, but which are not directly represented by intrinsic signals.

The *N*-system we assume to be a feedback control system which is organized to maintain the intrinsic signals at particular reference-levels. These reference-levels may be set by neural signals (as, perhaps, for sex or hunger signals) or they may be determined by the physical properties of the *N*-system functions. In either case, the reference-"signals" must be genetically determined, not determined by experience, for the *N*-system must be a complete control system (which implies reference-signals in existence) before any learned system can be developed. When all intrinsic reference-levels are satisfied by their respective signals, we say the organism is in its *intrinsic state.*

The overall operation of the *N*-system is thus very easy to describe (see Fig. 2). If some event occurs which makes one or more of the intrinsic signals depart from its reference-level, the *N*-system produces an output signal proportional (as a first approximation) to the error. Since the output signal has been defined as a rate of reorganization of neural networks, the net result is to establish a certain rate of attempting to learn. We would say "rate of learning" except that whether or not anything can be learned by reorganization depends to an important degree on the nature of the environment. If the reader

will keep in mind this hedge, we will after all use the more convenient expression "rate of learning."

Simply put, the rate of learning is approximately proportional to the intrinsic error signal, and this is a fundamental property of the human organism.

A *particular* organization will become a stable learned feedback system not because there is anything that "tells" the system to stop reorganizing, but because the lower-order systems and the environment are such that this particular organization produces behavior which results in a lessening of the intrinsic error, thus slowing or halting the reorganization process. If the same organization proves to have an intrinsic-error-reducing effect several times, then reorganization will stop with the new higher-order system in approximately the same form several times, and we suppose that this will cause the organization to tend to persist,[2] or even to become a semi-permanent part of the hierarchy of learned systems. This kind of learning has many evolutionary advantages; for one, a new system will not be fixed for every chance arrangement of the environment, but only for situations which tend to repeat. Another advantage is that while reorganization will stop with the new system in *approximately* the same form as before, there will tend to be differences in detail, so that the "noise level" is reduced, much as one eliminates irrelevant variations from planetary photographs by superimposing many negatives to form a composite print.

## Modifications of the Basic Feedback Unit

Our model so far has many properties like those of human beings, but we are lacking several important ingredients (at least!). The model has no memory for past experiences, it cannot use past information in present actions, and it is incapable of imagining (which we define as the ability to perceive sensory events generated internally rather than generated by present-time interactions at the input boundary of the whole system). As we consider them, memory and imagination are fundamentally related.

To see how we propose to introduce the function of memory, refer to Fig. 3. A new block has been added labelled "*R*," which stands for the *recording function.* We assume that there is a recording function associated with *every individual feedback subsystem* (associated functionally, not necessarily in space).

This recording function has an input which is the same feedback signal used in the local feedback loop and sent to higher-order systems. The function *R* receives this signal and by some means neither we nor anyone else understands,

[2]Because this form is, therefore, approximately adequate for control of the existing environment and hence will be changed further but little. This does not imply or deny a frequency theory of learning, for each organization that exists when learning ceases has been "learned," whether or not it is learned completely and whether or not it is an appropriate form. In this sense, learning is always complete, but perhaps does not match what *E* has in mind as the "proper" final organization.

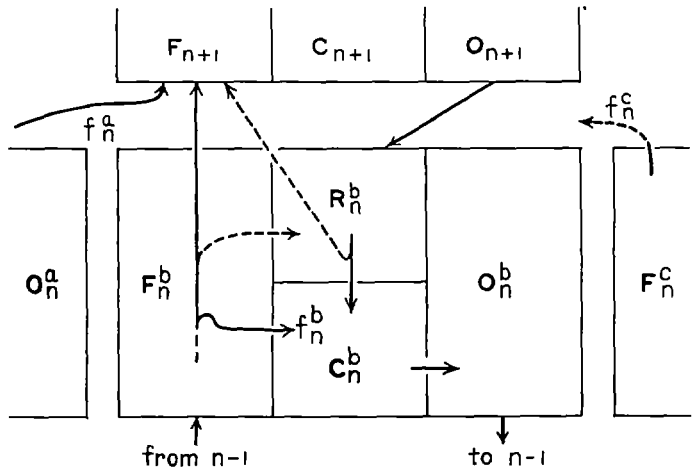


FIG. 3. Relationships among orders

records the information carried by it. The result is a set of recordings which may be permanent or which might have some finite half-life. (There is no present way to tell whether forgetting is due to fading of the recordings or to failure of the recovery apparatus.)

The recording function has the further property that when it is selectively stimulated by a signal external to the local system, it will produce a signal which is a facsimile of the signal that was recorded. This reproduced signal carries the same information, or some significant portion of it, that the original feedback signal carried. To all intents, it is a sensory signal, but one arising from a past event rather than a present one. Current experiments in brain stimulation tend strongly to suport this view of memory.

It will be noticed that the signal from a higher-order system in Fig. 3 no longer serves directly as the reference-signal for the pictured system. Rather, the higher-order output signal stimulates a memory-trace in $R$, which in turn produces a signal that is used as a reference-signal in the associated subsystem. Thus, the reference-signals which control a given feedback unit are examples of its own past sensory signals, and one could now express the task of the control system as being that of reproducing in present-time experience some previously-experienced perceptual field, or portion thereof. To some degree new perceptual fields could be demanded and brought about by stimulation of combinations of memory-traces. Rote learning could occur in the form of new recordings and hence an enlarged repertoire of reference-signals.

The process of selecting a memory trace and stimulating it might be a function of $R$, or it might result from some property of higher-order output

functions. We have not tried to specify the processes involved any further than our statements about what we assume to happen. In either case, the overall effect is that higher-order output functions act by stimulating memory-traces in lower-order recording functions.

We have come to associate perception with feedback signals, and specifically *not* with output signals. A moment's introspection will convince the reader that he *never* perceives an output signal in his own system. Even muscular forces are perceived as proprioceptive sensations. Thus, if the objects of perception must all be the signals *f*, our model still cannot remember! It cannot, that is, perceive signals arising from its memory-traces, because as we have drawn it so far, the reference-signal that is the remembered feedback signal enters the comparator function, which is associated with *O*, not *F*. We have a situation of some psychological interest wherein our model can reproduce a past experience without being able to perceive that experience.

The reference signal carrying "imagined" information cannot be properly interpreted by the feedback function *F* of the associated system of the same order, at least not in general. This is best demonstrated by an example.

Suppose that *F* receives a single variable $x$ and squares it to produce a new variable $y$:

$$y = x^2. \qquad [4]$$

If this new variable $y$ were to appear at the input of *F*, a new variable $y^1$ would be generated, equal to $y^2$ (because the function always performs the same operation on its inputs). Thus, we would have

$$y_1 = y^2 = x^4. \qquad [5]$$

We see that the new variable $y_1$ represents $x^4$, which is not the same "interpretation" given to other lower-order signals received by *F*. Thus, the system could not act correctly with respect to such a twice-processed variable if it were set up to handle variables representing $x^2$.

It is true that certain functions will not introduce such a distortion if applied to their own output signals (e.g., if $y = x$, then no distortion will result from any number of reprocessings), but the general structure of the model cannot be made dependent on such special cases; the way the model is to handle the imagination information must work for *any* form of *F*.

If the reference-signal is indeed a reproduction of a past feedback signal, then it bears the same relationships to lower-order signals as do present-time feedback signals in the associated system. Therefore, in view of the previous paragraph, if the reference-signal were to enter a feedback function of the next higher-order, it would always be interpreted properly, just as are the feed-

back signals currently present. Consequently, we introduce into the hierarchy what we call the "imagination connection," shown in Fig. 3 as a dotted line splitting off from the reference-signal in one system and entering the feedback function of the controlling higher-order system.

This connection is shown dotted; its introduction must be qualified because of the effects of having this connection present.

Note that the higher-order system would find its feedback signal at the required reference-level solely on the basis of the imagination signals from lower-order, even though the lower-order signals might be quite far from the reference-levels in the lower-order systems. This could occur if the higher-order *F* received imagination-signals *in preference to* feedback signals; a condition like dreaming or fantasy would occur, in which every goal set for the lower-order systems appeared to be immediately satisfied—in imagination, of course. This might seem clearer if it is remembered that normally the higher-order system specifies a reference-signal which the lower-order system matches with its own feedback signal; if the reference-signal substitutes for the feedback signal, the "match" is automatically ensured.

The imagination signal makes it possible for our system to perceive reproductions of past perceptual signals (that is, to remember as well as record), to plan an action "mentally" without actually performing it, to hallucinate, and as mentioned, to dream.

Obviously, the hierarchy could not perform very reliably in a real and sometimes dangerous environment if its actions were completely "short-circuited" by the imagination connections. Somehow this configuration must contribute more information to the perceptual field at some times, less at others. Under conditions of sensory deprivation, it apparently provides a great deal of information, while under conditions of, e.g., immediate danger (barring pathology) it contributes little. Everyone knows that the more thoroughly one wraps himself in perception of internal events—thoughts, memories, daydreams—the less sensitive he becomes to the present environment. There appears to be a kind of mixing control, which can be adjusted to full imagination (as when asleep) to full present-time perception. This might be a property of the feedback functions, corresponding to a shift in perceptual attention, or of the manner in which output functions stimulate lower-order recordings. We are open to suggestions.

The normal condition is probably one in which most information is present-time perceptual information, and small errors are filled in by the imagination connection—this would be a pro-survival property, in that it would allow the feedback systems to be very exact in their control-actions, while not tying them up over trivial discrepancies. The phenomenon of "filling in" small discrepancies is well-known under the label "closure."

## SUMMARY OF PART I

What has been presented so far is a model, a collection of functions which handle signals, arranged into a hierarchical structure and composed of elementary feedback control-systems of the external-loop type. For the feedback systems of any one order of control, the environment consists of a set of feedback signals, the same ones used in the control-loops of the next lower order; this environment is controlled by means of signals sent into the lower-order recording functions.

This set of systems is controlled by signals from higher orders or from random reorganizations of potential higher-order output functions in the bed of uncommitted neurones; such control signals stimulate the recording functions in the controlled system so as to give rise to reference signals, reproductions of past feedback signals produced by the local feedback functions.

The rate at which reorganizations take place in this hierarchy is proportional to the degree of intrinsic error existing in the *N*-system, which is a feedback control-system of the external-loop type concerned with maintaining a set of intrinsic variables at their genetically-determined reference-levels; the function of the *N*-system is to maintain the organism in its intrinsic state, or as near to it as possible. The output action of the *N*-system is conceived of as essentially random.

While we have made occasional reference to psychological or neurological properties of human beings as a means of making certain points more acceptable, this portion of the paper has been primarily concerned with presenting the structure of our model, not its application to understanding human behavior. Part II will deal with the problem of translating from this functional scheme to terms appropriate to human beings. The two parts are (understandably) reversed from the order in which this whole picture was developed.

The operation of this model can be summed up perhaps more clearly in plain language. A system at a given order has goals given to it by higher-order systems. These goals are in the form of perceptual images of past experiences or combinations of past experiences. The system acts to make its present perceptual field match the goal-field as nearly as possible. It does not act directly on the external world, but on the only environment with which it is in immediate contact, the set of next-lower-order systems. Its action is that of selecting and stimulating goals for lower-order systems; it is capable of perceiving the signals (either feedback or reference) resulting from its selection, so a set of lower-order signals can be specified which, if achieved, would be interpreted by the system's own feedback function as the required magnitude of perceptual variable.

Only first-order systems act directly on the (non-CNS) environment.

## COMMENTS

To an external observer the behavior of this model could, in principle, be interpreted at many different levels, each quite correctly. This follows from the fact that the feedback signals at a given order are variables which represent the collective behavior of some set of lower-order variables, and so forth down the chain of command, so that at each order we find the feedback signals corresponding to variables abstracted farther and farther from the original raw sensory data and individual environmental events. Each order of system acts on the lower-order systems until it perceives its own kind of variable as being at the required reference-level. It will alter its outputs to the lower-order systems to counteract environmental events which have, via intermediate perceptual interpretations, a disturbing effect on the feedback signal.

Thus, if one knew the kinds of transformations that characterized the transition from perception at one order to perception at the next order, he could observe the environment of the system under study and make parallel abstractions of his own; he could thus define *n*th-order variables in the environment of the other system, and watch how the other system interacted with those perhaps quite abstract variables. He could tell if those variables were actually under feedback control by the other system simply by applying forces to the environment which tended to alter those variables (but not inexorably, else no feedback action could occur) and watching to see which if any were maintained constant by the behavior of the system. He could tell whether he had abstracted correctly (to any desired probability of correctness) by applying all the different kinds of disturbance he could think of; if the variable were maintained constant or nearly so against all these disturbances, he could be fairly sure he had abstracted properly; that is, in the same way that the subject system's feedback functions abstract. By the same token, he could discover the reference-levels at which these variables are being maintained.

Given enough acquaintance with the system under study, the observer would see that the system is *always* maintaining *all* orders of perceptual variable at some momentary reference-level, by an active error-correcting process, except when its abilities are overwhelmed by superior forces in the environment. Even then, the higher-order systems will compensate by readjusting the reference-levels of lower-order systems, which might be seen as a drastic shift in the whole mode of behavior—from fighting to fleeing, perhaps. Whether fighting or fleeing, however, the lower-order systems would still be seen to control successfully patterns of movement, coordinate spatial relationships, produce vector forces, and so forth in a stable and disturbance-resistant manner.

If a human being is indeed this sort of functional being, we can find out more about what is going on inside him if we can learn to understand the

various classes of perceptual variable which are involved in his feedback control-systems. The method of disturbing and testing, which we call the "test of the significant variable," is one method, and it is wholly scientific in its procedures, but fortunately we need not go through this tedious process to obtain every bit of information we are going to accumulate. Both the human subject and the investigator are presumably similar creatures, and the investigator can often find short-cuts by an introspective analysis of his own perceptual methods. This, of course, cannot be done in the sense that the investigator cannot perceive his own perceptual apparatus. He can, however, attempt to discover those variables which in his experience are *self-evident* classes, that is, which to his knowledge and belief are the forms in which he must perceive and always has perceived his universe. This approach is naturally subject to errors of idiosyncrasy, but the results, in the form of classes of variables which should be significant to other human control-systems, can be subjected to the test of the significant variable, and false or inaccurate guesses eliminated.

To give the reader an advance notion of what we mean by a "self-evident class of variable," consider the referent of the term "sequence." This is one of the self-evident classes. We do not mean that everyone calls this part of his experience by the term "sequence," or even by any related term. That is part of verbal behavior. What we mean is that we think every human being can perceive the difference between experience A occuring before B, and B occuring before A, provided the limits of perception are not approached. He can set up a control-system that is capable of reproducing a past sequence of simple events correctly, in the same order as originally. If he cannot do this, he cannot talk, he cannot reason, he cannot even detect the passage of time. If he did not perceive and control variables of sequence, he could not be sure of walking forward rather than backward, and although he might be able to recognize his telephone number visually, he could not dial it.

Furthermore, "sequence" is a unique category, qualitatively different from other categories. A simple sequence (the least element in a sequence of sequences) is perceived as an entity different from any of the individual static *configurations* of which the sequence is always composed. A sequence can be maintained even though the individual configurations used to produce it change. I can hum "Shave and a haircut, six bits," or I can drum out the rhythm on the table, or I can reproduce the rhythm by generating nine different sensory impressions in the right order (pauses and sounds): 0—&@%,— —½¼. But I must always employ some set of static configurations, for that is another self-evident class of perception, and it is the *next-lower order* of perception.

Discussion of these categories of perception (which is sufficient to define categories or orders of control-system) will occupy most of the second section of this paper.

*Accepted June 19, 1960.*