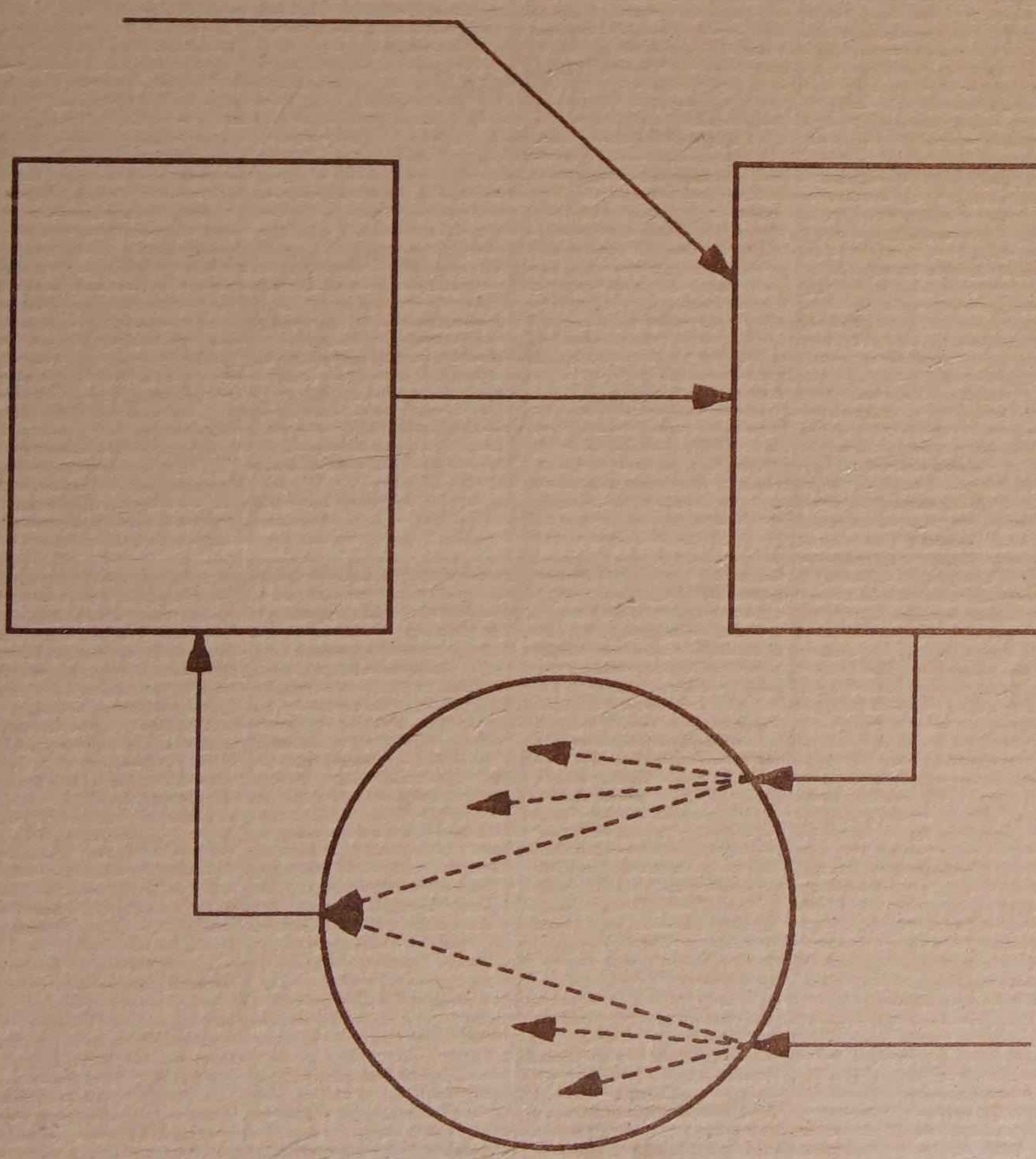
LIVING CONTROL SYSTEMS II

Selected Papers of William T. Powers



LIVING CONTROL SYSTEMS II

1. There have been two paradigms in the behavioral sciences since 1600 AD. One was the idea that events impinging on organisms make them behave as they do. The other, which was invented in the 1930s, is control theory. We are going to explore the record of these paradigms.

2. Control theory explains how organisms ~~~~ control what happens to them. This means all organisms from the amoeba to ~~~~ human being. It explains why no organism can control another without physical violence. It explains why people deprived of any major part of their ability to control soon become disfunctional, lose interest in life, pine away, and die. It explains why is is so hard for groups of people to work together even on something they all agree is important. It explains what a goal is, how goals relate to behavior, how behavior affects perceptions, how perceptions define the reality in which we live and move and have our being. Control theory is the first ~~~~ scientific theory that can handle all these phenomena within a single testable concept of how living systems work.

B.B. Powers

1. There have been two paradigms in the behavioral sciences since 1600 AD. One was the idea that events impinging on organisms make them behave as they do. The other, which was invented in the 1930s, is control theory. We are going to explore the second of these paradigms.

2. Control theory explains how organisms ~~ean~~ control what happens to them. This means all organisms from the amoeba to ~~mankind~~ Humankind. It explains why one organism can't control another without physical violence. It explains why people deprived of any major part of their ability to control soon become disfunctional, lose interest in life, pine away, and die. It explains why it is so hard for groups of people to work together even on something they all agree is important. It explains what a goal is, how goals relate to behavior, how behavior affects perceptions, how perceptions define the reality in which we live and move and have our being. Control theory is the first ~~seien~~ scientific theory that can handle all these phenomena within a single testable concept of how living systems work.

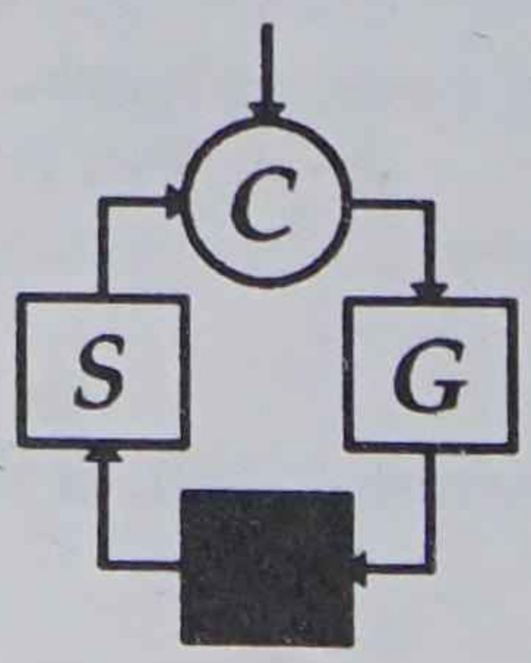
Bill Powers

Holograph written on November 3, 1991, at Valencia, California. Courtesy of Dag Forssell.

About CSG

The Control Systems Group (CSG) is a membership organization supporting the understanding of cybernetic control systems in organisms and their environments: *living control systems*. Academicians, clinicians, and other professionals in several disciplines are members of CSG. Annual meetings of the Group have been held since 1985. CSG publishes books, a newsletter, and *Closed Loop*, a quarterly compilation of discussions on CSGnet, an electronic mail network for individuals interested in control theory as applied to living systems. For additional information, contact The Control Systems Group, Business Office, 73 Ridge Pl., CR 510, Durango, CO 81301; phone 303-247-7986.

The CSG logo, designed by Mary A. Powers and Gregory Williams, shows the generic structure of cybernetic control systems. A Comparator (*C*) computes the difference between a reference signal (represented by the arrow coming from above) and the output signal from Sensory (*S*) computation. The resulting difference signal is the input to the Gain generator (*G*). Disturbances (represented by the black box) alter the Gain generator output signal on the way to Sensory computation, where the negative-feedback loop is closed.



LIVING CONTROL SYSTEMS II

Selected Papers of William T. Powers

The Control Systems Group, Gravel Switch, Kentucky

Production Coordination: Gregory Williams

Printing: Diamond Graphics, Lexington, Kentucky

Typeset using *PowerForm* and *PictureThis* on a DTK Tech-1230 computer, with camera-ready output produced by a NEC Silentwriter LC-890 printer.

Library of Congress Catalog Number: 89-62634

ISBN 0-9624154-2-1

Copyright © 1992 by William T. Powers

All rights reserved

The paper in this book meets the guidelines for permanence and durability of the Committee on Production Guidelines for Book Longevity of the Council on Library Resources.

Printed in the United States of America

1 2 3 4 5 6 7 8 9 0

Contents

Foreword, *W. Thomas Bourbon* *ix*

Preface, *William T. Powers* *xv*

A Note on the Text, *Gregory Williams* *xvii*

Some Implications of Feedback Theory Concerning Behavior *1*

The Wiener Feedback Model— A Strategic Error *11*

Emotion *31*

An Experiment with Levels *41*

The Illusion of Control *55*

Control Theory for Sociology *67*

CT Psychology and Social Organizations *91*

A Bucket of Beans *129*

Deriving Closed-Loop Transfer Functions for a Behavioral Model, and Vice Versa *145*

Learning and Evolution *161*

An Agenda for the Control Theory Group *171*

Control Theory, Constructivism, and Autopoiesis *175*

Control Theory: The Road to Utopia *189*

On Violence and Aggression *197*

The Good, the True, and the Real *201*

A Manifesto for Control Theorists *207*

After Galileo *213*

The Epistemology of the
Control System Model of Behavior *223*

The Fallacy of Behaviorism *233*

Human Nature from the Standpoint of Control Theory *241*

Standing at the Crossroads *259*

The X-Phenomenon *269*

Published Works by William T. Powers
on Living Control Systems, 1989-1991 *275*

Foreword

In 1979, Bill Powers wrote a prophecy: "A scientific revolution is just around the corner, and anyone with a personal computer can participate in it.... [T]he particular subject matter is human nature and in a broader scope, the nature of all living systems. Some ancient and thoroughly accepted principles are going to be overturned, and the whole direction of scientific investigation of life processes will change." (William T. Powers, "The Nature of Robots: Part 1: Defining Behavior," *BYTE* 4(6), June, p. 132) Powers foresaw the overthrow of the idea that either stimuli from the environment, or commands from the mind or brain, are sole causes of behavior. In its place, he offered the concept that people (and in their own ways all other organisms) intend that they will experience certain perceptions and behave to cause the perceptions they intend. The social, behavioral, and life sciences had simply missed the fact that living things control many features of their environments. Powers acknowledged that fact, and he realized that to an organism the environment exists only as perceptions, hence his insight that organisms act to control their own perceptions. His formal statement of the new concept was control theory, and he said amateurs, working with personal computers on their tables at home, would be major players in the revolution. Thirteen years later, the revolution is not accomplished, but it is underway.

Powers' perceptual control theory is new, but he is not the first to describe many of the key ideas in the theory. Over 2200 years ago, Aristotle wrote about intention—"that for the sake of which," the desire or wish that causes actions that result

in a particular end. Aristotle used many examples in which a person acts to produce an intended object, such as a bed, statue, tray, or house. The person's intention to create the object is the "final cause" of the actions that produce the object. Aristotle wrote that, depending on the condition of the world and the intention of the person, the same actions sometimes produce different ends, and different actions sometimes produce the same end. All of that sounds like good control theory, so why are those ideas considered revolutionary today?

For many centuries, Aristotle's ideas disappeared from Europe and were preserved by scholars in the Arab world. They returned, in altered form, to a Europe dominated by Christian theology. Theologians changed "final cause," which to Aristotle often meant only a person's intention to manufacture a bed out of wood, into God's original plan for the linear unfolding of history, from creation, to Calvary, to Apocalypse, to the end of time. Aristotle's original idea was unrecognizable.

Most early European scientists worked within Christian theology, embracing its notion of linear time and its implication of linear cause and effect. Many of these scientists mistakenly assumed that the original concept, that a final cause is a goal, implied that the future influences the present—a clear violation of the assumed linear flow of cause and effect. Eventually, potentates of The Church and potentates of Science came to a falling out over dogma. Those who established the canon for Science had yet another mistaken reason to reject final cause: they said it represented an appeal to the supernatural, in the form of God as agent. The idea that there is purpose or intention in the behavior of any living thing was pronounced "unscientific." Most aspiring behavioral and biological scientists still affirm that credo.

When William James wrote one hundred years ago, the ideas of purpose and intention were popular again. James said purposive behavior is the distinguishing feature of intelligence —of life. He said that in a variable world an organism's behavior necessarily varies to produce unvarying intended results. James wrote that people do not intend their specific actions; they intend to experience perceived consequences of their

actions, then they vary their actions any way necessary to produce those perceptions. For a while, it looked as though the idea of intention might take hold, but once more the idea was purged from the sciences of behavior and life. Orthodox scientists asserted that intention implies final cause, which necessarily implies an appeal to supernatural forces and to a temporal reversal of causality. Purposive behavior was banished, on the one hand by behaviorists, environmentalists, and reflexologists who claimed that events in the environment determine behavior, and on the other by those who claimed that instincts acting as internal stimuli cause behavior. People on either extreme believed their positions were dramatically different, but they all portrayed behavior as the end result of a linear chain of cause and effect.

Powers writes at a time when purpose and intention remain unacceptable to most scientists. Behaviorists still believe environmental "stimuli" have the "power" to control behavior; and most cognitive scientists and neuroscientists say the mind-brain issues "commands" that cause muscles to produce appropriate behavior. Cognitive-neuroscientists frequently claim behaviorism is dead and a cognitive revolution has swept the behavioral and life sciences; in return, behaviorists pronounce themselves very much alive, and some portray cognitive theorists as "creation scientists," bent on keeping alive the concept of soul-as-mind. Once again, each camp believes its views differ markedly from those of the other, but both embrace the wearisome model of linear cause and effect —a model that was necessary a few hundred years ago to establish the physical sciences, but a model that mistakenly rejects what Powers recognizes as the defining properties of life. Neither wing of the cause-effect orthodoxy recognizes the abundant evidence that organisms control many parts of their world. But revolutions have a way of changing the minds of the orthodox.

Powers turned the millennia-old idea that living systems act to produce intended perceptions into a formal theory of behavior: perceptual control theory. Perceptual control theory identifies behavior as the necessarily variable means by which organisms control their perceptions of the world. Working first on a build-it-yourself computer (the one he used when he

wrote his prophecy), then on a first-generation IBM personal computer, Powers created elegant demonstrations in which the simple-idea-turned-formal-model generates remarkably accurate quantitative simulations and predictions of behavior and its consequences. He identified a first principle for behavioral, social, and life sciences and showed the way to a new foundation of theory and method.

For several years, only a few people followed Powers' lead, and even fewer gathered the data and performed the modeling that could establish control theory as an alternative to traditional science. But interest in the theory grew—a tribute to the dogged efforts of William and Mary Powers, over three decades, to maintain the visibility of the theory. During most of that time, Powers published only one book and a few papers. More recently, information about control theory burst into wider circulation through two functions of personal computers that no one predicted in 1979: desktop publishing and electronic-mail networks. Those applications freed perceptual control theory from the heavy hands of editors and reviewers who routinely rejected manuscripts on the theory. They were true defenders of cause-effect orthodoxy, rejecting control theory as uninteresting and unnecessary, or as merely another way to describe things that were already understood. The new media let many people see control theory, then judge it on its own merits. The once-small circle of people aware of the theory grew into a network spanning the world, including people from many disciplines, specialties, and professions. And the demand for Powers' writings grows.

In the Foreword to the first volume of *Living Control Systems*, Richard Marken wrote about the difficulty he experienced several years ago when he tried to locate published material by Powers. Volume I was a collection of Powers' published work. But Powers has written far more than he has published. When he writes, Bill does not revise his drafts. If he encounters a block or is dissatisfied, he starts over. He has cast aside several beginnings of books and many drafts of chapters and papers that he never submitted, or that were rejected by editors and reviewers. Most of us would be happy if *any* of our publications equalled the quality of the work Bill put away in drawers and boxes and, more recently, on disks.

Over the years, only a few people have had a chance to read parts of Bill Powers' unpublished work. The opportunity to rummage about, looking for those gems, was at least part of "that for the sake of which" some of us travelled to his "laboratory" in the back room of his home in Northbrook, Illinois. When Mary and Bill decided to move to Colorado, Edward Ford, a counselor in Arizona, suggested that the mandatory gathering of possessions into boxes provided an excellent chance to select part of Bill's unpublished work for an edited volume. Greg Williams, a frequent visitor to Northbrook, journeyed there from Kentucky for the last time to gather the pages and disks and take them away so he could select the pieces in this volume.

This volume contains a small sample of the previously unpublished material from the years when Bill and Mary Powers were in Northbrook. If you want to rummage through the next accumulation, you must travel to the new site of The Laboratory of William T. Powers. That is the locus of many of today's clearest insights into purposive behavior. Over the millennia, that locus has moved from Aristotle's Lyceum, to James' Harvard, to Northbrook, and now to a house atop a ridge near Durango, with a view of the San Juan Mountains, located only a few miles away, across a broad valley—a view that, years ago in Illinois, Mary and Bill Powers said they intended to see out their back door. Stated intention, actions, and perceived consequences that match the intention. It looks like control to me!

W. Thomas Bourbon
Nacogdoches, Texas
February 1992

Preface

This is a volume of explorations, all unpublished and some, quite possibly, unpublishable in the ordinary course of things. A good many of them, such as the paper on evolution, are the result of following my nose, seeing where control theory will lead if given its head. Others have resulted from people asking me questions I couldn't answer, enticing me to think about subjects—like epistemology—that I wouldn't have thought interesting without some nudging.

Many of these papers were written as a form of thinking out loud, trying to put some order into disordered thoughts by arranging them as if for publication (with no intention of seeing them in print). In some cases, as for the paper on transfer functions (somewhat shortened because the data were taken four computers ago and are no longer at hand), the ideas were set down as a promise to myself; a promise to continue the work and get it into some finished form before offering it for publication. A few of those promises may yet be kept; most, I hope, will be kept by others.

Anyone with the will to do so is welcome to finish any of these unfinished thoughts if they look interesting. At this stage of my life, that would be a much more satisfactory outcome than doing it myself; the point of all this is to see others grasping and using control theory in the study of living systems.

If it hadn't been for Greg Williams, most of this material would have sat yellowing in my files or awaiting a cosmic ray to delete it from old computer disks. Greg went through mountains of my old papers and numerous disks simply because he thought it ought to be done before, in my irre-

sponsible way, I got rid of them in a fit of neatness. He has judged this collection to be worth preserving; I can't tell if this is true, because my mind is on what to write next, not those old words which have solidified into print. But I trust Greg Williams' judgment on scientific matters better than I trust that of almost anyone else, and certainly more than my own.

I must also acknowledge the material contributions to the conception and production of this volume by Edward Ford, at this writing the president of the Control Systems Group. Ed is a down-to-earth practical person, who deals daily with real human problems. For some reason he gives boundless encouragement to my theorizing, even though he then has to ask me what it means so he can translate it into something usable.

My thanks to Greg and Ed, good friends, good critics, and persistent benefactors.

William T. Powers
Durango, Colorado
February 1992

A Note on the Text

As in the previous volume of *Living Control Systems*, I have minimized alterations to the original manuscripts, attempting mainly to provide consistency of spelling and grammar within each paper. All deviations from the originals, for better or for worse, are solely my responsibility.

In coordinating this project, I have been aided by several individuals, including my wife Pat, who typed several of the papers and wrote the software used to redraw all of the figures; Dag Forssell, who provided—and typed—the elusive complete manuscript of "CT Psychology and Social Organizations" (which he had originally obtained from Rick Marken) and also provided the frontispiece holograph; Clark McPhail, who provided the manuscript of "Control Theory for Sociology"; and Mackenzie Yearsley, who provided the manuscript of "A Bucket of Beans." I thank all of them.

Ed Ford got this endeavor started, then made sure it was completed. Many, many thanks, Ed, for your ongoing support!

Most of all, I thank Bill and Mary Powers, who graciously permitted me to stay at their home, rummage through their files, photocopy whatever I wanted, and even cart precious manuscripts around the country. For the record: Bill and Mary are simply great folks!

Gregory Williams
Gravel Switch, Kentucky
March 1992

[1959]

Some Implications of Feedback Theory Concerning Behavior

In the course of developing a general feedback theory of human behavior, we have come across an interpretation of goal-directed behavior which seems to offer a departure from some concepts of stimulus and response currently employed by many experimentalists and theorists in psychology.

This interpretation is based on three facts about feedback control-systems: (1) Feedback systems are always goal-seeking; (2) feedback systems determine their own input-signals, and not their output-efforts; and (3) the output-efforts of feedback control-systems have the "purpose" of preventing environmental disturbances from affecting the input-signals.

Let us first show that these are valid statements about how all stable feedback control-systems operate, and then show how they might influence current interpretations of behavior if the systems involved should prove to have the properties of feedback control-systems.

Behavioral properties of control-systems

Diagram One is a very general representation of any feedback control-system, showing all the variables essential to such a system and all the functions performed on these variables within the system and its environment. Any feedback control-system can be represented in this form.

The oval labelled E represents the environment of the system. This environment consists of environmental variables in

Unpublished manuscript.

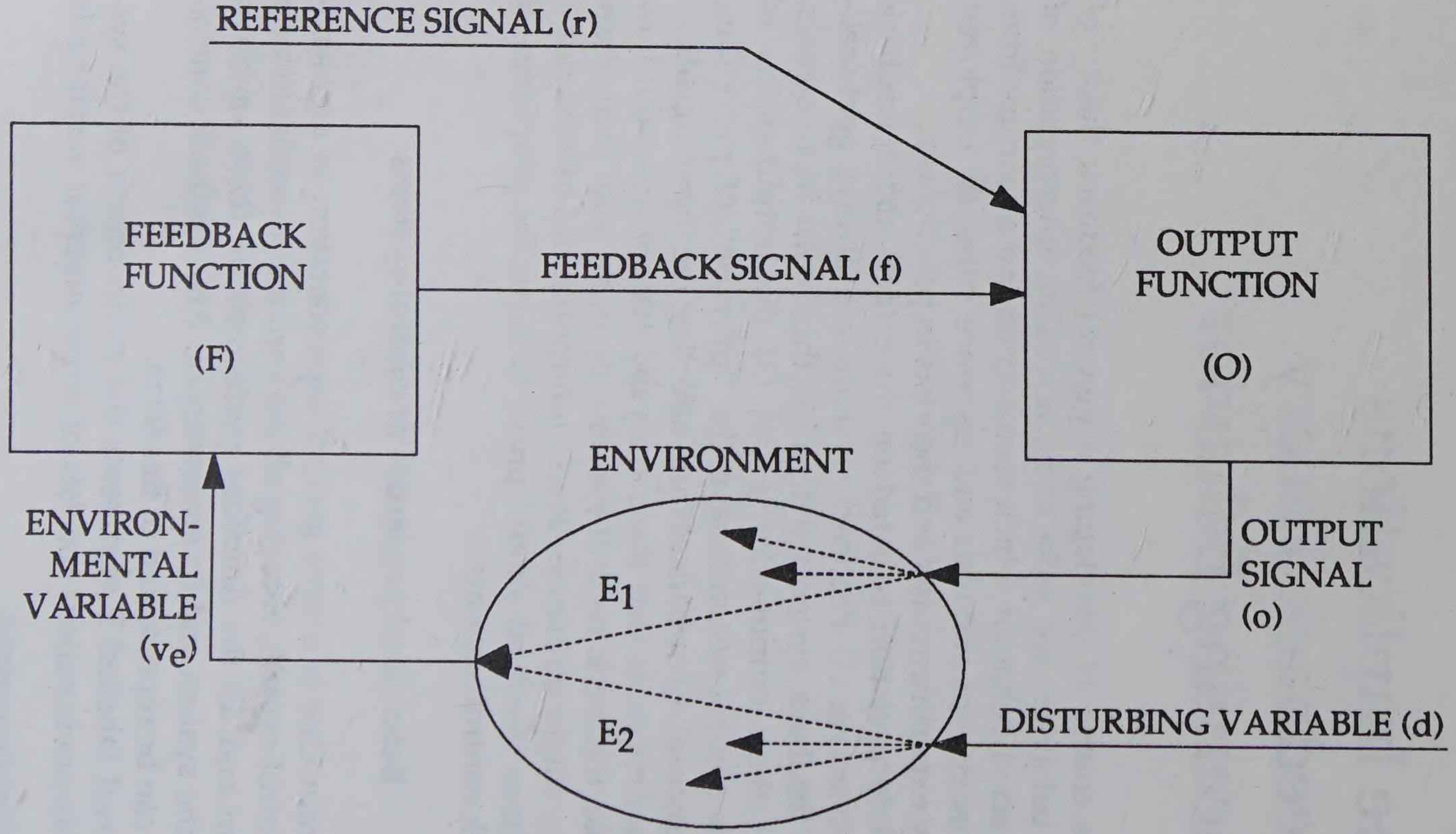


Diagram One: Functions and Signals in a Feedback Control System

relationship, some of which, labelled v_e, affect the feedback system's input apparatus or "sensory equipment."

The system has some kind of output (labelled o) into the environment; this output acts via the environment to produce among other effects, changes in v_e. All *other* environmental events which can affect v_e we sum up in the symbol d, standing for "disturbing influence." Thus v_e is functionally related to both d and o:

$$(1)\ v_e = E_1(o) + E_2(d)$$

The feedback system contains "transducers" which convert v_e (which need not be simple) into an internal signal f, the "feedback signal"; these "transducers" we lump under the label "Feedback Function," which may be complex or simple, and which we label F. Thus the feedback signal is functionally related to the environmental variables:

$$(2)\ f = F(v_e)$$

All feedback systems receive, at least implicitly, a signal which we call the "reference-signal" (labelled r). This signal (or a physical structure equivalent to such a signal) is compared with the feedback signal, and the remainder of the system acts *only on the basis of the difference* between reference and feedback signals. Thus, if the feedback signal exactly matches the reference-signal ("is at its reference-level"), the output of the system is zero, regardless of the implied magnitude of the feedback signal.

All that remains now is the function which converts a specific difference between r and f into an output o which will have the ultimate effect of reducing that difference. This final function we call the "output-function" and label O. A complex O can respond correctly in many environments; a simple one only in a restricted range of environments. We see that this output-function establishes a functional relationship between the "error" (r - f) and the output-signal o:

$$(3)\ o = O(r - f)$$

If we solve for the feedback signal f, substituting from (1) and (3) into (2), we have the general equation for a stable feedback control-system:

(4) $f = F\{E_1(d) + E_2[O(r - f)]\}$

If the system is stable and is subjected only to small disturbances, we can approximate all the functional relationships by constants. The restriction to small disturbances will be removed later. If we represent each function by a multiplying constant F*, E*, and O* (so that, for instance, $f = F^*v_e$), we can rewrite equation (4) (expanding at the same time):

(5) $f = F^*E^*_1d + F^*E^*_2O^*r - F^*E^*_2O^*f$ or, solving for f:

(6) $f = (F^*E^*_1d + F^*E^*_2O^*r)/(1 + F^*E^*_2O^*)$

For any control-system to have significant response to v_e and r, the product $F^*E^*_2O^*$ must be considerably larger than unity; under this condition, the "1" drops out of the denominator, and cancelling we find

(7) $f = (E^*_1d/E^*_2O^*) + r$

The condition stated previous to (7) is satisfied if O* is large enough; it then follows that (approximately)

(8) $f = r$

Equation (8) contains no mention of *any* of the functions: this does *not* mean that one could dispense with the physical functions and have the system work as well as ever! It only means that if the product $F^*E^*O^*$ is always much larger than unity, the specific values of the individual factors have no significant effect on the relationship f = r. A control-system satisfying this condition is called a "tight" system. Thus *if we know that a system is "tight" and stable,* non-linearities (even severe ones) in the functions have little effect on the condition f = r, as long as $F^*E^*O^*$ remains above some minimum value.

The validity of our first statement, that feedback systems are

goal-seeking, can now be demonstrated. The magnitude of the signal r (or its physical equivalent) represents a goal for the system. That is, it determines a state of the feedback signal f at which the system comes to equilibrium. Since r (if it is a real signal) is compared with f, it must have the same dimensions, and hence it must resemble a feedback signal—it represents a "goal-perception" which the system is "trying" to bring about in its actual perceptual field.

Our second statement, that feedback systems determine their inputs, not their outputs, follows also from the statement f = r. The system acts, as proved above, to maintain f at the value determined by r (remember that both d and o have dropped out of this expression) regardless of what the environment may do or what output is momentarily required (within the sophistication of the system, of course). If the system is overpowered momentarily, it will restore f to the value r as soon as it is free to do so.

The third statement, that the output of the feedback system has the "purpose" of preventing environmental disturbances from affecting f, is implied by the above statements; however, it can be substantiated directly by a different manipulation of equations (1), (2), and (3). If we combine them so as to leave the signal o explicit, we have (again with both the product F*E*O* and O* itself large):

(9) $\Delta o = -(E^*_2/E^*_1)d$ ("Δo" is the change in o resulting from introduction of d)

Remembering that $v_e = E^*_1 d + E^*_2 o$, we see that the effect of d on v_e is just cancelled by the effect of o on v_e. Since E^*_1 and E^*_2 take care of possible differences in dimensions, we can say that the system's output is equal and opposite to disturbances *but only with respect to* v_e. In general both d and o will cause changes in other environmental variables, but these other changes will not necessarily "cancel"—indeed, they might even add, but since they are not monitored by the control-system, they are not directly meaningful with respect to it.

The term "purpose," of course, is not meant metaphysically: the system has the "Purpose" of bringing f to the value of r because that is how it is built. However, this term is very

useful *with most of its usual connotations,* when one speaks of human control-systems controlling abstract perceptual variables, and acting with respect to correspondingly abstract goal-perceptions.

Control-systems and behavior-systems

The particular "goal" being sought by a feedback system is determined by the reference-signal, which usually can be altered by something outside the feedback system. In human neural feedback systems, this reference-signal might well be a neural signal (from some "higher" system) which, because of the point at which it enters the feedback system and *not* because of its "nature," sets the goal-perceptions of the feedback system. It is plain that the reference-signal "controls" the feedback system in a meaningful way, since it makes the system produce the required perception-signal f. It does not "control" the output, for that will have a value appropriate to the environment as well as to the reference-signal. Likewise it is clear that the externally-originating disturbance d does not "control" the system in any meaningful or regular way because the manner in which o varies is determined partly by r, and because d has little permanent effect on the critical variable f.

The disturbance d can *appear* to have direct control over the output o if the following conditions are all applicable:

(1) The reference-signal of the system under study does not change during the course of the experiment,

(2) The environment is so restricted that only one output-behavior o can possibly satisfy the requirements of the reference-signal, and

(3) All the variables studied bear an unchanging relationship with the pertinent v_e.

Practically every study of behavior based on the S-R viewpoint is conducted under these conditions. The "stimulus" corresponds to what we have called "d," but includes not only v_e but many other variables affected by d. The "response" corresponds to o, or rather (usually) to some set of environmental variables (not necessarily extra-organismic) linked directly with the outputs of the nervous system under

study.

The reference-signal "r" corresponds to different things in different experiments: in animal experiments concerning "drives," r corresponds to real or hypothetical signals determining the state in which the organism attempts to keep a set of basic neural signals which we call "hunger," "pain," "thirst," etc. The reference-levels for these signals are stable enough over a single experiment that one can usually rely on an animal to avoid pain and seek cessation of hunger, maximize pleasure-signals (or bring them to match high-intensity reference-signals), and so forth.

In many experiments with human beings, however, the signal r is determined *by the instructions*, and more important, by the subject's way of interpreting these instructions *in terms of the specific goal-perceptions he is to seek*, whether they be concrete or abstract. Misunderstandings at this level, and subsequent attempts by a subject to seek some irrelevant perceptual field, plague all experiments with abstract human behavior. However, it is often possible to explain instructions sufficiently, and to make sure that they are interpreted in terms of the proper goal-perceptions, so that during the course of an experiment, the reference-signals governing the various control-systems employed by the subject remain relatively constant.

Whether the reference-"signal" is constant because it is built-in or because it is held constant by instructions, its constancy is a prerequisite to any S-R investigation.

Even so, one cannot expect to find a given behavioral output in response to a specific stimulus in an environment where many different outputs might be relevant: consequently, it is the custom to restrict the ways in which the animal or subject can manipulate his environment, so that the "proper" response becomes the *only probable* response. Under these conditions, one can only measure the rate at which the response is learned or unlearned, and any statements that the stimuli "cause" the response, implying that they have meaningful control over the response, are based on too limited an analysis of the relevant variables. Any "control" exerted by the stimuli on the response becomes trivial in comparison with the overwhelming control of response exerted by the experi-

mental conditions.

All these considerations are meaningful only if we can show reason to think that the behavior under study is involved in a feedback control-system: any behavioral system in which the components of Diagram One can be defined or inferred reasonably is very likely a feedback control-system.

If we wish to further test the feedback hypothesis, there is a good way of identifying feedback systems experimentally, which involves an approach to behavior foreign to the customary S-R approach. Relatively moderate stimuli, or "d," are used, chosen to have a *tendency* to affect many variables in the vicinity of the animal or subject. Then, *instead of concentrating on observing the resulting behavior*, we look for environmental variables (for simple systems) or patterns of them (if we are interested in systems controlling abstract variables) *which*, (because of the behavioral resistance) *either do not change or are quickly returned to their initial values*. To discover such variables or patterns of variables requires, of course, guesswork and insight, but a proposed variable or pattern must satisfy this condition if it is to be considered significant. Only variables or patterns which pass this "test of the significant variable" are directly relevant to the behavior of any feedback control-system.

One can define many patterns of variables perceivable in an organism's behavior and imputable to its perceptual abilities. Since the environment is highly patterned by natural laws, chances are that any definition of "stimulus" and "response" will enable one to find that R is at least a *statistical* function of S, and to formulate statistical "laws" of behavior. But only in terms of variables which pass the "test of the significant variable" can one find laws of behavior containing the least possible statistical variability. If one is studying feedback systems, it does not matter whether a variable is significant to the Observer: it matters crucially whether it is significant to the Subject.

Summary

A very general analysis of the behavior of feedback control-systems is used to show that such systems act to maintain a feedback (perceptual) signal at the same value as a reference

(goal-perception) signal, and that they accomplish this by producing an output (behavior) which cancels environmental disturbances tending to interfere with this action, as well as carrying out the "error-correcting" action. It is then shown that if organic behavioral systems should prove to be feedback control-systems, one cannot consider "stimuli" (disturbances) to "control" "responses" in any regular or meaningful way. It is shown that S may *appear* to control R in a feedback system if the reference-signal (goal) is held constant for the system, and if the environment is such as to allow only one response to succeed in making the perceptual field match the goal-perceptions. All S-R experiments are carried out under these conditions, and so (under the feedback hypothesis) any "control" of stimulus over response is an artifact of the experimental situation. An experimental criterion is outlined by which one can test for existence of a feedback control-system in behavior-systems by the "test of the significant variable."

This paper is a preface to the general feedback theory of human behavior being developed by myself and four associates. Preparation of an introductory paper is now nearly complete.

Acknowledgement: R. K. Clark and R. L. McFarland, both of the Veterans Administration Research Hospital, Chicago, Illinois, have shared the labor with me in the development of both the viewpoints expressed here and the general theory now in preparation for publication. They have contributed essential parts of the whole theoretical structure to which this paper is a preamble.

Bibliography

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

2. Bertalanffy, L. V. (Ed.) *General systems yearbook*. Ann Arbor: Braun-Brumfield, Inc., 1956.

3. Frank, L. K. et al. Teleological mechanisms. *Ann. N.Y. Acad. Sci.*, 1948, *50*, 187-278.

4. Korn, G. A. & Korn, T. A. *Electronic analogue computers*. New York: McGraw-Hill, 1952.

5. McKay, D. M. Toward an information-flow model of behavior. *Brit. J. Psychol.*, 1956, *47*, 30-43.

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

7. Spence, K. W. *Behavior and conditioning*. New Haven: Yale University Press, 1956.

8. Wiener, N. *Cybernetics*. New York: Wiley, 1948.

[1971]

The Wiener Feedback Model—A Strategic Error

The way in which feedback was presented in *Cybernetics* has kept psychologists from seeing how to apply feedback concepts to behavior. In effect, the concept of feedback was presented in such a way that it could be assimilated into the stimulus-response concept of behavioral organization: the idea of successive-state analysis, basic to Wiener's "stationary time series," may be responsible for erroneous concepts of how feedback operates, and his basic feedback diagram, often reproduced, is inappropriate to most feedback situations. As a result, the fact that there are feedback loops from output to input via the environment has been ignored, or else dealt with in inappropriate terms.

Successive-state analysis

In the 1964 edition of a basic psychology text, D. O. Hebb explains feedback as a succession of stimuli and responses, each response modifying the next stimulus for the next response. (Hebb, 1964)

This analysis uses the concept of *successive states* of a system-environment relationship. There are difficulties with this approach even in neural-net models, since the nervous system must be assumed to be "clocked," which it is not; in dealing with overall behavior the difficulties are worse because neither stimuli nor responses are discrete point-events, and they do not neatly alternate in time.

Unpublished manuscript.

A successive-state model, as it is used in cybernetics, represents a system as a collection of variables which assume specific values that depend on the previous values of the variables; the time-scale, however, is discrete rather than continuous, and the value of a given variable is computed by applying a transformation rule to the values which all the relevant variables had one time-quantum ago. It may seem that this type of representation, given small enough time-quanta, will converge to the equivalent continuous representation, but that is more likely not to be true than to be true.

Although time does enter into a successive-state analysis, it enters in a peculiar way that is not equivalent to the way in which it enters into a differential equation. Because of this fact, a successive-state analysis can make situations involving feedback appear to involve instability, when in fact there is none in the real system being modelled. This can be shown very easily in a two-variable model; after showing how the apparent instability can be removed by more realistic treatment of time, I will repeat the analysis for a behavioral thought-experiment in which the approach advocated by Hebb (and many others) is shown to have unacceptable implications.

Consider a two-variable system; the variables might represent the input and output of a behaving organism. The input variable X affects the output variable Y, and due to connections outside the system proper, the output Y has an effect on the input X.

The transformation rules are as follows: to compute the value of Y_t (in time-state t), multiply X_t by k_1; to compute the value of X_{t+1} (in time-state t + 1) multiply Y_t by k_2.

Let $k_1 = -2$ and $k_2 = 1$: the net sign of the factors in the closed loop of transfromations is negative, so there is negative feedback. Applying the transformation rules, however, we find that this system is unconditionally unstable for any initial value of X not equal to zero:

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4	
+1	-2	-2	+4	+4	-8	-8	+16	.....

If the intention was to model a system-environment relationship that is observed to be stable, then this analysis has clearly

failed, for it predicts that both variables will show ever-increasing oscillations about zero. Yet each relationship can be verified, the input-output relationship by application of arbitrary inputs and the output-input relationship by adequate physical knowledge of the environment.

The reason for the failure of this analysis is that the "time" separating two successive states of the variables is not really time at all: it is merely an ordinal number indicating the sequence in which computations are performed. If one performs one computation of a new value for X and Y every second, the system will oscillate at one-half cycle per second. If one programs the computation and runs it on a computer at 100,000 computations per second, the oscillations will occur at 50 kilohertz. The time-scale is completely arbitrary, and so the intervals between states are not time-intervals at all. Plotting the states of the system at "t_1, t_2, ..." may produce a chart that *looks* like a time-plot and is *labelled* like a time-plot, but which in fact is only a list, not a time-function.

In order to use the successive-state approach properly, it is necessary to determine the way in which the transformation rules are affected by the chosen interval between time-states. In all situations involving real physical variables, at least on the macroscopic scale of classical physics, there will be some variables that cannot change value instantaneously: one may determine experimentally that when the variable X is given a value, the value of the variable Y will indeed become -2 times X, but a close inspection will almost surely show that Y does not reach this value at the same instant that X begins to have its effect on Y: rather, the value of Y begins to change toward its computed value, and gets there by a smooth transition requiring a length of time that is fixed not by the speed with which we do computations but by nature.

This means that when we select a time-interval for a successive-state analysis, we are in effect determining how far toward its computed ultimate value the dependent variable will actually move during one computing cycle. As we shorten the interval between computations, the amount of change of the variable that takes place during one computing cycle becomes smaller and smaller, and in the limit goes to zero. Once that limit is reached we have a differential equation, not

a successive-state model.

The stimulus-response approach to feedback as exemplified by Hebb's analysis is incorrect because it is a successive-state model applied in a situation where such a model *must* make a wrong prediction of oscillatory instability. To be sure the nervous system involved in behavior seems to be made of components that show an all-or-none response and which respond in succession, but when there is feedback from the output of an organism to its inputs other kinds of events are involved, and these events involve continuous variables. Even the very first event in the output-to-input connection, the muscle response, is best described in terms of smoothly-varying forces applied to the bones, not discrete contractions of the individual muscle fibers. The net output force generated by a muscle cannot change instantaneously; even less so can the limb actuated by that muscle alter its position instantaneously or change its ongoing relationship to the rest of the environment instantaneously.

However one treats the events inside the nervous system, therefore, the overt behavior that results involves physical components that introduce a dynamic constraint on any successive-state model. Following a discrete stimulus, there may in fact be a brief lag before the muscle receives a signal at the start of a response, but whatever the neural message is telling the muscle to do, the muscle cannot assume the implied state of tension instantly. Rather, it begins to contract and both directly and indirectly to alter the stimulation of sensory nerves. Thus the successive-state computations must be modified to take the slow muscle response into account.

The effect of this dynamic constraint is to change the successive-state model to a continous variable model. This can be seen in the X-Y system we have already used, if a suitable dynamic constraint is added to the output variable, Y.

Previously, the variable Y was computed by multiplying the value of X from the preceding time-state by -2. We shall add a constraint by specifying that in the interval between two computations, the variable Y can alter by only 1/4 of the difference between its current state and its next computed state. For a model applicable to the nervous system, we would naturally have chosen a computing interval equal to the time required

first for a change in muscle tension to alter the stimulation of sensory nerves and then for that stimulation to change the message reaching the same muscle.

Now the computation proceeds quite differently:

1. Initial value of X = 1, Y = 0.

2. Compute new value of Y: $Y_{t+1} = 2X_t$, so Y_{t+1} (computed) = -2.

3. Allow Y to change by 1/4 of the difference between present value and computed value: Y_{t+1} (actual) = -1/2.

4. Compute $X_{t+2} = Y_{t+1} = -1/2$.

5. New computed value of Y = +1; 1/4 of change (1-1/2 units) = +3/8; actual $Y_{t+2} = -1/2 + 3/8 = -1/8$.

6. Next values of Y: -1/32, -1/128, ... 0 (limit).

Now the model shows stable behavior—after an initial stimulus, the response and the effects of the response on the input both die rapidly away to zero. The successive values of Y lie on an exponentially decaying curve, just as if the system had been analyzed with a continuous first-order differential equation, or a critically-damped second-order equation. The initial successive-state model of the feedback situation failed because it did not treat time properly; introducing time in the form of a dynamic constraint resulted in a model without the instability that made the other model incorrect.

In precisely the same way, the stimulus-response analysis of feedback is incorrect because it fails to account for the fact that response variables cannot change instantaneously. Consider a thought experiment similar to many actual experiments in the measurement of human control actions:

A man grasps a handle; by a twisting effort he can cause a spot of light to track a moving target. If the target is made to jump randomly back and forth, a system property can be measured relating distance between spot and target (error) to initial twist force. Suppose a roughly linear relationship is found, such that

$$\text{(twist) ft-lb} = -10x \text{ (error), inches}$$

Knowledge of the apparatus tells us that, conveniently,

$$\text{Spot movement, inches} = \text{twist, ft-lb}$$

If we now say that the error is a stimulus that causes a twisting-effort response, and apply the successive-state analysis as Hebb does, we compute successive values of twisting force after the target jumps to the right 0.1 inch creating an error of -0.1 (spot left of target). The initial force is zero, so we obtain, in foot-pounds,

Twisting force = 0, +1.0, -8, +81, -810.1 ...!

Obviously this model is qualitatively wrong.

To introduce a dynamic constraint, let the twist-generating muscles be such that they can change their output only 1/20 of the way from the current state to the next computed state, in the time required for the twist to affect the spot and hence the sensory feedback motivating the twist. In that case, we obtain the following series of force outputs:

0, 0.05, 0.0725, 0.0826, 0.0871, ...

Now the model is qualitatively correct.

When dynamic constraints are introduced, the successive-state analysis approaches a continuous-variable analysis. The hand-eye coordination task just analyzed can be cast in the form of a system equation relating spot position to force output, and an environment equation relating force output to spot positions.

Let t = target displacement
s = spot displacement
f = twisting force output
k_1 = system sensitivity to error
k_2 = spot response to twisting force

then:

(1) $f = k_1(t - s)$
(2) $s = k_2 f$

This leads to an equation for the output force:

$$f = \frac{k_1 t}{1 + k_1 k_2}$$

Letting k_1 = +1 ft-lb per inch of error, and k_2 = 1 inch per ft-lb

of twist as before, we compute that f = (0.1)(10)/(1 + 10) = 0.0909 ft-lb. This is essentially the limit toward which the previous solution was converging.

This analysis has no apparent dynamic considerations in it at all. They are there, however, in the form of the tacit assumption that the equations hold true *simultaneously*. This amounts to the assumption that time-lags in the system are short in comparison with the time required for some variable (say, f) to change appreciably. Another way to express the same assumption is to assert that transient terms in the differential equations that actually describe this system die away to zero before the state of the system is observed. In short, the system is *stable*, an assumption that is clearly valid in a model for behavior that is *observed* to be stable (non-oscillatory).

Thus the successive-state analysis of a real feedback situation can only be made to work by introducing dynamic constraints, and the constraint needed has the effect of making time-variables unnecessary for computing the steady-state condition of the system. Not all feedback situations can be attacked in this way, but when the behavior being modelled is known by observation to be non-oscillatory, no other approach can lead as directly to a corrrect quantitative prediction as a simple algebraic computation.

A feedback model for behavior

This brings us to the second aspect of Wiener's *Cybernetics* that has had lingering misleading effects, Figure 5 on page 132 that shows how to apply continuous-variable feedback analysis to behavior:

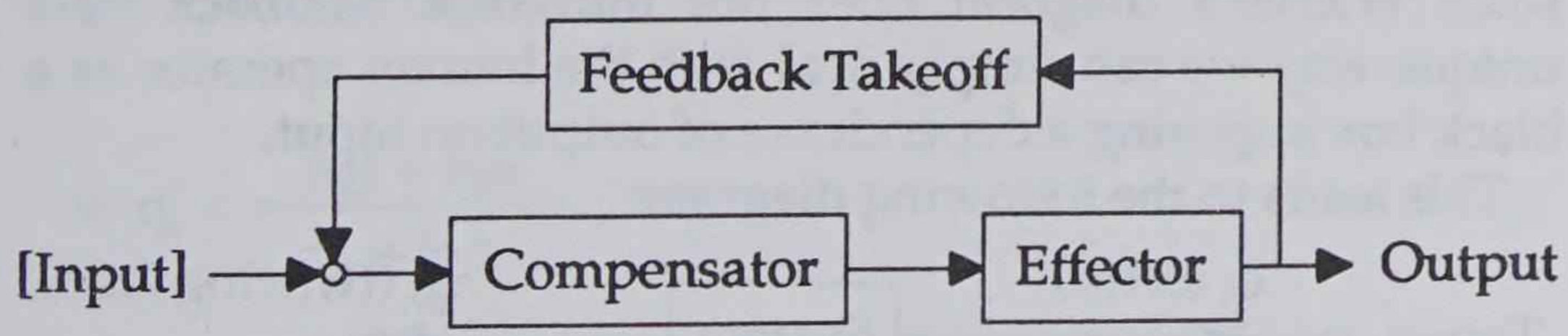


Figure 1

This diagram cannot be applied to the tracking experiment already described. In that experiment there was certainly feedback from output to input, but it did not occur *inside* the man; it existed in the environment. If we try to apply Wiener's diagram to this situation in the way that is commonly done (and that has been used for the last 24 years), we get an analysis that is ambiguous and incomplete.

The *input* by common usage would be taken to be the error between spot and target, since that is the "stimulus" that correlates with the man's motor "responses." Going forward through the diagram we arrive at the "effector," which corresponds to the man's arm muscle: the "output" would be the torque applied to the handle.

In Wiener's diagram, the "feedback takeoff" samples a portion of the output and feeds it back to subtract from the input so that the output is actually caused by the discrepancy between input and feedback information. Whatever the transformation T involved in the feedback pickoff unit, this connection will tend to force output to be the inverse T function (T^{-1}) of the input (exactly so if the effector amplification is infinite and free of lags).

Exactly the same overall input-output relationship could be produced by an infinity of combinations of feedback takeoff functions, compensators, and effectors, or by a straight-through active filter of suitable form, as Wiener himself pointed out. This means that as far as our thought-experiment is concerned, Wiener's diagram adds nothing—it only offers a selection of ways in which the observed input-output function might be accomplished.

It is still necessary to analyze the overall feedback loop, and since Wiener's diagram does not introduce feedback in a unique way we can simply deal with the human operator as a black box imposing a dependence of output on input.

This leads to the following diagram:

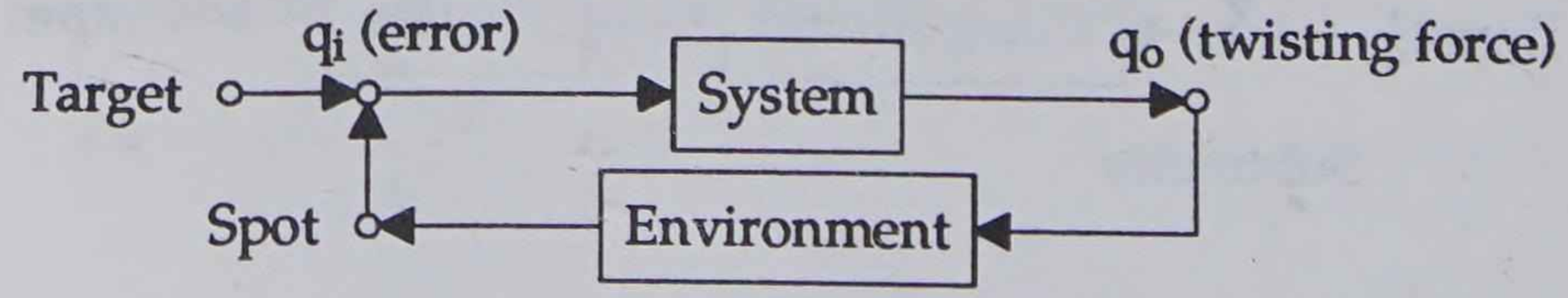


Figure 2

The input quantity q_i is the error between spot and target; the output quantity q_o is the twisting force applied to the handle. The "environment" box converts the twisting force to a given spot position, and hence a given amount of error. Whether there is feedback inside the "system" box makes no difference, and cannot be discovered by a behavioral experiment.

There is an implicit assumption in the way we have applied the diagram above, that zero input leads to zero output. Even if transient terms are ignored (meaning that we have reason to believe that they quickly disappear after any change) the steady-state part of the system equation might still be non-linear. In general it could be represented as a polynomial,

$$q_o = -a + bq_i + cq_i^2 + dq_i^3 + \ldots$$

(the negative sign makes later discussion easier).

If we ignore all terms of second and higher order, we have the system equation as:

$$q_o = -a + bq_i$$

The environment equation, as before, can be given as

$$q_i = k_2(t_{-1} - q_o)$$

where we let q_o be measured in terms of equivalent spot position.

Solving for the input and output quantity, we have

$$q_i = k_2t + k_2a - k_2bq_i$$

$$\text{or } q_i = \frac{k_2t + k_2a}{(1 + k_2b)}$$

$$\text{and } q_o = a + b[k_2(t - q_o)]$$

$$\text{or } q_o = \frac{a + bk_2t}{1 + bk_2}$$

These equations lead to the previous result, *but only if* a *is set to zero*. If a is not zero, meaning that the system produces some output when the input is zero, a different picture emerges.

The easiest way to see intuitively the effect of a nonzero constant term is to deal with a feedback situation in which the cumulative product of all multiplying constants around the closed loop is much greater than unity. This product is k_2b; if we say that $1 + k_2b = k_2b$, the two equations reduce to:

$$q_i = t/b + a$$

$$q_o = \frac{-a}{k_2b} + t$$

A further simplification comes from assuming that the system is very sensitive to small variations in its input—that b, in other words, is a large number.

This permits all terms divided by b or k_2b to be dropped; and we have

$$q_i = a$$

$$q_o = t$$

If a is zero, the equations tell us that the system will keep the spot on the target, and will do so by making its output (in terms of spot movement) match the target movements. In fact, because the system is not infinitely sensitive to spot deviations, q_i - a will not quite reach zero, and the spot will not quite exactly be on the target. This approximation, however, gives the correct idea of the behavior involved.

When the system constant a is non-zero, an interesting change occurs. The spot will no longer lie on the target, but a distance a away from it. However the target moves, the spot will track a point a units to one side of it. The effect would be

that of telling the subject, "keep the spot 3 inches to the left of the target." (The output behavior must now be inferred from the exact equations.)

Clearly the subject could be given other instructions: 1 inch to the left, 5 inches to the right, and so forth. The effect would be exactly like changing the value of a. Thus a is seen to have special significance, and the diagram should be modified to show it explicitly.

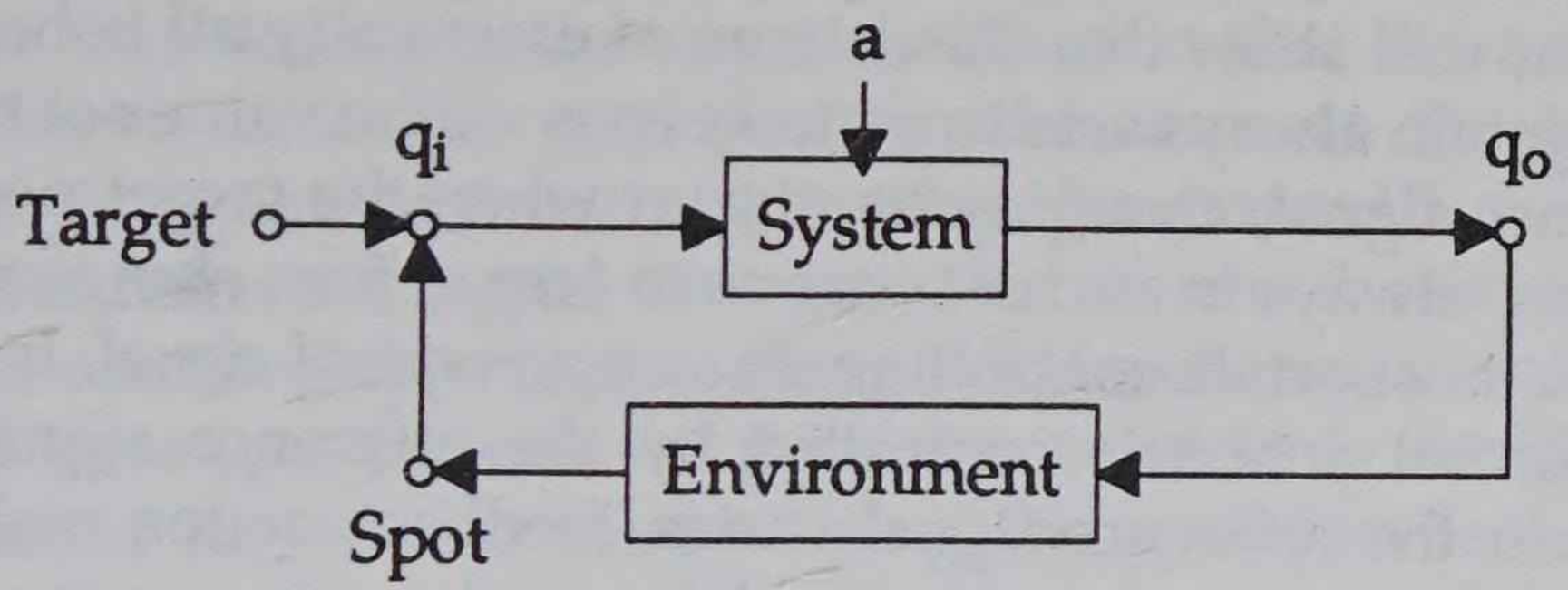


Figure 3

This diagram, while useful in behavioral analysis, is not philosophically satisfactory because the input quantity is arbitrarily defined. The "errors" between spot and target cannot excite sensory endings, but must be computed by the subject's nervous system from the actual input information, the visual images of spot and target. To eliminate this failing of the diagram above, the system must be taken apart somewhat, just enough to show how q_i has to be defined.

The crucial change is to indicate an input computing function which receives detailed sensory information and computes the value of q_i:

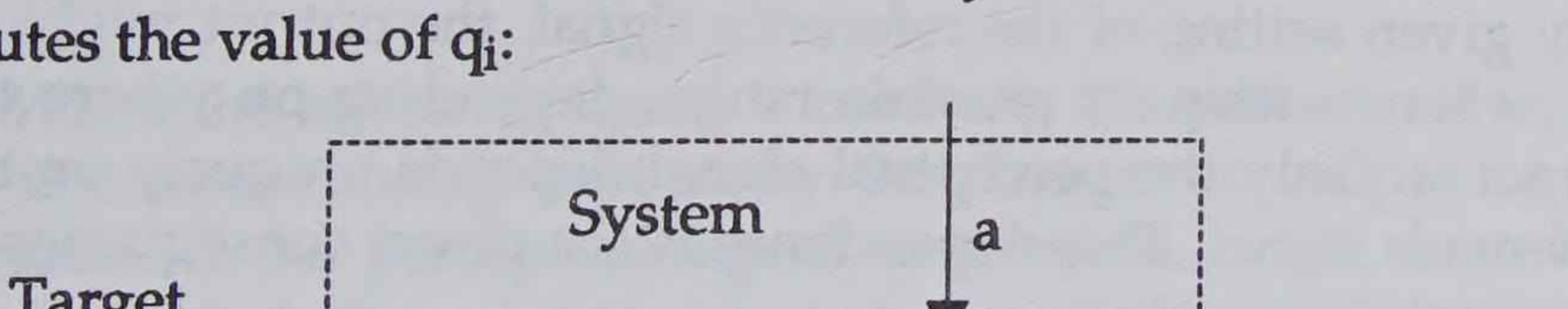


Figure 4

Assuming that q_i and a are expressed in units of q_o (thus taking care of all scaling factors), the system can now be seen to act on its environment in such a way as to make the *perceptual signal* q_i track the value of a; a acts as a *reference signal*, for the output response depends on changes in q_i *with reference to* a (which determines the zero-point: when q_i = a there is no output).

This analysis points to a radical change in the way we think of behavior when feedback plays a prominent part (serious thought will show that this is true of essentially all behavior). This system always acts so as to keep q_i at the value set by the reference signal, changing its output when the target position changes, always in such a sense as to *keep* q_i *from changing*. The system, in short, is controlling *its own perceptual signal*. In turn, this control action is controlled by the reference signal, for however the reference signal varies, feedback action makes q_i vary in exact correspondence. We are ignoring, remember, transient terms.

The reference signal controls the perceptual signal: this is counter to appearances. Tracing cause and effect would show that variations in the reference signal cause the output to change—it would seem that the reference signal controls the *output* of this system. Using the time-scale in which transient phenomena damp out essentially instantaneously, however, we see that if the reference signal varies *slowly*, the output is affected equally by the perceptual signal.

The output is therefore affected by changes in the target position just as much as by changes in the reference signal, and hence is not uniquely dependent on the reference signal. For any given setting of the reference signal, the output might be anywhere within its possible range, depending on where the target is. Only the perceptual signal depends uniquely on the reference signal. This dependence is the direct consequence of feedback.

If the reference signal remains constant, then the output is completely determined by the target position; this is easiest to see with a reference signal of zero, meaning spot on target. As long as the system continues to behave as a good feedback control system, the spot will be on the target ($q_i = 0$) and the system's output will always be such as to produce that result.

In a traditional behavioristic analysis, the target would be assigned the role of "stimulus" or "independent variable," and the subject's motor output (or equivalently, the spot position) would be called the "response." Indeed, the response *will* depend on the stimulus in this situation and the behavioristic interpretation is compelling. It seems as if the stimulus acts on the subject to make him produce the observed responses.

On the basis of purely empirical observation, one can deduce a stimulus-response law enabling one to predict the response from knowing the stimulus. The fact that the response happens to make the spot position the same as the target's cannot be explained (but it *can* be taken for granted!).

If the reference signal changes, the observed stimulus-response relationship will also change. If the change in reference signal were spontaneous—if the subject suddenly decided for reasons of his own to start keeping the spot two feet to the right of the target—this change in the observed relationship would be inexplicable. If the subject made a series of such arbitrary changes in reference, the stimulus-response law would begin to show variability, noisiness. In that case, the only recourse would be to repeat the experiment over and over, and extract an *average* stimulus-response law.

Such statistical treatment of behavior in a feedback situation, however, effectively discards all information about changes in the reference signal. What is left is a picture of behavior with some average and constant setting of the reference signal. The average relationship of response to stimulus might prove reliable, repeatable, and accurate, but it nevertheless leads to a false conclusion—that something has been learned *about the system*.

When the reference signal is constant, the perceptual signal q_i is constant, being kept so by the changes in output of the system. If the perceptual signal is constant, what does that imply about the relationship between "stimulus" and "response"?

Recall that q_i is defined by the system's input function, which derives q_i from detailed sensory inputs. Some of these detailed inputs are affected by the target position and some by the system's own output via the spot position. If q_i is known to be constant, or very nearly so, then the observed relationship

between the target position and the system's output "makes sense" only if the nature of the input function is known, because for every target position, there is one output that will leave q_i unchanged.

Furthermore, the target position affects q_i in the opposite direction to the effect of the changes in system output; if q_i remains constant, then in effect the sum of target and output effects on the input function is a constant. Therefore, the observed relationship is determined *completely* by properties of the environment. The primary changes that take place are those in target position and in system output which are connected to the system's sensory inputs by properties of the environment:

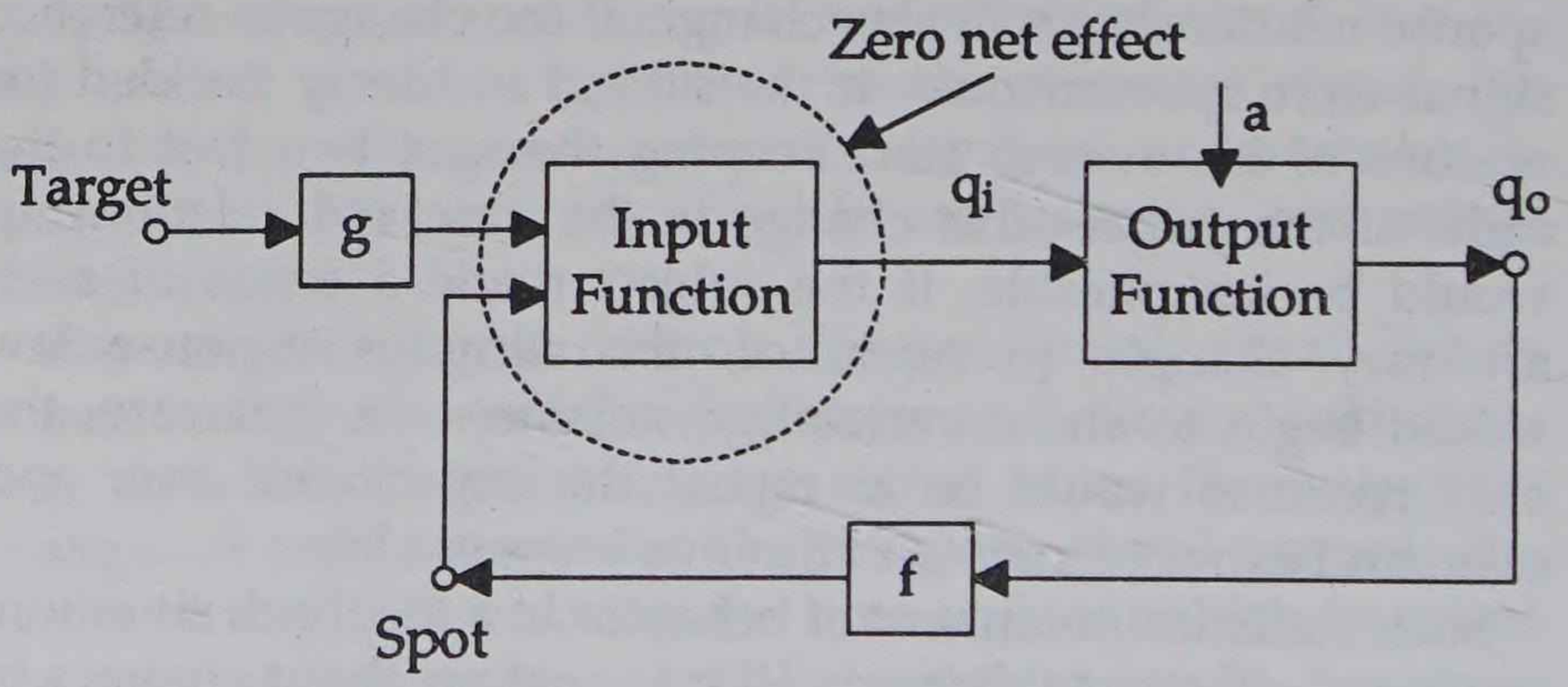


Figure 5

If we say that the *net effect* on the input is zero at all times (or very nearly zero) then the stimulus-response law can be written in terms of environmental functions f and g:

$$f(R) = -g(S), \text{ or}$$

$$R = f^{-1}[-g(S)]$$

Letting the sum of effects be some constant other than zero introduces no conceptual difference. The observed stimulus-response law is still shaped mainly by properties of the environment and not of the system. Therefore in all feedback situations such as this, and they are very common, the stimulus-response approach yields a result that can be very reliable, but

nonetheless trivial.

The example used here involved a tracking situation, but the applicability extends to many other situations. The input quantity q_i need not represent the distance from a spot to a target—it can represent anything at all that is affected by the system's output forces.

Precisely the same diagram would apply if q_i stood for, say, the contrast of a television picture, and q_o was the manual output that turns the contrast control. In that case, the entity we have called the "target" becomes *any source of disturbance of* q_i. This disturbance need not be sensed at all: if picture contrast drops *for any reason* it will be corrected.

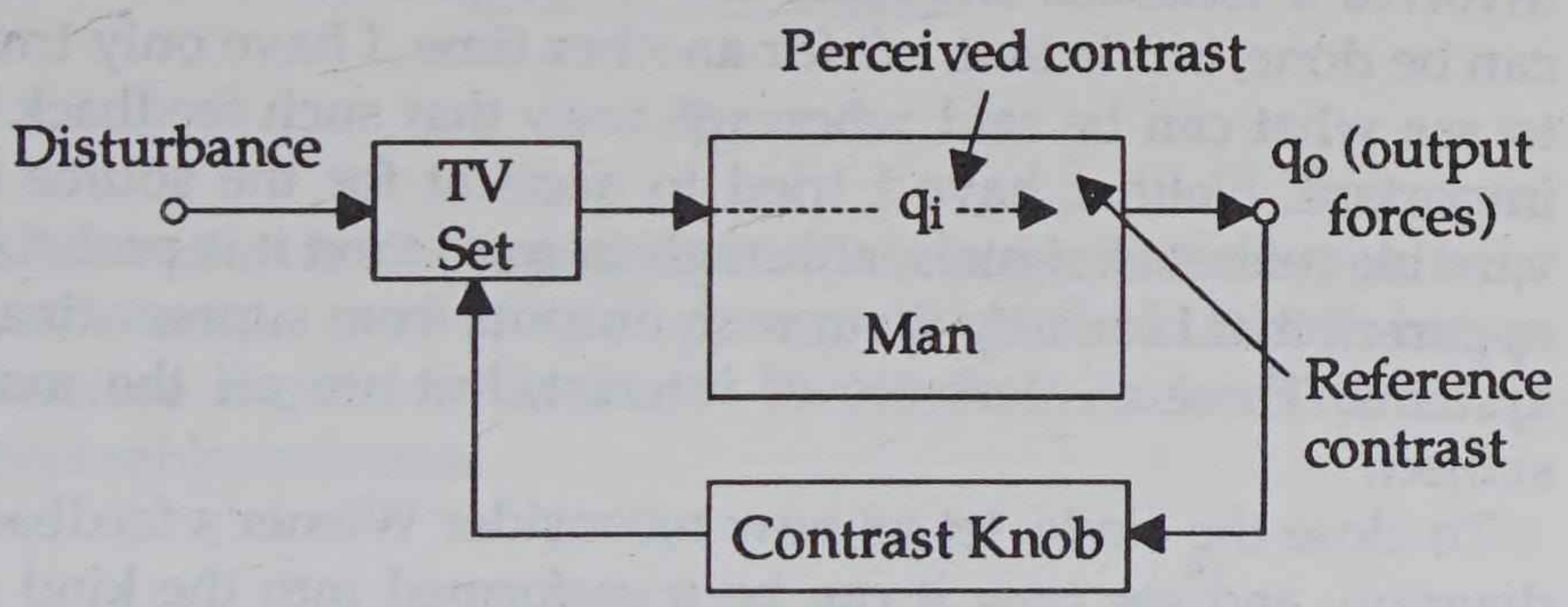


Figure 6

Traditional cause-effect concepts of behavior have led to ways of thinking that are solidly entrenched but which in feedback situations are nevertheless wrong. This can be seen by considering a person adjusting the contrast control, and asking "What is he doing?"

The normal answers would be "He is adjusting the contrast," or, if one considered himself to be rigorously behavioristic, "He is turning the control knob," or "His muscles are tensing in such a way as to grasp and turn the knob."

In fact, the closer one's answer came to a factual description of muscle behaviors, the more his answer would tend to be viewed, even begrudgingly, as "scientific." The present development shows just the contrary: the more closely one approaches a description of muscle behavior, the farther one gets from a proper description of what the subject is actually doing.

The only reliable consequence of the man's nervous-system activity is the *control of his perception of contrast.* His motor output will fluctuate with every random effect that can change the picture's contrast, and the relationship between these effects and his output is determined almost entirely by properties of the world that lies outside his nervous system. What this man "is doing" can be given a reliable definition only after one has discovered the nature of the input function that is involved and hence the nature of q_i—the way it depends on external events. No behavior of the man's output can reveal the essential information about *what he is doing.*

I have not attempted to demonstrate that behavior always involves a feedback situation like those described here—that can be done, but is best left for another time. I have only tried to see what can be said when we *know* that such feedback is important. Neither have I tried to account for the source of variable reference signals, although from context it is probably apparent that I identify them with outputs from supraordinate systems. These matters are of interest, but are off the main subject.

To close the circle, let us now reconsider Wiener's feedback diagram, and see how it can be transformed into the kind of diagram developed here. First let the comparator—the subtracter—be absorbed into the "compensator-effector" functions, thus:

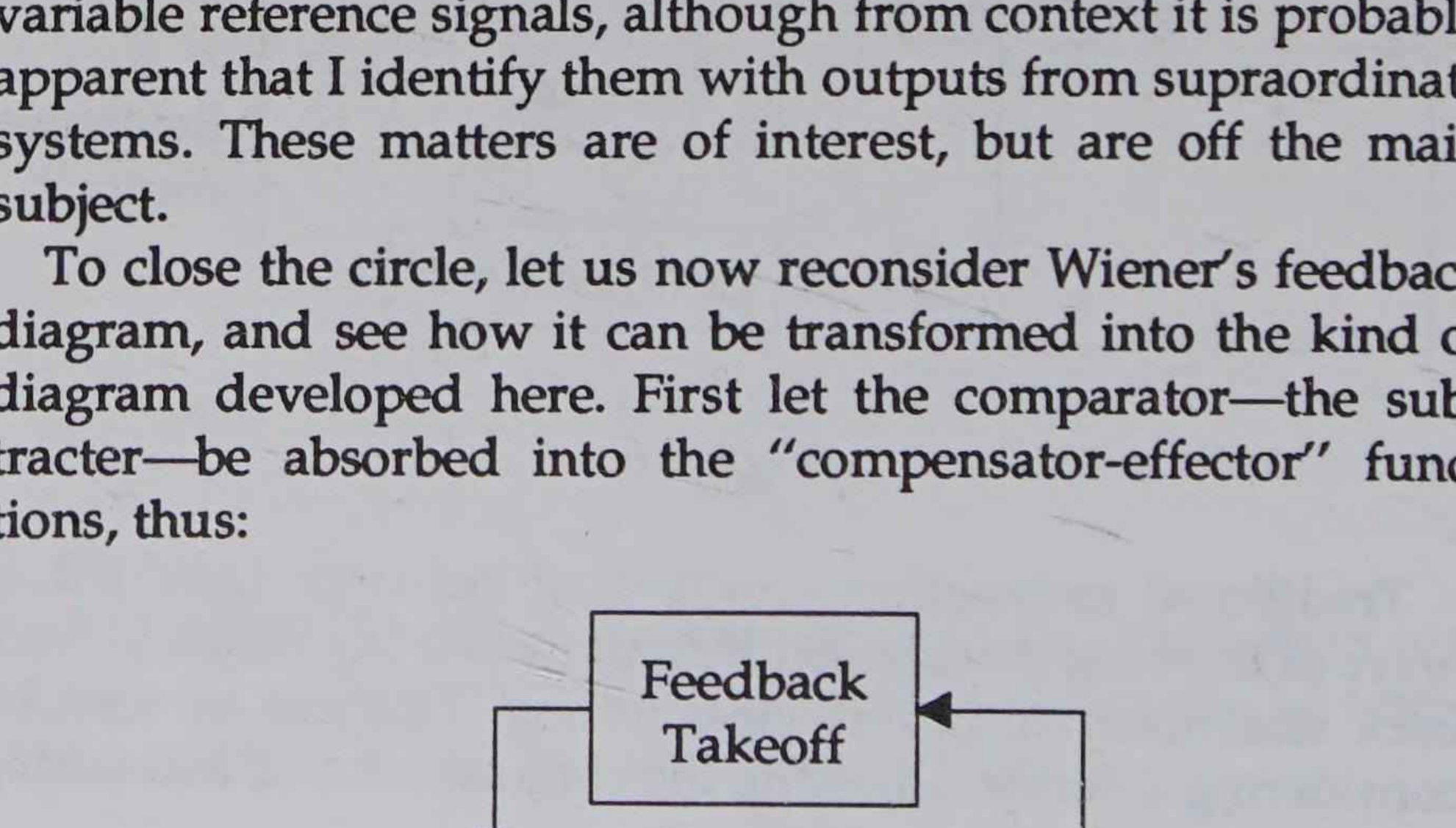


Figure 7

Next, let us find where the output of the system affects the environment at a place that lies outside the active and computing parts of the system; here:

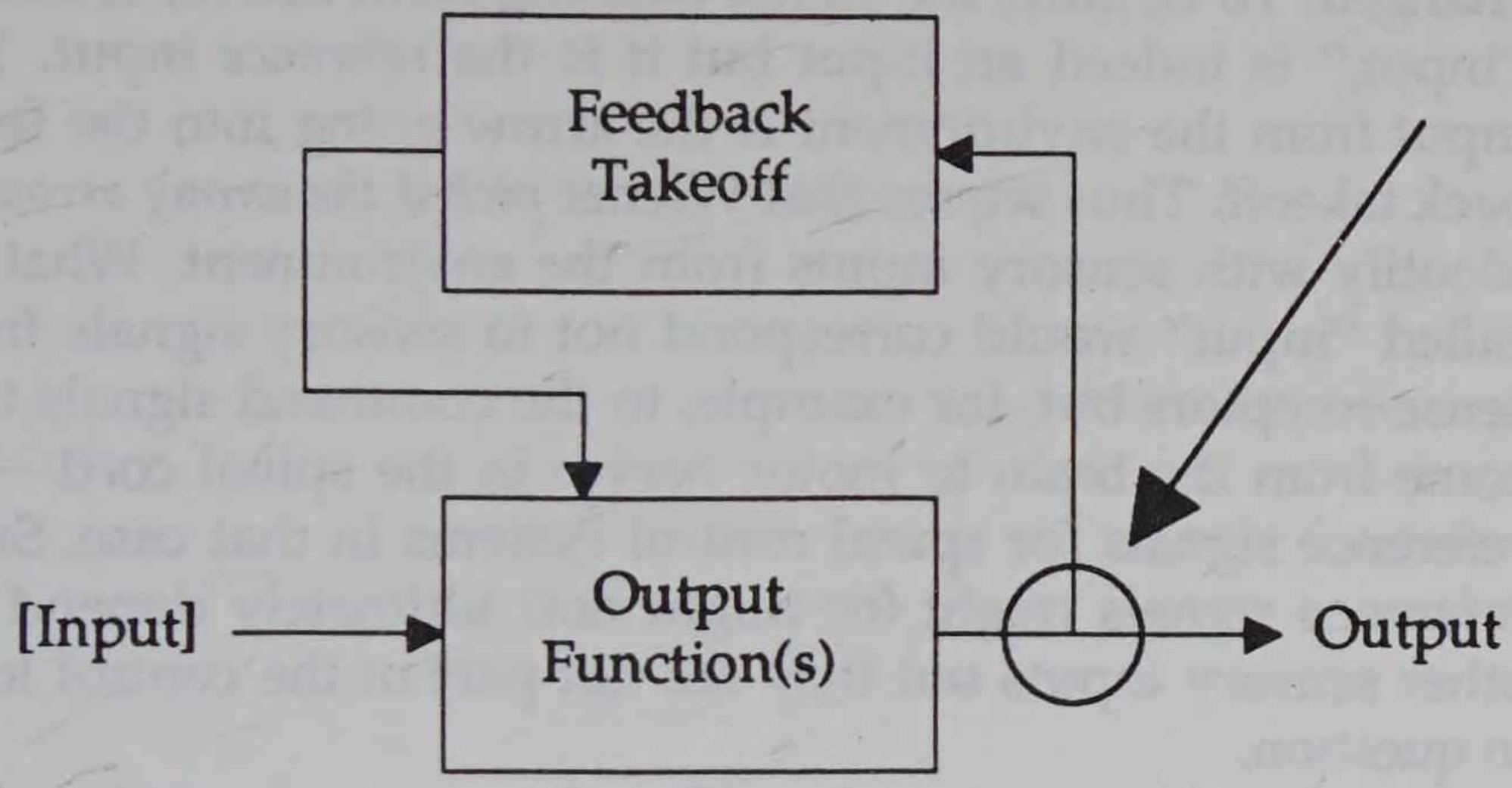


Figure 8

This location is where physical events occur that affect an *input* to the active part of the system, the *feedback takeoff*. Let us therefore expand what lies in the circle to include the entire observable universe.

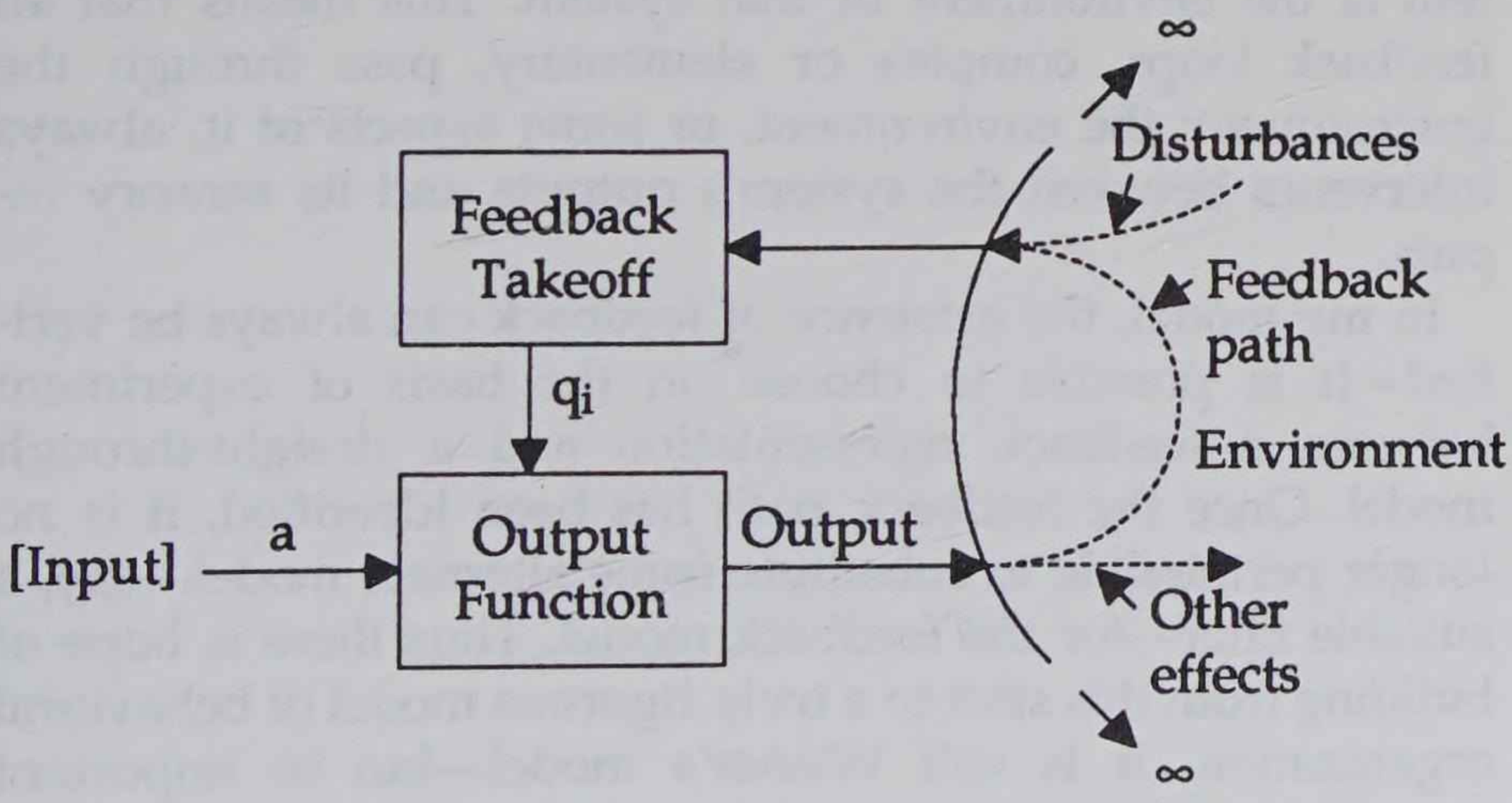


Figure 9

The length of the dotted path is immaterial to the organization. The final step is to rotate the diagram 90 degrees and turn it right-for-left; we see that it is now essentially the diagram we have been using (e.g., Fig. 4).

The identifications of the pathways, however, are now changed. To be sure, the signal entering from above, Wiener's "input," is indeed an input but it is the *reference* input. The input from the environment is the arrow going into the feedback takeoff. Thus we see that Wiener *picked the wrong arrow* to identify with sensory inputs from the environment. What he called "input" would correspond not to sensory signals from sense-receptors but, for example, to the command signals that come from the brain to motor nerves in the spinal cord—the reference signals for spinal control systems in that case. Such reference signals might (or might not) ultimately depend on other sensory inputs but they are not part of the control loop in question.

This error, if it ought to be called that, arose simply because of what Wiener considered to be the behaving system. He considered the muscles and the tissues connecting them to kinesthetic sensors, and those kinesthetic sensors themselves, to be *inside the behaving system*. What I have done is to draw the boundaries of the behaving system so that the behaving system is *entirely neural*—everything that is not nervous system is the environment of that system. This means that *all* feedback loops, complex or elementary, pass through the environment: the environment, or some aspects of it, always intervenes between the system's outputs and its sensory inputs.

In my model, the existence of feedback can always be verified—it is possible to choose on the basis of experiment between a feedback representation and a straight-through model. Once the feedback path has been identified, it is no longer permissible to substitute some alternate model—say, a suitable filter—for the feedback model. Thus there is hope of building from this start to a truly rigorous model of behavioral organization. It is still Wiener's model—but in important ways, it is no longer his model.

References

Ashby, W. R., *Design for a Brain*, New York, Wiley (1952).

Wiener, N., *Cybernetics*, MIT Press (1948).

Hebb, D. O., *A Textbook of Psychology*, Philadelphia and London, Saunders (1958; 1964 Edition).

[1972]

Emotion

There are three major categories of sensations. One category is associated with what we learn to identify as events that take place outside the body; sensations of light, of smell, of sound, of temperature can arise or change without our taking any action at all. In control system terms, these are the quantities that seem especially susceptible to disturbance by outside agencies. Then there is a set of sensations that seem to be under more or less direct voluntary control: sensations of effort, of joint angle, of facial expression, of touch, of movement—these sensations hardly ever appear unless one wishes them to appear, and external events have very little power to affect them in comparison with our own control over them. The third category is odd in that it seems to have its location entirely inside the body, and yet at times seems to be under control by the external world, or by factors in ourselves over which we have no voluntary control at all: these are *feelings* or *emotions*. This chapter represents an attempt to bring the third category of experience into the scope of this model.

Emotions are well-known to have two components, one thought-like and one sensation-like. Physiologists have long identified the sensation-like component as arising from changes in the biochemical state of the body—alterations in the levels of action of the major organ systems. In fact for all strong emotions, there appears to be a more of less common physiological state, termed the "general adaptation syndrome." The adrenal cortex releases adrenalin into the blood-

Chapter 26 of the preliminary version of *Behavior: The Control of Perception*, used for teaching a course at Northwestern University.

stream, causing a release of stored energy, constriction of the peripheral capillaries, an increase in heart-rate (and a rise in blood pressure), increase in respiration rate, and so on—the whole balance of the biochemical system seems to change in such a way as to prepare the body for immediate action.

The important fact here is that this change in biochemical state *can be felt*. When the balance changes, we perceive a change in our own internal state—it can be felt very clearly and strongly.

The general adaptation syndrome is not the only such change that can be felt. When one becomes ill, he begins to notice a difference in the sensations from his body; he may feel aches, or nausea, or simply an undefinable "discomfort" from various parts of his body. When a person overeats, he is aware of a sluggishness, an overstuffed feeling; when a fever appears, he feels chilly inside, or feels "odd" in his stomach, chest, limbs, head. The general adaptation syndrome is merely one configuration of altered sensations from the body; other types of malaise represent other configurations. I want to suggest here that these different feelings are not caused by signals arising from special receptors for each new sensation, but represent varying patterns perceived in the *same* set of signals that represent the state of the organism.

An odd fact about these sensations is that while they seem to involve a great many receptors inside the body, and can be arranged in a very large number of patterns, it seems that under ordinary circumstances *these signals are all zero*. If one is not sick, or excited, or afraid, or upset in any of the myriad possible ways, he feels *nothing*. I do not mean that he feels numb but that he refers to whatever sensation may be in existence as if nothing were happening. One knows his body is there, he can feel it when he does anything; he can detect certain events going on such as breathing, but he would not say that he is experiencing any *feeling state*. That is the normal state of affairs, that is the state one is in when nothing in particular is going on. Emotions and feeling states appear to be identified with deviations from this "nothing-going-on" state.

In the model that has been put together so far, we have a set of signals that seem to have exactly the properties that are seen in the properties of feeling-sensations: the *intrinsic error signals*.

The intrinsic signals themselves, which supposedly represent the output of special body-status sensors, were spoken of as continuously-variable signals which could at various times have values greater than or less than some set of intrinsic *reference signals*. It was noted that these hypothetical reference signals might not exist for some kinds of intrinsic signals; a receptor having a high threshold could generate an intrinsic error signal directly, the threshold-setting determining the reference level, and any amount of super-threshold stimulation constituting an error: pain signals, at least of certain kinds, would fit this category.

It is still not necessary to try to guess whether real reference signals exist to determine "intrinsic state"—the state in which all intrinsic error signals are zero. But clearly the way in which we sense our bodily status resembles the behavior of the intrinsic error signals, and not signals that represent intrinsic state directly. We sense an increase in activity of the adrenal cortex as a heightened state, a feeling of excitement, or fear, or anger. We sense a lowering of its activity as dullness, lack of ambition, tiredness, depression. I over-simplify, of course. But my point is that the ordinary, the average, perhaps the optimum condition of this variable is sensed as *no* sensation. That is what we would expect if somehow the output of adrenalin, or some effect of that output, was monitored by a sensory nerve, and the sensory signal was compared with a reference signal, so that there would be no error when the sensory signal matched the reference signal. A mismatch of either sign would produce an error signal, and would be sensed; remember that neural comparators work only for one sense of error, so two comparators, one working with inverted signals, would be required to detect both too much and too little of the sensor output level. Perhaps that is one reason why we give different names to "excitement" and "depression," even though they may represent errors in the same variable.

On this basis I am going to propose that the feeling-component of emotion is precisely the set of intrinsic error signals. This does not imply that all intrinsic error signals that may exist and influence learning are sensed; it implies only that those that *are* consciously sensed are to be identified with the sensation part of an emotional state.

This immediately ties feelings and emotions to learning. If this theory is correct, then learning—the kind of learning that involves a change in one's properties, not merely acquisition of new information—is motivated under circumstances where one is feeling something different about his intrinsic state. Of course I have not ruled out intrinsic errors that are not sensed, and it is possible that one's awareness may not be directed to such error signals even if they are present as inputs to the learned hierarchy, but putting the case positively, when one does detect a feeling-state, *reorganization is probably taking place.*

Strong emotion has been considered "disruptive" by psychologists; that is almost an acceptable definition of emotion, as opposed to gentler feelings. If the sensation-component of emotions is in fact produced by intrinsic error, then this is understandable, for reorganization can destroy as well as build. Whatever the strategies of reorganization that are built into the organizing function, they are not rational, for rationality consists of operating according to logic and plan, and those are activities of the learned hierarchy of systems, not the system that is responsible for that learning. "Reorganization" does not have to mean "progress."

The other aspect of feeling states, of emotions, is the "cognitive" side, as it is called. There is a mental state accompanying the physiological state, and it is this mental state that gives character to the sensations of emotion. It is known that the state of anger is very similar to the state of fear in its physiological aspects: in either case the general adaptation syndrome appears. In fact, the logic behind that term can be seen in another name for it: the "fight-or-flee" reaction. The changes in bodily state that occur are appropriate for *any* behavior that involves strenuous activity. We feel differently when we are angry from when we are afraid primarily because of what is going on upstairs, in the brain, and not because the actual sensations are so different.

I think this provides a very simple way to define an emotion. In terms of this control-system model, when one initiates an action, he selects a reference signal for some high-order system: "I'm attacking!" the reference signal, in effect, says. The lower-order systems then do what they have learned to do in order to make the appropriate perceptual function say, "I'm

attacking." (The Russian language uses this mode of expression to indicate immediate action: if one wants to indicate that he is leaving *right now*, he says, "I'm gone." Even in English, when one wants to indicate his intention to comply *immediately*, he says, "Done!")

In short, the prime initiator of an emotional state is the selection of a goal for action, as an immediately effective reference signal. How this might work, in terms of the whole model, we will get to in a moment, but given just this much we can put together a reasonable definition of the term "emotion."

An emotion is the combination of a goal and a feeling. If the feeling is akin to the general adaptation syndrome, then the goal of attacking produces anger, the goal of fleeing produces fear, the goal of rejecting a perception entirely becomes horror or hate, and so on. If the opposite sort of syndrome is in effect, that is a general depression of functions rather than a heightening, then the goal of escaping becomes despair or anxiety, the goal of attacking becomes resentment, the goal of rejecting a perception, apathy, and so on. It makes a very neat system, which I am sure can be worked out better than this.

There is one difficulty with this sort of definition, however. If one springs to the attack, he most probably would not notice that he was angry; if one were startled by a lion and instantly ran toward safety, he would probably not say he felt afraid *at the time*. If one successfully rejects a perception, he does not feel what he calls horror—the effect of the action is to *prevent* that feeling from arising, in each case. One truly *feels* an emotion when he has decided to attack, and then is prevented, when he is on the point of fleeing and discovers that he cannot, when he wants to stop something from happening and is powerless to do so. It seems that there is a third factor: one must select a goal, and feel the physiological state, *and have the action blocked*, before he would truly call his state *emotional*.

Sometimes this blocking of action is due to external circumstances. More often, at least in a civilized society, and far more destructive of organization, it can arise from internal *conflict of goals*. One wants to attack, but wants to avoid hurting anyone. One wants to flee, but wants to be thought of as brave. One wants to stop seeing unacceptable things happening around him, but dares not lose contact with what is going on. One's

own goals prevent action from occurring. In that situation, the poignancy of emotional states is sharpest, and the emotional state goes on and on, because the error never can be corrected unless somehow the environment cooperatively changes itself so that the goals no longer appear necessary.

Normally, however, one cannot stay in a continual state of emotion. Emotion is, according to the present discussion, a state of intrinsic error as well as a state of chronic error in the hierarchy of learned systems. Therefore, reorganization will be taking place as long as emotion persists. The result is that *somehow* one will change so as to restore the intrinsic errors to zero—to make the feelings go away. If one is chronically angry, he finds a way to become less angry—he stops going to places where he becomes angry, he stops seeing people who make him angry, he stops wanting to attack and tries wanting something else. He learns to predict situations that make him feel fear, and avoids them, or alters his judgement of what is threatening. Somehow, he will find a way to restore intrinsic error to zero; he will keep reorganizing until he does, or until he reorganizes himself to death.

As people grow older, we say that they mellow, or become more tolerant, or else lose interest in things and narrow their lives more and more. These are ways in which people deal with emotion; these are the signs of successful reorganization. Not every way of correcting intrinsic error, however, leads to what one would call the good life. All that is guaranteed in the long run, if the approach is random, is relative equanimity.

There is still one very important question left unanswered by the discussion so far. We have accepted as fact that an emotion involves both a goal and a feeling, but there has been no suggestion as to how the feeling state happens to arise along with the goal. In order to deal with this question we must make some hypotheses, but they are in line with what is generally accepted as *probably* true; the state of knowledge about the relationship between brain and body is still not all we could desire.

In earlier parts of this book, it was suggested that overt behavior was under control of a hierarchy that progressed from spinal systems (first order), to brain-stem systems (second order), to thalamic systems (third order), and thence to higher systems. This chain, however, accounts for only part of the

functions known to be carried out by the nervous system at these levels. Recall Hess' cats: in addition to the postural control he found when the thalamic regions were stimulated, he found other effects; nearby regions in the sub-thalamus produced effects that are normally associated with *autonomic* or vegetative functions. Stimulation in various places could cause the bowels or bladder to empty, the hair to stand on end. In other brain-stimulation experiments it has been found that stimulation of the hypothalamus can produce pleasure and pain, can simulate hunger, can alter the temperature of the bloodstream, and so on—in short, can affect many of the factors we are associating in this theory with the organizing system, particularly the intrinsic signals. Clearly the sub-thalamus is connected with the substrate of control systems we would associate with emotion; with intrinsic signals, too. The limbic system, the "endbrain" functionally just above the hypothalamic systems, appears to have to do with recognizing and controlling the "appropriateness" of feelings, organized feelings that would tend to be given emotion-type labels.

And nestled just *under* the hypothalamus is the hypophysis, containing the pituitary gland, which is half neural and half biochemical in structure. At this point the hormones are secreted which govern large subsystems in the hierarchy of chemical systems that make up the regulators for fundamental life processes; there appears to be the connection between neural organization and the "chemical muscles" of the body. Here, it may well be, neural signals act as reference signals that govern the major chemical feedback control systems; as is the case with the control of the thyroid gland, it is known that substances secreted by glands stimulated by hormones from the pituitary *feed back to the pituitary* to inhibit production of those very stimulating hormones; the hormones emitted by the pituitary appear to be the error signals of a group of biochemical negative feedback control systems.

Finally, in the brainstem that seems to be functionally below the thalamic regions we find automatic control centers, such as those that control respiration rate, heart rate, and vasoconstriction, and which send control signals to every major organ of the body. The whole system is utterly fascinating; a multi-levelled system that begins in the brain and continues down—

who knows how far? Perhaps the first-order systems in this biochemical chain are inside the cells themselves, throughout the body.

Not too sharply, I am afraid, but most tantalizingly, a picture begins to form of a second hierarchy of control that splits off from the behavioral control systems at about order two or three; this other branch is concerned with the sensing and control of quantities derived from sensors and from chemical messengers throughout the body. When a higher-order behavioral system acts, the downgoing chain of reference signals determines two complementary types of behavior; the behavior of the muscles which position and move the limbs, and the behavior of the biochemical systems which brace to meet the demands of the behavioral systems just as the muscles of legs and trunk brace to meet the forces created by the arms. Both branches are hierarchies of feedback control systems; one branch deals with the quantities that can be derived from sensory endings affected by the external world and the mechanics of behavior, and the other branch deals with quantities that are involved in control of the life support systems.

Setting a higher-order goal for behavior, then, above the level where these two major branches of the hierarchy meet, has the effect of altering reference-signals in both branches. In the behavioral branch, the effect is to produce patterns of perception via action of the muscles; in the biochemical and autonomic branch, the effect is to produce patterns of feeling-states, that arise as the biochemical balances in the body change in response to the commands.

Thus the model accommodates the relationship between "mind" and "body." Chemical disturbances will be met by feedback control action in exactly the same way that mechanical disturbances are met by muscular control actions; complex disturbances are met by higher-order control actions in both branches; it is not too hard to see how certain drugs might induce errors so complex that they reflect clear up the biochemical hierarchy into the higher-order control systems of the brain—that alter perception not necessarily by affecting the perceptual apparatus of the visual, auditory, tactile, or olfactory systems but by altering the feeling component of these perceptions, contributed by the lower-order systems of the

biochemical hierarchy.

Above the joining of the two main branches of the hierarchy, perceptions will contain information from both branches. At these higher-order levels, where one perceives relationship, program, principle, and system concept, a feeling component comes in that is absent in the lower-order perceptions. One does not see beauty in the brightness of light or the loudness of a sound, in a single chord, or an arbitrary configuration of visual patterns, or an isolated event. The beauty seems to appear only when one begins to appreciate relationships and such higher-order concepts. Beauty is in the eye of the beholder—or better, if less beautifully; beauty is a feeling-state that is perceived as part of a higher-order version of experience.

This brings us to the final subject, that of subjective value, of esthetic judgments. These are properly dealt with under the heading of emotion, because such values are felt primarily as an emotional response, either to their presence or (more likely) to their unwelcome absence or violation. The most puzzling aspect of value-judgments and esthetic feelings is the apparently arbitrary given-ness of the judgment. One may not know much about art, but knows when he sees something ugly. There is no *reason* behind such judgments, no practicality in the sense of saying "I like this the way it is because having it that way serves such-and-such a higher-order goal." At best, one can say only, "I can't help it—that's how I feel about it."

If there is any primary difference between the way the organizing system works and the way the hierarchy works, it is in the fact that intrinsic reference-levels are *not* learned, but are part of the inherited workings of our bodies. We have no choice but to consider all pain-signals as errors, and by that token we have no choice but to consider them *bad, unpleasant, unwanted*. Pain is not bad because of any other consideration —it is bad because it hurts; it hurts because it is bad. If we have been starved for salt, a little bit tastes extremely good; if we get too much salt in the soup, it tastes very bad. What tastes like the "right amount" of salt is something beyond our control, and beyond naive understanding.

Intrinsic state is our fundamental definition of good and bad. I am not trying to reduce esthetic judgments to the level

of a judgment of "not enough salt," but only to indicate that esthetic judgments are based on values that are every bit as built-in, and which constitute part of the definition of intrinsic state.

An esthetic error feels just like an emotional error; there is no reason to put it in a separate category. Reorganization can be motivated just as much by an attempt to produce a proper esthetic feeling as by an attempt to avoid pain or get enough to eat. Indeed, the esthetic drive is stronger in some people than the hunger drive; if, as can happen, there is a conflict that prevents simultaneous correction of both kinds of intrinsic error, the esthetic drive can cause reorganization to keep going even though the changes have an adverse effect on the state of hunger.

But as before, I would judge that behavior aimed at producing some esthetic effect would not be accompanied by what we term "emotion" unless the preparation went all the way to the point of action and then action was blocked, perhaps by conflict.

One last thought, lest anyone think that this is supposed to be the last word on emotion. It could be that these two hierarchies, the biochemical and the behavioral, are simply highly modifiable assemblies of control systems, and that the true nature of intrinsic state—the real variables that must be maintained at inherited reference-levels—have nothing at all to do with either. Certainly every organ system is highly affected, even to the point of determining the physical size of organs, by the exercise that the organ gets. Given a constant infusion of thyroxin into the bloodstream, so that the output of the thyroid is shut down via feedback through the pitutitary, the thyroid gland itself will shrink in size; this is just as much "learning" as would be the alteration of an output function in the brain. Every organ behaves like this. The physical makeup of the biochemical systems is not laid down in an ironclad pattern by the genes. So it might easily be that I have completely missed the mark in identifying the critical variables that *control* reorganization as being those involved in emotions. Perhaps emotional states themselves come to have the values they seem to have because of their effects on still other, and unsuspected, quantities that are even more deeply a part of our inheritance.

[1972]

An Experiment with Levels

A truism accepted by practically every working therapist is that progress in therapy depends on the patient's becoming aware of the true nature of his problem. This observation, indeed, was one of the reasons for which I suspected that awareness is associated intimately with the functions of the organizing system. If awareness is directed to the functioning of some subsystem in the hierarchy, this notion implies that the subsystem is being modified, or can be modified if there is intrinsic error. The requirements for successful therapy are quite clear if one accepts these ideas: *awareness must be directed to the systems that are in conflict, while they are in operation, and there must be intrinsic error.* The awareness alone is not sufficient: there must also be what is termed "affect"—an experiencing of the emotional consequences of the conflict. Since one does not seek help in order to become free of pleasant emotions, the consequence is that one must experience unpleasantness in order to make any important change in his own organization. It is not enough to achieve "insight" into oneself; the conflicted systems must be brought into operation, both at the same time, and the consequences of the conflict must be experienced as a disturbance of intrinsic state while awareness is directed to the conflicted systems. In terms of having more subjective meaning, one must consciously experience both or all of the desires that are in fact causing the conflict, so that the emotional consequences become plain in relationship to the action of seeking both conflicted goals.

Chapter 27 of the preliminary version of *Behavior: The Control of Perception,* used for teaching a course at Northwestern University.

I have not recommended any particular method for resolving the conflict that is discovered. Rather, the conditions for its resolution would be stated: be aware of the conflicted systems in action, and suffer the disturbance of intrinsic state. That initiates the process of change, which will continue until the intrinsic error no longer occurs when the systems operate. There is no prescription for what must be changed—perceptual interpretation, goal, or method of seeking the goal—but if the conditions are brought about, the use of "pseudo-resolutions" is ruled out automatically. As long as one's intention is to *experience the conflict*, not to avoid it, imagine it away, or fight it, changes will occur and they will not stop until intrinsic error is reduced below the level that calls for change. The only criterion for success is that the systems become capable of operating—that one be able to seek and reach goals of the type formerly involved in the conflict—without causing intrinsic error.

All that a therapist can possibly do is to persuade his patient to do what is necessary. The therapist cannot (or so I think right now) do very much to make change begin unless the patient himself understands *and practices* the necessary attitude. The therapist can, however, given the patient with the right intentions, interact with the patient in such a way as to bring the conflicted systems into operation. The therapist, being generally more experienced than the patient and understanding more of how the mind is organized (ideally, that is), is in a position to see when the patient has his attention fastened on consequences, symptoms, instead of on the higher-order goals that are the source of difficulty. With the patient's permission and cooperation, the therapist can in fact help to direct the awareness of the patient to level after level of system in the patient until the level is found where the conflict originates. With this concept of therapy in mind, I have developed the beginnings of a formal therapeutic technique. I am not sure whether what follows is on exactly the right track, but I have tried the system with several people and have experienced it as "patient" myself, and the results have been interesting.

The "levels" method

To set the stage for this discussion, I must first describe an informal mental model I use when thinking about how awareness is related to the operation of the learned hierarchy. This model represents an attempt to understand how it is that awareness can be directed to one aspect after another of the same perceptual field, so that without changing any of his inputs one may change the perceptions of which he is aware. One may, for instance, read these words, or while continuing to read them become aware of how he is sitting, or breathing, or of the fact that the letters are printed in black ink, or that "t" is taller than "a." The substrate of perceptions from which these abstractions are taken were there all the time, of course, so one may wonder what happens to them while one is *not* being aware of them. What I think is that at least some of them are in fact present as perceptual signals even when one is not aware of them, and so for the time being they constitute *unconscious* perceptions. Certainly when one is attending to some high-order perception, such as the meanings of the words on this paper, he must necessarily have all the lower-order signals present, even though he is not being aware of them at the moment.

Consider, then, that at any particular moment some of the systems in the hierarchy are active, but that awareness is involved only with a fraction of those active systems.

The first part of my mental model consists of the idea that awareness is mobile, can participate in the operation of *some* subsystems simultaneously but not usually *all*. What directs awareness to this subsystem or that, and what widens or narrows the scope of awareness in terms of the number of different perceptions involved at one time, I do not know. I know only that these phenomena do occur.

The second part of the little model is an interpretation of the way one *experiences* shifts of awareness. It seems to me that when awareness is involved with some particular system, the result of the combination is that one *takes the point of view of that subsystem*. In other words, one's conscious point of view at any given moment is defined by the subsystems with which awareness is involved, and however one's viewpoint shifts he

always ends up in some *familiar* point of view. If one restricted his awareness entirely to third-order systems, then he would consciously experience a world in which the only property was configuration—he would be aware of familiar shapes, contrasts, sizes, distances, patterns, and so forth, paying no attention to intensities, sensations, sequences, and the rest. Without altering the content of his sensory field, he could transfer his attention to fifth-order aspects of the same environment, and then he would experience the world as a set of dynamic changes: running, walking, flying, turning, shrinking, growing, and so on, and he could do this at the expense of the third-order viewpoint he had a moment before. Or he could encompass several levels, being aware for example of the consequence of movements involved in dialing a telephone number, and at the same time of the shape of the numbers and of the sensations from the dialing finger.

So when one is operating normally, his awareness is involved with some subset of the systems in the hierarchy, and these determine his point of view. The organization of the input functions determines what and how he will perceive, the goals given to the subsystems as reference signals become the subjective desires and expectations of the person, and the organization of output functions and their relationships to lower-order systems constitute what the person "knows how to do" at the moment. The person will experience himself and the rest of universe of perception only as these systems are capable of perceiving.

If one shifts his awareness now, so that new subsystems are involved, he experiences this as a shift *in his world*, not in himself. New aspects of the environment come into awareness and others drop out. The person feels this not as himself moving from one viewpoint to another, but as appearance and disappearance of aspects of the external world, with himself unmoving and stationary in the center—and unchanging as well. When I mention breathing, up to the moment I mention it, it does not exist in the conscious perceptual field; as soon as I mention it, it pops into awareness as if it had just been created out of nothing. We are all quite used to this phenomenon, and think nothing of it, but as with many such things we take for granted there is considerable mystery behind it. My

little model does not dispel the mystery, but it does provide a way of thinking about what is happening.

Now to the "Method of Levels." The objective of this method is to drive awareness to higher and higher levels of subsystem, with the ultimate aim of finding the level at which a conflict originates. Even if one is not concerned at the moment about a conflict, the method is still of interest simply as a way of getting to know oneself better, to see oneself more as a whole.

Two people can carry out this method better than one person can alone, although a solo performance may be possible. Let us designate the two people as Subject and Experimenter; the roles may be exchanged freely. The method involves a conversation between the Subject and the Experimenter, but a conversation of a rather exasperating type, for the Experimenter is listening to the subject not to discover what he is talking about but to discover the viewpoint *from which he is talking*: to identify, if he can, the level of subsystem with which the Subject's awareness is associated at the moment. To identify the level does not require using the levels I propose in this book; it requires only reminding the Subject at strategic moments of the real purpose of the conversation. Both Subject and Experimenter will frequently forget what they are trying to do.

When one's awareness is involved with some particular system, he is not aware *that* it is so involved. The world available to his awareness *is the world as interpreted by the subsystem*, and constitutes *the totality of that which is available to experience*. It is as if one peers out of a box with various kinds of windows in it; the window-glass distorts, filters, and changes the light, but since one has nothing to compare the scene with, he is not aware of the glass, but only of the scene. One cannot even back away from the window, because if he does there is nothing at all to see. The analogy to shifting awareness up a level would be for a person to perform some indescribable maneuver and suddenly see *the whole box he had been in*, with all its windows, and a quite-different-looking scene, from another viewpoint, surrounding that box. One would spend some time examining his former "viewpoint" before it began to dawn on him that he is still looking through a window, this time in a *different* box.

This realization will not come from backing away from the window and trying to look at the glass, because again there would be nothing to see. Rather, it comes after one finds first one window, then another, and then another in this new box, and realizes that the scene has a similar appearance through each one. Then he may be able to perform that indescribable maneuver again, and see *both* of the former "boxes"—viewpoints—again, of course, as he will eventually discover, through a new kind of window. These peculiar windows, which do the seeing as well as the filtering, always remain invisible themselves, and one is aware only of the scene they transmit.

In the same way when one becomes aware that the trees have begun to sway, he has no way of being aware—of sensing—that his awareness has suddenly become involved with a fourth-order system, and he is seeing the world as that system interprets it, in terms of motions. All he knows is that motion has suddenly appeared in his world. He may not even realize that the trees have been waving all day, and may think that a wind has just sprung up.

The "Method of Levels" begins as any normal conversation. The Subject chooses any topic at all, and speaks on it for awhile—he may decide to talk, for example, about what he is trying to do, about the meaning of "levels" and so forth, or he may start anywhere else.

There is, in fact, no "right" way to start, since any beginning point is as good as any other. If one finds it difficult to start, for example, then the conversation might proceed as follows.

Experimenter: So say something.

Subject: I don't know what to say. How do I start?

Experimenter: Talk about not knowing how to start.

Subject: But I don't—oh, I see. Well, I suppose you want me to think about why I'm having difficulties, or what is keeping me from getting started.

Experimenter: In other words, you're trying to do what I want you to do?

Subject: Well, why else would I be sitting here? Of course I am. I don't know what you're getting at. Do you want me to talk about that?

Experimenter: Are you waiting right now for me to tell you

what to say?

Subject (getting more irritated): Well, isn't that what I said? I'm confused. What *am* I supposed to do?

Experimenter: You say you don't know what to do, you want me to tell you what to say, and you're confused. Is that a fair description of your attitude right now?

Subject: It sure is. I'm sorry, I'm not doing very well.

Experimenter: Well, now that you are looking at those things, your attitude, what do you think about them?

Subject: I don't know what to say.

Experimenter: That's one of them.

Subject: Well, it's true. I don't.

Experimenter: And what do you think about that fact?

Subject (agitated, more irritated): What do I think about *what*? I haven't the foggiest idea of what you're getting at.

Experimenter: That's the same attitude, isn't it?

Subject: All right, *you* tell *me*. What *is* my attitude?

Experimenter: That's the question. What your attitude is toward the things you have been expressing to me.

Subject: Well, right now, I'm annoyed. I feel like giving up.

Experimenter: Anything more?

Subject: And you're not being much help.

Experimenter: Uh-huh.

Subject: I feel as if I'm being tested or something. I was trying to do what you wanted, but you wouldn't tell me. I was waiting for you to tell me what to say.

Experimenter: Now what do you think about having done that?

Subject: I suppose it wasn't very bright of me. But it's true, you can't just expect me to dive into something completely new like this and know right off the bat what to do.

Experimenter: Does that explain why you were getting irritated?

Subject: I guess I was irritated. But I felt you were being unfair, to expect me to know what to do without any explanation. It's like getting a question on a test about something you never had in class. I suppose I shouldn't get upset so much, but that has always bothered me.

Experimenter: This isn't a therapy session, exactly.

Subject: Oh, yeah.

Experimenter: Where were we?

Subject: Talking about tests. This being a test, and me being annoyed. I feel sort of blank.

Experimenter: I can wait, there's no hurry.

Subject (after a long pause): Am I supposed to be doing something? (Laughs.) That's where I was before, isn't it? I mean, that's where I was before. Period. I don't have to ask you, do I? (Laughs again.) I mean, I don't have to ask you. I can do it myself, mother. Mother—why should I think of that?

Experimenter: Free association?

Subject: Yeah, sort of. I suddenly thought of asking my mother if I could do something or other, and that TV commercial...

Experimenter: Uh-huh.

Subject: Well, uh... All right, I'm thinking about my mother for some reason, so I'll see where that leads.

Experimenter: How do you feel about seeing where that leads?

Subject: Huh? I mean... well, anyway I'm thinking about how she always used to tell me I had to think for myself, and then she'd land on me when I didn't guess what *she* wanted me to do...

Experimenter: What are you doing?

Subject: ... which makes me think of my father, too, and an ugly priest I knew when I was a kid...

Experimenter: Wait a minute. What are you doing?

Subject (irritated again): Well how can I do anything if you keep interrupting?

Experimenter: But can't you just tell me what you are doing? Describe it to me?

Subject: Well I sure can't describe it and do it at the same time?

Experimenter: Uh-huh.

Subject: Well, I'm in the middle of tracking down a lot of thoughts that came popping into my head. I'm trying to see where the trail leads. I'm remembering things. Or was. (Suddenly laughs.) I'll be damned. That was another one, wasn't it? I was all wrapped up in something or other—now let's see, what was it? I was sort of exploring, waiting, looking, trying to get somewhere—I felt I was doing something familiar; I

thought I understood a little of what was going on. I was remembering, a lot of things. Just to be doing it, because it was a relief to have found something to do.

Experimenter: A relief?

Subject: Well, I'd been feeling that I wasn't getting anywhere, and I wanted to get somewhere, and it was a strain.

Experimenter: What kind of strain?

Subject: Well, as if something was wrong with me—as if I wasn't very bright, as if I should be getting it and wasn't.

Experimenter: Something was wrong with who?

Subject: Me.

Experimenter: Is that the way you feel now?

Subject: No, not any more.

Experimenter: Are you different from what you were five minutes ago?

Subject: Well, I *feel* different. I don't feel the same way as then. It's surprising, to change like that. I was really irritated with you, and now I'm not. And I don't seem to feel that I don't know what to do, either. I still don't know what we're doing, but we seem to be doing something, anyway.

Experimenter: What do you think about changing like that? So fast?

Subject: I guess it's all right.

Experimenter: How do you mean that?

Subject: Well, I mean I'm not used to moods coming and going so fast. I'd like to stick with one thing for awhile, you know, not let go of something as soon as it gets interesting.

Experimenter: Stay with it, talk about it longer?

Subject: Yeah, explore it. You get curious.

Experimenter: Suddenly I've forgotten where we are.

Subject: Me, too. I was talking about... that's funny, I can't seem to keep my mind on *anything*. Oh, that was it! I was saying (breaks into laughter). I was saying I didn't like to flit from one thing to another. I guess I'm *really* thinking that I shouldn't be so—well, unstable. As if it's a personal fault, you know. But you said this isn't a therapy session.

Experimenter: Not supposed to be, anyway.

Subject: I guess you can't help getting involved, though.

Experimenter: You feel you're getting involved? What do you mean?

Subject: Well, like suddenly this has to do with *me*, personally. Not just ideas and attitudes and like that. I guess I'm bothered a little about finding *myself* changing so easily.

Experimenter: Yourself? Wasn't it yourself you were talking about before?

Subject: Oh, I guess so, but—well, yes and no. I'm the same now as I was then, but I feel different now.

Experimenter: What's different?

Subject: My feelings.

Experimenter: Do you mean you were the feelings or that you were aware of the feelings?

Subject: Aware. But I was them, too. It's hard to explain. I don't even understand it myself.

Experimenter: What feelings are you having now?

Subject: I don't know. Let's see—sort of curious, maybe a little edgy, too. I don't know where this is leading. I'm not too sure what's going to happen next.

Experimenter: Is that a feeling?

Subject: Well, there's a feeling with it—I suppose those are more like thinking.

Experimenter: Are you aware now of watching your feelings and thoughts and trying to decide which is which?

Subject: Yes.

Experimenter: Why? I mean, why are you doing that?

Subject: Uh... well, it's as if I'm sort of *watching*.

Experimenter: My mistake. Go ahead.

Subject: I don't get it. I mean, I see what you mean, I'm looking at thoughts and feelings, but I'm not in them—they're mine, but they're not me. I mean... I mean, I mean, I mean. Why do I always say that? And that's a thought, too. I'm watching it, too. And that, too. And... (Silence.) I see. (End of session.)

In a much-abbreviated and unnaturally-smooth-running form, that is how the "Method of Levels" proceeds. It ends where there is no place left to go.

At first, the subject was operating from a particular point of view, based on an attitude of wanting to be told what to do, and full of feelings of being lost and disoriented. It took quite a while for the experimenter to find a way to get the subject to begin talking *about* that attitude instead of *from* it. Typically the

subject resisted attempts to make him move up a level, and centered his resistance on the experimenter. The experimenter, remaining non-directive (relatively), simply kept trying to get the subject to describe his current attitude, the attitude *immediately* behind what he was saying. Instead of talking *about* the attitude, the subject *demonstrated* it: "What do I think about *what*? I haven't the foggiest idea of what you're getting at." The experimenter's response was not a response to that comment (such as, "Well, are you trying?" or "You will if you keep with it"); rather, it was something intended to bring the fact of the attitude to the attention of the subject: "That's the same attitude, isn't it?"

Then the subject began, in effect, summarizing all the topics which arose from his attitude: I'm annoyed. I feel like giving up. You're not being much help. I feel as if I'm being tested. I was trying to do what you wanted. You wouldn't help me.

The experimenter responds again not to those points, but to the attitude: *Now* what do you think about having done that? The subject responds in past tense, a pretty good sign that he has left or is in the process of leaving the former viewpoint. After a bit of fumbling, he does indeed leave it—and enters a higher level in which the issue is trying to do something, remembering, playing at self-directed psychoanalysis, and so on. When the experimenter interrupts him, the subject once more resists being dislodged. He objects that he can't describe what he is doing and do it at the same time, which is exactly the reason for the interruption. He cannot describe his viewpoint and hang onto it at the same time. By this time the subject is getting freed of involvement with many layers of viewpoint, and the process goes faster—at least the subject now recognizes as familiar the experience of popping out of an old viewpoint and seeing it before him where he can describe it. Near the end, he recognizes, for example, when he has slipped back *into* a viewpoint that he had been talking *about*: "I can't seem to keep my mind on anything. Oh..."

Then finally levels are approached which are behind all of what has been going on—the subject's self-awareness, his attitudes toward his own characteristics. The experimenter pushes him with unseemly haste (to keep the example short) into realization of the significance of all the shifts that have

been going on: that something has remained the same throughout, and that the various thoughts and feelings and attitudes are what have changed. When the subject finds himself contemplating those thoughts without thinking about them any more, he has followed this "tree" as far as it goes. I don't know just why this experience results in such a marvelous state of mind for awhile afterward; perhaps it is the fact that in that silence one can see the entire trail he has just followed; somehow to grasp it as a whole, even without analysis or thought, seems useful. Whether there is any truth or not, the feeling one is left with is one of understanding his own organization better than he did before—how he came to be the way he is.

Of course the process is not generally smooth. The experimenter gets drawn into the conversation when he pays too much attention to what the subject is saying, and forgets that he is trying to grasp the attitude, the point of view, from which stems all that is being said. He can get a fixed idea and push in the wrong direction. He and the subject can get going around in circles. The subject can run into attitudes that have tremendous emotional charge in them, and may become so involved in them that his resistance to moving on overwhelms the experimenter's efforts. The subject can operate from attitudes that the experimenter simply does not understand and so cannot help jog awareness. Of course these are all the same difficulties that arise in any honest attempt to unravel the organization of anyone's mind, and are familiar to every successful therapist. What the "Method of Levels" introduces, however, is a model that can guide the process of self-inspection, and can suggest to the experimenter how to keep the process headed where it must head: toward the highest levels of organization that are responsible for what all the lower levels are doing. This is indeed a shortcut, not without pitfalls and difficulties, but a shortcut. Best of all, it does not superimpose any arbitrary theoretical framework on what the experimenter says to the subject, or on what the subject believes about his own organization. Even though the idea of levels of organization springs (for me) out of the theoretical model in this book, it does not require one to believe that there is any specific structure for these levels, such as the five low-order

levels I tried to define earlier. Enough experience with this method, indeed, may give us the information we need to reach a decision about higher levels of organization—whether there are categories as universal as that of, say, sequentiality.

[ca. 1974]

The Illusion of Control

Any reader of Harvey Wheeler's compendium, *Beyond the Punitive Society*, must, if at all analytically inclined, wish he could add a chapter of his own. I will try to avoid the temptation, even though the liveliness and variety of the contributed chapters makes the temptation strong. Throughout this volume the discussions revolve around critiques, modifications, and applications of B. F. Skinner's provocative views—but this is not a book review, either. Let us say that Wheeler's volume has inspired and focussed my own analytic response to Skinner's work, by bringing together all the most cogent arguments concerning it. My contribution here will not be a 20th chapter, but a meta-chapter derived from a point of view not completely shared by any of the contributors. An uninvited meta-chapter, I hasten to add.

The title of this article suggests, I hope, that there is one serious question that has not been raised by Skinner, by his critics, or by his proponents either whole- or half-hearted. That is the question of just what we mean by "control of behavior," and whether what we have in mind by this concept is *possible*, much less practical or desirable. Nearly everyone assumes that it *is* possible, and bases his arguments on that axiom. My purpose here is to cast doubt on that axiom.

What is behavior?

Before we can speak of controlling behavior we must define

Unpublished manuscript.

what it is that is to be controlled, a step which behaviorists resolutely avoid taking. For scientific purposes it is not enough to have a sort of informal understanding that the meaning of the term is self-evident. Without a consistent formal definition of behavior, we can have no consistent formal theory about behavior.

The difficulty with reaching agreement on a formal definition is that behavior is so strongly influenced by the observer's way of perceiving it. As an example, take a man standing on a grassy field scratching his nose. What is his behavior? Is he moving a finger? Waiting for somebody? Creating a bit of reinforcement for himself? Not at all, as any decent baseball fan would point out in disgust: he is initiating an attempt at a double steal. The behavior one sees is determined *mainly* by what he knows and by the level of perceptual abstraction he applies to the raw stimuli affected by another organism's motor activities.

This fact has long been recognized, but has been treated by behavioral sciences primarily as one of those inconveniences that Nature inflicts on scientists. A large part of the philosophy of behavioral science has been concerned with the problem of how to construct an objective science of behavior that can somehow bypass this unfortunate effect of subjectivity. And in the background has been the assumption (often, apparently, quite unconscious) that there *is* a way to bypass human perception and study Reality Itself—what else could justify any degree of reductionism? What other basic assumption could have been behind Skinner's justification of forcing a child to ask for and consume yogurt; namely, that yogurt is actually a delicious food?

The behavioral model I am developing proposes nine levels of human perception: nine hierarchically related classes of perception, each derived from those of lower order: intensity, sensation, configuration, change, event/sequence, relationship, program, principle, and system concept. In effect, the assertion is that a human being placed in *any* environment, whatever its "real" character, will experience it in terms of intensity, sensation, configuration, ... etc. Such perceptions are the consequence of placing a human perceptual system (I cannot yet speak confidently of other species) in contact with

an environment. Naturally, we all agree, as a result, that the environment contains entities of just these classes.

A moment's thought will show that these classes of perception encompass all levels of behavior, all levels of reductionism or globalism (all, that is, to the degree that I have accomplished my purpose). No level is the "right" level for viewing behavior; they are all ways of representing in our heads something we cannot actually perceive at all, the *cause* of these perceptions.

This picture, which seems forced on us by all we know about the nervous system, implies that whatever else behavior may be, it is certainly a phenomenon in the brain of an observer; my theory adds that it is one or more of nine distinct and hierarchically-related classes of phenomena. Within any class there is room for great variety, but the classes themselves, I propose, are universal.

Taken by itself, this theory of perception seems only a fancy way of restating the well-known fact that interpretation enters into perception of reality, although of course my theory goes farther than that cautious (and rather wishful) statement (to say "enters into" implies that objectivity is *possible*, even if difficult). The perceptual model, however, is only half of a general *control-system* model, and when the whole model is considered, this subjectivity, layered or not, takes on a new significance.

Behavior, according to this model, is the process by which the brain *controls* those perceptions with respect to which it has goals. As far as the brain is concerned, its outputs are the neural signals that are sent out of the nervous system into an environment of muscles, glands, and other things. The brain's *perception* of its behavior occurs in terms of intensities, sensations, configurations, change, and so forth, in any sensory modality or combination of modalities.

The activities of an organism that occur during this process of controlling perceptions are all that an external observer can see of behavior. If the behaving system and the observer share common classes of perceptions, however, the observer will see certain regularities in the *consequences* of motor behavior, although not necessarily in the motor behavior itself. These regularities are seen *as* regularities not just because they re-

peat, but because they repeat against the background of a *changing* environment. If a rat consistently finds its way through a maze to a goal, we might *suspect* that this regularity is a feature of the rat's behavior. This suspicion would disappear if tilting the maze caused the rat to end up, however consistently, in a different place. It would also disappear, and indignation would take the place of suspicion, if we discovered a slot in the floor of the maze and a projecting rod fastened to the rat's body. We will not accept a regularity associated with motor outputs as behavior unless it shows *some* degree of independence of environmental conditions, and *some* dependence on motor activities.

The main point is that whatever regularity we *do* accept as behavior, it is going to be a regularity of intensity, sensation, configuration, etc. And if the behaving organism is in fact *controlling perceptions* of those same classes, there is no surprise in discovering that *our own* perceptions of these same classes are stabilized against disturbances by the activities of the organism.

Behaviorists treat behavior as being the *output* of an organism; the very idea of an organism "emitting" responses says so. The control-system model, on the other hand, suggests (persuasively, to me) that behavior as we commonly define it is a *consequence* of whatever an organism actually "emits," and a *perceptual* consequence at that, for observer and behaving system alike.

To understand a behavior, then, is to share a perception, or at least think one is doing so. If one chooses to call that perception the organism's *output*, one is simply misleading himself and creating for himself and others immense philosophical and practical difficulties. Prior to Norbert Wiener, there may have been an excuse for doing so, but there is none now.

What is control?

I have already used the word "control" in speaking of a control-system model, but present development is not confined to the technicalities of control theory (although it must not violate the rules). In our common usages of this term are all the elements of a definition that is consistent with control

theory and which asserts nothing that is impossible for a control system to do.

We must first distinguish between two sets of terms that are often used interchangeably when the speaker intends to say "control." Going with the word "control" are terms like "regulate, adjust, set, stabilize, direct, guide." All these terms apply directly to the behavior of a negative feedback control system. Another set of verbs is often used loosely when "control" is intended: "affect, influence, change, alter, impinge upon, stimulate, initiate, drive, cause, force."

The first set necessarily implies the second: to "regulate," for example, one must necessarily "affect." But the second set does *not* entail the first. We would agree we could *influence* the behavior of a watch with a hammer, but we would not accept a hammer as a means of *regulating* the watch.

The difference between the two sets of terms is that the second applies to *consequences* of actions, while the first, while including consequences, adds a *goal*-consequence to the picture. When one *regulates* the flow of water from a faucet, he does indeed *change* the flow, but he does so relative to a *particular* flow that he has in mind. Regulation implies a *directed* change, a change toward some predetermined state.

It is clear that even B. F. Skinner uses this meaning of "control." When he speaks of using shaping procedures, for example, it is clear that he accepts existence of some *particular* state of the pigeon's behavior toward which the experimenter continually guides actual behavior. No behaviorist would expect his colleagues to permit him to predict the final state of "controlled" behavior *after* it had occurred.

In every moderately careful usage of the term "control," or its associated terms, two entities are implicit: (1) a perception of the way things *are*, and (2) a "perception" of the way they *are to be*. The latter, of course, has been the source of heated debate in psychology. The old debates, however, centered around the unspoken assumption that by a goal we must mean the way things *will* be, which is not at all the same thing I said above. Intentions can be confused with prediction of an objective state of affairs only if one believes that his perceptions are really Out There. The control-system model renders this debate obsolete, by providing a real present-time *reference*

signal, or perception-like signal generated in the brain against which the actual state of a perception is compared. Behavior is driven by the magnitude and direction of the discrepancy. Thus the behavior of a properly designed control system will always bring the system's *actual* perceptions into congruence with its *intended* perceptions, and maintain the congruence despite external disturbances or variations in the reference signal that is the intention. It will do so, that is, when circumstances do not rule out success.

Control always implies some kind of action. It is easy, and indeed customary, to allow one's attention to be captured by the action or some of its consequences, and forget the necessary relationship of the action to both perception and intention. Thus we can speak of controlling the path of a car by turning the steering wheel, but we would mightily resist being asked to do that blindfolded or in an opaque fog. Cutting off vision would in no way prevent one from turning the steering wheel, but it would certainly prevent him from using that action *as a means of control*. Thus everybody really knows that controlling means controlling perceptions, although not everybody puts it just that way.

It can be said that everybody really understands reference signals, too. If I am taking someone to an unknown destination and he is driving, I do not direct him by telling him how much to turn the steering wheel. I describe goal-perceptions to him: "See that schoolhouse? Turn right at the gravel road that runs around behind it, not the blacktop road." I give him enough information to permit *him* to construct the required specific reference-perceptions. We all do that, although, again, we don't all describe it just this way.

How we talk about control is a matter of choice, but the components involved are rigid requisites. There can be no control without (1) an inner representation of its intended aim, (2) perception of the actual state of affairs, (3) comparison of intention and perception, and (4) action properly based on the discrepancy resulting from (3). Control can occur if and only if the controlling system has the above components and functions. Just try leaving one out.

Parenthetically, there is one usage of control which fails to fit this definition. Before feedback theory was invented

(around 1940), the only explicitly known way of controlling a process automatically was to fix all the initial conditions precisely for a precision mechanism, and then give the mechanism a shove. Assuming constancy of the laws of physics, trusting in one's workmanship, and doubting that any mechanism could have purposes of its own, one could then confidently expect and predict that the mechanism would behave in a foreordained manner.

The emphasis in this view of control is on its "hands-off" aspect, especially in science. What many scientists have meant by a "controlled" process is one that will proceed in exact accord with prediction *without human intervention*. One feels he can "predict and control" a phenomenon if he can say what will happen, all by itself, after he pushes the "go" button and sticks his hands in his pockets, grinning. To reach out and twiddle a dial or add a pinch of reactant *during* such a demonstration would be grounds for serious accusations of cheating. Of course *any* experiment can come out right if the experimenter is allowed to influence it during its course (that is, act as a control system). Needless to say, this concept of control is irrelevant to this discussion, if not to others.

What controls behavior?

B. F. Skinner says that the consequences of behavior control behavior. While we can forgive some psychologists for delighting in phenomenon-names that seem to thumb their noses at physical causality (e.g., "retroactive inhibition"), Skinner has gone too far in this case. In any operant-conditioning experiment, it is plain to the eye that behavior is causing reinforcement; one must have an active imagination to assert that he sees reinforcement causing behavior. I am not *against* an active imagination; I just prefer people to label their "facts" as metaphors when that is what they are.

To be sure, when a contingency is established, a schedule of reinforcements, reinforcement is made to depend on behavior in a *particular* way selected by the experimenter. The organism is generally not free to seek an environment organized more to its tastes, and so, if the reinforcer is to be brought into range for eating, drinking, or otherwise having, must tailor its out-

puts to the rules that are in effect. I would not be surprised if this change of behavior sometimes actually required the interior organization of the organism to change in some way, so we could say that "learning" had happened.

If the subject organism is organized as a control system, however (and I flatly assert that to be the case, always), a change in its output accompanying a change in the schedule of reinforcement is no indication that anything at all has changed inside the organism. After initial acquisition of the necessary new control system or activation of an old one, behavioral output will quite automatically adjust to a great range of disturbances and changes in the feedback path (which is where we find schedules of reinforcement implemented).

Consider the well-known experiment described by Skinner in which pigeons trained to peck for food had the number of pecks required per pellet progressively increased. Acquiring this mode of getting food, I would agree, probably entailed learning. But as the number of pecks required increased, there was a very nearly proportionate increase in the rate of pecking, which a moment's thought will show must have resulted in a very nearly *constant* food input. From the experimenter's point of view behavior—that is, pecking output—was certainly controllable by varying the schedule. But that was possible only because the pigeons were *not* controlling their pecking output. They were *varying* it as a means of controlling *food input*.

If in the middle of this experiment all the pigeons had suffered an internal change that halved their goals for food-input, pecking rate would have abruptly halved *without* any change in the schedule. The food-input would have faithfully followed not the schedule but the inner reference signal, probably to the experimenter's dismay. Plenty of experiments of this nature have been done, although not for the purpose of proving *my* point, using surgical intervention in the hypothalamus. Reference signals control inputs, and inputs are all that control systems know or care about.

A simple control system with fixed properties *could* display exactly the changes of behavior Skinner described. Learning is not *necessarily* demonstrated by any such experiment.

None of this is meant to imply that B. F. Skinner has been wasting our time. Indeed, he has brought a segment of psy-

chology halfway to control-system theory by inventing exactly the *type* of experiment that has always been used by control engineers to analyze control behavior. If I come down hard on some of his notions, that is only to prevent their continuing to mask Skinner's real contributions, the ones that are permanent.

It is instructive to consider what is *not* controllable in an operant-conditioning experiment. First, establishing a schedule of reinforcement does not *ever* enable one to predict whether the "reinforcer" will ever be produced, or if produced by accident will come to be produced by behavior consistently. If we arranged for a pigeon's pecking of a key to connect the pigeon with the telephone time service, I doubt that the bird would exert much effort to cause that consequence. Experimenters do not choose reinforcers at random; they try to guess what will *matter* to the subjects. In other words they try to find inputs that the subjects will control relative to some fairly stable goal-state. This leads to confusion: some goals, as for breathing, are built-in, but others are learned. Both kinds look alike in an operant-conditioning experiment.

The experiment does not control the *status* of a consequence as a reinforcer. Neither does it control *how much* reinforcer will be generated by behavior. The schedule merely establishes how much effort is required to maintain *any* given level of input, or how the input depends on the time-pattern of activity. The amount of behavior, and thus the amount of reinforcement, turns out to be whatever it turns out to be.

The best the experimenter can do by way of predicting behavior is to say that if this pigeon is like other pigeons, it will eat grain, and if it ate a certain amount of grain every day when free-feeding, it will endeavor to eat the same amount during the experiment. Since those predictions apply to all pigeons in all experiments, there is not much left to predict about the results of a given experiment.

What is left to predict is removed by application of elementary control theory. It can be shown that a control system meeting a few simple criteria (high error sensitivity, proper rolloff in frequency response) will maintain its input at a steady reference level despite disturbances or a considerable range of changes in the nature of the feedback path. Predicting

the output behavior of such a system is trivially easy. If a disturbance would affect the relevant input variable by an amount D, if unopposed, and if an unopposed change of output Δq_o would affect the same input by an amount $E = f(\Delta q_o)$, the actual behavior will be $\Delta q_o = f^{-1}(D)$. In other words, if you want the output to be a predetermined function of disturbances of the input, you just place the *inverse* of that function between the output and the input. Want the output to be the time-integral of the input? Put a differentiator in the feedback path. Want the output per unit input to double? Halve the amount of input created by a given output. This is known, in other circles, as analogue computing.

Of course in order to use this sort of experiment to explore the characteristics of the control system itself, one must be sure the system can operate normally, i.e., not be subjected to extremes of disturbance or very complex feedback functions. That is hardly ever done in behavioral experiments. An animal normally capable of controlling its body weight within a few percent for weeks at a time, through varying its reference levels for food input (my model is hierarchical), is quite commonly prepared for an operant-conditioning experiment by starving it down to 80% of its free-feeding weight. The major control system in charge of eating, therefore, must be operating throughout such experiments in a condition of extreme overload. As a result, subordinate control systems, too, must be showing gross distortions both in goals and in actions. Surely one cannot say that a pigeon is showing normal behavior when it pecks so hard and fast for its meager pittance of grain that it wears its beak down to a stub! One can say that the pigeon has a hell of a control system, but he can't say anything interesting about how it works. One can only say, "Gee whiz, will you look at what that bird has done to itself!"

Operant-conditioning experiments done under more normal conditions can reveal a great deal about the internal organization of behaving organisms. Under the *usual* conditions in the laboratory, they reveal little information, and much thoughtless cruelty.

So what *does* control behavior, the question that began this section if anyone has forgotten? My answer, which I think can be fully justified in every detail, is "organisms control their

own behavior." They do not, of course, control what they are not controlling, which includes many aspects of the environment and very often aspects of their own *outputs*. They do control a great array of consequences of their own outputs, perceptual consequences which we have, through an understandable error, called behavior and thought of as outputs.

To be more precise, perceived consequences of outputs are controlled, in effect, by the organism's goals or intentions which exist in the physical form of reference signals in the organism's brain. Somewhere in the background must be *inherited* reference signals that tell the organism, in effect, what to like and what to dislike. This is not the place, however, to follow *that* long path. The point is that the *environment* does not tell the organism what it will find reinforcing or how much reinforcer to want, what perception to prefer and what *state* of that perception to seek. In the final analysis those crucial factors are determined inside the organism and cannot be controlled from outside. Most of the situations to which we now so carelessly apply the term "control" require quite a different explanation from that now offered by behaviorists.

Conclusions

I hadn't meant this paper to be so theoretical, but there was little choice; control theory is not yet a taken-for-granted tool of the behavioral or social scientist. The final paragraph above says, essentially, what I had hoped to say, but it wouldn't have made much sense by itself. Control-system theory has immense implications for the way we think about ourselves and others, for the way we attempt to structure our society, even for our concepts of our relationship to other living things.

Control-system theory implies nothing incongruent with the goal in the title of Wheeler's book. Neither, for that matter, does Skinner's approach in which the substitution of reward for punishment, a revolutionary improvement in human relations, is a central theme. But I think control theory answers most objections by Wheeler and his colleagues, and takes us toward a concept of human freedom that Skinner can only point to with a wavering hand and a self-contradictory motto. Behavior is *not* controlled by its consequences. Quite the oppo-

site.

One last word. A theoretician who works slowly and often in isolation sometimes generates new ideas that others, it turns out, published some time before. There is something of an anomaly in my case. This theory was first published in 1960—and the earliest citation of it in the literature, as far as I know, concerns a 1957 paper. I saw the citation in a Russian book. But due to a number of quite ordinary reasons, the 1960 paper generated little general interest, and in effect I gave up for ten years, consoling myself by writing the book. While I was recovering from my rather unrealistic disappointment other researchers began to arrive at one or another of my earlier conclusions, and there were even some popularizations that seemed, although I doubt they were, lifted straight from the 1960 paper. As a result of my disappearance from the scene for so long, I found myself being accused, in 1970 or 1971, of being behind the times, not having heard of this or that fragmentary theory that my words seemed to resemble (and no doubt did). I hope these remarks will explain why I am a little hesitant now to relate my work too closely to that of others, even though some of Wheeler's authors—Rosen and Pribram, for example—are far from neonates in control theory. I do not believe that the general picture I present here has been exactly duplicated before.

[1979]

Control Theory for Sociology

As a theoretician and an interloper in this volume, I will not attempt here to translate from control theory into Meadian sociology. That task is better left to the real sociologists who, seeing that such translations were possible, invited me to provide some background at a 1978 meeting at Kent State University. Having benefited from discussions at that meeting, I will not follow the format I brought to Kent State, but will try to speak directly to what now seem the most germane points.

The basic issue involving control theory is that of the nature of sociological principles and laws. Control theory is above all a theory of *individual* organization; it purports to describe how any one person is put together. If it has anything to say to sociologists, its contribution will be in the form of suggestions about the kinds of interactions that are to be expected when living systems organized as control systems encounter each other. I understand that this was one of Mead's concepts as well: that social laws are emergent, and arise from individual characteristics.

In the following I will first talk *about* control theory and its relationship to other theories, and only afterward explain what it is. Understanding what it is not is of considerable help in interpreting its meaning, and avoiding the trap of reinterpreting it in terms of something more familiar, but quite different.

Written for an unpublished compendium of papers given at a conference on George Herbert Mead at Kent State University, Kent, Ohio, 1978.

Control theory and cause-effect theory

To grasp the broader meaning of this theory, it is necessary to compare it with the right level of other theoretical concepts in psychology. Control theory does not compete with theories such as the proposal that people express emotions most strongly on the left sides of their faces, or any proposition of the form, "Under condition A, people tend to show behavior B." In fact there is really only one theory that is directly competitive with control theory.

That theory can be stated, "What an organism does is determined by what has happened to it and what is happening to it." This is the fundamental thesis of behaviorism, and to many life scientists it is a statement of a basic scientific principle. It is thought by many to be no more than a statement that the universe, including organisms, obeys regular and discoverable natural laws. Once the control-theory model is grasped, however, one can see that the theory above *is* a theory, a particular model of organization, and not the only possible one. Seeing what seemed a basic premise as a testable hypothesis opens a host of questions once thought to be settled, even if no alternative is proposed.

One way to state the traditional model is to say that the behavior of organisms is caused by past and present stimuli: this is so-called SR theory. When an explanation is called "mechanistic" (by friend or foe), the particular mechanism envisioned is one in which causes enter the system from independent external sources, and through some mediating train of events, produce effects in the form of actions. While one can certainly not blame life scientists for not being engineers, the fact is that this concept of mechanism is wrong, if it is taken to mean that this is the *only* kind of behavior of which mechanisms—organizations constructed of matter—are capable. It was the only kind known (formally, to academicians) before the 1930s, but it is no longer the only kind known. This means that the basic cause-effect concept behind most life sciences is not basic, but amounts to choosing one type of mechanism —for no good reason—over another, as a model of organisms.

Control theory is concerned with a type of mechanism whose actions depend (in part) on stimuli from the outside

world, and *simultaneously* are strongly determinant of the effective stimuli which act on the mechanism. This sort of mechanism is built of the same kinds of materials as the other sort, and obeys all the same physical laws, but obeys quite different *organizational* laws (laws characteristic of the arrangement of components rather than of any individual component). In fact, it shows that calling an explanation "mechanistic" means nothing more than that it concerns something made of matter. Matter can be organized in ways that the pioneers of explanation in the life sciences never imagined.

The principal difference between a control system and an SR or input-output system is that the SR system can have no internal purposes, but the control system can. The actions of the SR system are made to happen by whatever stimuli impinge on it; the system itself simply does what it is caused, from outside, to do. The actions of a control system, on the other hand, are organized to *control* the effective stimuli reaching the system: the system itself determines what it wants to sense, and acts in whatever way is necessary (and possible) in order to create the preselected inputs. The SR system *obeys* the implied commands of the outside world; the control system *resists* them.

Clearly, a control system has exactly the properties that were thought, during the formative years of all the life sciences, to be metaphysical fantasies. A control system can select some state of affairs that does not yet exist, and under favorable circumstances can bring about and maintain that state of affairs even in the face of pronounced and unpredictable disturbances. It does this by monitoring some aspect of itself, the outside world, or both, comparing what it senses with an internal reference, and using the discrepancy at each moment as the basis for directing and changing its present-time actions. Those who rejected the idea of purposive behavior did so for invalid reasons; they thought that this kind of action required seeing into the future, which it does not. It requires only the ability to sense the present state of affairs, compare it with an internal model of an intended state of affairs, and use learned skills that will reliably eliminate the difference. Since the late 1930s, the principles of operation of systems that can act in this way have been quite thoroughly understood—by engineers.

There is thus a dividing line in the history of psychology

and all the other life sciences, located somewhere around 1939. Before that date, all life scientists had an excuse for assuming that no physical system obeying natural laws could seek goals, have intentions, or otherwise manifest the symptoms of internal purposes. After that date, the only excuse was ignorance; continuing to base fundamental thinking about organisms on a simple cause-effect model had become an anachronism. I do not pretend to know why such anachronisms are defended so energetically, but they are and always have been. Control theory cannot do anything about them; it can only exist, and wait to be understood.

Control systems were invented in order to imitate the human ability to act on an external variable so as to maintain it in a preselected state; they were intended to be labor-saving devices, not models of living systems. Nevertheless they *were* models, working ones at that, of purposive organisms. Before the invention of control systems many people had insisted that human beings and perhaps other organisms were purposive, but nobody had the faintest idea of how purposiveness worked. Now we know at least *one* way.

How control systems work

All control systems we now understand can be divided into three major functions, no one of which could be called a "control" function (Control is an emergent phenomenon). I will explain these functions in some detail, with particular reference to features that would be used in a model of human organization. Then we will put this kind of system in relationship to an environment and an observer outside the system.

Perception

One function is *sensing*. To an engineer, sensing means creating a *signal* inside the system that in some way represents something outside the system. This does not mean that the signal looks like what it represents. On the contrary, the signal is merely a convenient physical variable (say, a train of neural impulses) that the control system can handle, and which, through use of suitable transducers, can be given a magnitude

that corresponds at every moment to the magnitude of some external variable. All sensing, however complex, can be reduced to a set of *one-dimensional* signals, each representing one degree of freedom of whatever is being sensed. Even if the system does not work this way, there are tremendous advantages for the modeller in transforming its sensory signals to an equivalent collection of one-dimensional signals (taking interactions, of course, into account).

There is more to sensing, however, than merely converting the magnitude of one variable into the magnitude of an analogical internal variable. Even artificial systems, which are very simple compared to living ones, do better than that. It is possible to equip a system with a multitude of sensing devices, each reporting the state of one localized physical variable, and then to *combine* the resulting sensory signals in a complex computing apparatus. This computing apparatus can continually calculate the value of some *function* of all the elementary sensory signals, creating a new signal that depends on complex relationships outside the system; the density of something, the entropy of a chemical reaction, or (to skip a few levels) even the value of an opponent's last move in a chess game.

The value of such a function exists, of course, as another signal, a signal that depends on a number of more elementary signals. If we think of the elementary signals as *first-order* representations of the external world, then a continuously-calculated signal is a higher-order representation. It represents, perhaps, phenomena that affect many sensors at the same time.

Given a set of such higher-order signals, we can obviously introduce still more computing devices, and derive signals of still higher order, and so on for as many levels as we need to model a particular organism. Thus the sensing part of a control system is potentially the source of a *hierarchy of perceptions*.

The term "perception" in this context means, of course, an internal representation in the form of a signal. Epistemological questions abound, and we will ignore them. As far as any person is concerned, perceptions *are* the outside world and one's body, and one does not need to settle any epistemological issues to live in terms of these perceptions. In artificial

control systems there are no such issues; the only world with which an artificial control system can deal is the one that exists as a set of signals inside itself; their relationship to the outside world is fixed strictly by the kinds of sensors it has and the way the sensor signals are combined by internal computations, but the system needs no knowledge of that relationship (which can be completely arbitrary and independent of what is "really" outside).

Comparison

Given only an internal signal standing for the state of some variable or complex relationship, the control system could experience the external world, but could not control it. Also required is an internal reference against which to compare the signal, and of course some apparatus for actually making the comparison. Since this model deals with simple signals, the internal reference can also be a simple signal. The internal reference is generated by processes independent of the perception; higher-level systems, perhaps. Since the internal reference and the perception to be compared with it are both simple signals (however complex their origins), the process of comparison itself is hardly more complex than subtraction. If the perceptual signal, to pick a specific possible model, *inhibits* the same neurone that is *excited* by the reference signal (as the internal reference is called), the discrepancy between the two signals is simply what is left over after the inhibition. The neurone emits a new signal that represents the excess of excitation over inhibition, quantitatively. This new signal is called the *error* signal in control theory.

Action

The error signal indicates the direction and amount by which the current perception fails to match the internal reference or reference signal. This error signal can thus be used to direct motor actions which, by affecting the outside world, alter the same perception that is involved in the comparison —and hence alter the magnitude and direction of the mismatch. In order for a control system to exist, the error must be

translated into just the action that will tend to *reduce*, rather than increase (or fail to affect) the mismatch.

Here again we see a tremendous advantage in dealing with *simple one-dimensional control*. Even a very complex control system that controls complex variables through complex combinations of actions can be represented as an equivalent set of simple control systems, each producing a one-dimensional action and controlling a one-dimensional aspect of the world. If nothing else, this picture makes the logic of control clear. If the perceptual signal is one-dimensional, it can only increase or decrease. An action on the outside world, therefore, working through external physical connections that ultimately reach the sensors, can have only the effect of increasing or decreasing the perceptual signal. Likewise, independent disturbances that also affect the sensors independently of the actions can only increase or decrease the perceptual signal. Thus choosing the right action which will tend to reduce the above mismatch becomes relatively simple: only the *sign* of the action has to be right. A positive mismatch should lead to one action (its amount proportional to the amount of mismatch) and a negative mismatch should lead to the *opposite* action. If the link from mismatch to the choice of one action or its opposite is selected properly, the whole system will tend to reduce the mismatch whether the perception is less than or greater that the reference signal. Since that covers all possible cases, the control system will operate properly for all states of the variable being sensed.

For a control system to continue working properly, of course, the effects of its actions on its perceptions must not change sign, or take on vastly different dynamic characteristics. There could be no control systems if the external world had randomly-shifting properties, or inconsistent relationships among causes and effects. Every control system must be adapted to its environment, in the sense that once it is set up its successful operation depends on certain basic constancies in the rules. This does not mean *absolute* constancy, but only that a given action should continue to affect the same perceptions in the same direction, at least in circumstances where a given control system comes into play. The possibility of successful control depends on the existence of natural regularities in

exactly the same way science itself does. In order for any control organization to be useful, it must work for at least a respectable range of the environments that are encountered, and by the same token, there must *be* some environments that are hospitable to control systems.

The control system from outside

The above picture is a model of the internal organization of a control system. I went through it first simply to show that there is a feasible and not too complex kind of organization that can account for control. It does not matter greatly whether an experimental or naturalistic scientist grasps every nuance of how such systems work, any more than it matters whether an SR theorist really can explain what stimulus was responsible for each "response," or how the internal connections work in detail. Such investigations can properly be left to full-time theoreticians, who will gladly check up on each other.

Psychologists, sociologists, and many others deal with the externally visible phenomena of behavior. We now consider how a control system appears to interact with the world in which it resides. There are typical symptoms of control that can be used to identify control systems, and to understand how they operate singly and on each other.

Control of perception

In the preceding exposition I have rather casually referred to control systems as controlling their *inputs*. This concept is not the one which most people—even engineers—have used. It is commonly said that control systems control their *outputs*, their *actions*. To see exactly why that is an incorrect interpretation one must come to understand how control systems work in some detail, and that requires something I can't supply here: study. I recommend any good text on introductory control theory, or an excellent book by Richard Jones (1973), or my own book (Powers, 1973). For the present, I will do my best to make sense of this somewhat odd proposition without getting into details.

I probably overuse this example, but it is familiar to more

people than perhaps any other simple one: steering a car. The action involved in steering a car consists of applying varying amounts and directions of twist to a steering wheel. Most experienced drivers watch the general appearance of the road and car framed in the windshield, and base their steering actions on that appearance. One appearance is "right," and deviations are corrected by means of twists on the steering wheel.

The "right" appearance does not exist in the windshield, but in the driver's brain. Most of us have at least a faint sense of what it is, although our attention is usually focussed much more strongly on the *actual* appearance. If asked why we twisted the wheel to the left, we would probably reply that the car was too far to the right, without explicitly answering the question, "Compared to what?" Yet with a little introspection of a simple sort, we can all see that the "what" is some conception of how the picture in the windshield looks when the car is in the right position.

No experienced driver would answer "Because the *steering wheel* was turned too far to the right." When we drive we are not concerned with the position of the steering wheel. We have no goal for that position. Instead, we tacitly agree to move the steering wheel to *whatever* position will make the picture in the windshield look right. In this case at least, it is clear that what we are controlling is not the action of steering, but a *perceptual consequence* of that action.

Furthermore, all experienced drivers know that there *is* no position of the car on the road that goes with any particular steering action. On one occasion keeping the car centered in its lane may require maintaining a constant twist to the right, on another occasion a constant twist to the left, and on still another, a constantly-varying and apparently randomly-changing position. A person who actually controlled his *actions* while driving a car—who maintained some predetermined or preferred position of the steering wheel—would not last long on the highway. All control actions while driving a car are centered around maintenance of a particular *perceptual* situation, which we have learned by slow experience corresponds to a desired physical situation.

Disturbances

The reason why control systems control their own perceptions and not their actions is precisely the reason why control systems exist at all. Organisms do not behave in order to have objective effects on the world outside them. They learn to behave only from effects of that behavior on themselves, and the whole point of behavior is to *control* those effects. Why else would any organism produce outputs? With perhaps a few esoteric exceptions that are still being debated, organisms evolve and learn to produce those actions that will have favorable effects on themselves, not to have particular effects elsewhere. Certainly the physical environment does not want organisms to affect it in any particular way.

If the environment spontaneously provided just those effects on an organism that promoted the organism's survival or short-term well-being, in every possible respect, there would be no need for behavior. Indeed, behavior would be contraindicated, since any action would then cause changes away from that ideal state. When one's children are behaving in ways that satisfy one's own reference levels for a happy family, one does not admonish them. When one's car happens to travel down the center of its lane while one is exerting no steering effort left or right, any steering effort would *cause* an error, and one leaves the wheel alone.

The environment is not so cooperative, nor so understanding of an organism's needs. It will not spontaneously assume the state that matches an organism's reference levels. Furthermore, even when behavior succeeds in bringing about the inputs the organism needs, there are disturbances from independent events which randomly tend to alter those inputs. The action which corrects an error today will enlarge it tomorrow and be insufficient the next day—or the next second. The inputs to an organism do not depend on its actions alone, but on the *sum* of effects of its own actions and disturbances originating elsewhere.

A control system is the only known kind of system that can maintain that *sum* constant, by varying its actions as independent disturbances vary. It is quite correct to say that the actions of a control system exist for no other reason than to

counteract external effects on its perceptions. That is only a description of the logical design of such a system. Those external effects range from the inherent reluctance of nature to drop what we want in our laps to active disturbances that strongly and unpredictably tend to alter what we sense.

A crosswind provides a familiar situation in which the present discussion should make sense: when in doubt, return to the idea of the crosswind. When there is no crosswind, the driver holds the wheel centered and the car is perceived as in the center of its lane (and, we hope, it is). When there is a crosswind from the left, the driver holds the wheel to the left, and perceives the car in the center of its lane, as before. When the crosswind varies, the driver varies the angle of the steering wheel, and continues to perceive the car as centered in its lane. Nearly all of the movements of the steering wheel relate to changes in the disturbance, but the controlled variable is the position of the car in its lane. That is why that position remains so nearly constant, despite the effects of the crosswind. The behavior *cancels* the effects of the crosswind, because the driver senses and reacts to small errors, preventing them from becoming large.

Disturbances as tools

Once this picture of disturbances is clearly understood, it can be used as a way of coming to identify and understand living control systems. This is not a matter of *assuming* that all behavior is control behavior (as SR psychologists *assume* that all actions are responses to stimuli), but of being aware of the possibility, and knowing how to test it.

If a particular kind of action is a control action, then it follows that somewhere is a variable that is being affected equally and oppositely by that action and the sum of all disturbances of that variable. To prove that the action is a control action, one must find the controlled variable, and demonstrate that it is protected by the actions being observed against disturbances of all relevant kinds. To make this proof complete, one must also show that the continuation of the cancelling relationship between action and disturbance depends on the continuing ability of the system to *sense* the state of the pro-

posed variable.

While this procedure seems rather elementary, it can tax one's ingenuity in practice. The control model shows that perceptions are derived from elementary sensory inputs by processes that may be complex. This means that perceptions being controlled may not correspond to anything obvious in the environment. The observer may, in simple situations, simply notice what is being held constant, as in the case of seeing a bird holding its head stationary while its body moves, or any biped standing upright on the deck of a ship. But in other situations the observer may have to learn a new mode of perceiving (or defining variables) before he can see anything being held constant by behavior. This is actually a commonplace procedure; we often wonder what a person "is doing" even though we can see every action; what we are really wondering is what that person is *controlling* with those actions, and to find that out we have to know what disturbances he will resist, and what he is paying attention to. When we casually speak about what a person is doing, we do not refer to the obvious visible motor activities (in most cases). We are trying to say, or guess, which of the many consequences of those activities is being perceived by that person, and what state of those perceptions is intended.

There has been a sort of informal consensus, quite spontaneous and arising in quite independent groups to which I have talked about control theory, about how to speak of controlled variables. I never used this term myself, but many others started talking about "controlling for" things. One eats a lot of candy if he is "controlling for sweetness." Another interacts with people in predictable ways if he is "controlling for status." I rather like this; it's hard to find comfortable ways of speaking about control theory, since it did get formalized by engineers and not social scientists. This way of speaking gets across the idea (if understood correctly) that one's actions are different from what one is controlling for, and that what one is controlling for is an *experience*. The problem in discovering living control systems is exactly that of deducing what experience a person is controlling for through his actions. And the most direct method of testing any hypothesis on this subject is to apply disturbances that ought to affect the thing being con-

trolled for, and seeing whether the person acts to correct or prevent the effects of the disturbance.

Contrast with a traditional method

The mainstay of nearly every experimental approach to human behavior is an experiment in which some few variables are manipulated, and then are correlated with ensuing changes in behavior. This is the approach that the old cause-effect model dictates; using it is simply an affirmation of belief in direct cause and effect. Under this approach, one assumes that the most pertinent manipulated variable is the one that shows the *highest* correlation with changes in behavior.

The procedure we have just looked at, however, leads to exactly the opposite assumption. Clearly, in the case of the driver in the crosswind, the *highest* correlation would be seen between the velocity of the crosswind and the angle of deflection of the steering wheel; that would be a highly significant, almost perfect correlation. But that correlation would not reveal at all what the driver is controlling for. It would only show the expected relationship between a control action and a disturbance. That relationship would hold between the control action and *any* disturbance acting on the same controlled variable; there is no special significance to the fact that on one occasion, it is a crosswind that is doing the disturbing. A low tire would have the same effect. What matters is the controlled variable that is disturbed, not what happens to be causing the disturbance.

In order to discover what the driver is controlling for, we would have to look for a variable that shows an *abnormally low* correlation, with both the crosswind and the angle of the steering wheel. Of course thousands of variables would show a low correlation, but only the controlled variable or a variable closely dependent on it would show an *abnormally* low correlation. What does that mean?

Normally—meaning, in the absence of any active control systems—the lateral position of a rolling car is quite strongly affected by a crosswind. If the steering wheel were fastened in its neutral position, the car would drift rapidly in the direction the crosswind is blowing, and would soon go off the road.

Likewise, in the absence of any crosswind, any significant deviation of the steering wheel angle from the neutral position would cause a rolling car to follow a curve instead of a straight line; again, it would quickly go off the road, in the direction toward which the top of the wheel is turned.

If there were no relationship between variations in steering behavior and variations in the crosswind, we would expect that the lateral movements of the car would show the effects of two uncorrelated influences; the variance of position would be the sum of the variances of the two influences. Since the two influences are pronounced, the expected variance under the hypothesis of *no* control would be *large*.

Suppose, however, that the driver were *perfect*: that his perceptions were perfectly calibrated in terms of the observable position of the car, and that he maintained those perceptions precisely constant. In that case, the variance of the car position would be *zero*. There would be *no* correlation between car position and either of the two influences on it, steering wheel angle and crosswind velocity. That is what is meant by an "abnormally low correlation": a variable that is known to be strongly affected by two influences shows essentially no correlation with either.

The only way in which this failure of correlation can occur (assuming we have eliminated trivial possibilities) is for the two influences to be not just highly correlated, but quantitatively equal and opposite as measured in units of effect on the controlled variable. This can lead to some rather strange situations if these correlations are seen in terms of the old cause-effect model (Powers, 1978).

Thus control theory suggests a completely new approach to the analysis of behavior, and to the design of experiments. While some experiments themselves might not change greatly, control theory shows that once a significant relationship has been found between a manipulated variable and a measure of behavior, there still remains the step of asking *why* there is such a correlation. The answer must be found by discovering a variable that is being affected equally and oppositely by the manipulated variable and the behavior. As this search is narrowed, modifications of the manipulations will suggest themselves, and at the end of the procedure will be found an

essentially *perfect* correlation—and a controlled variable that would never have been revealed under the traditional method, that shows essentially *zero* correlation with the manipulated variable, for good reason.

The significance of control theory for sociology

As someone once said, more or less, I want to make one thing perfectly clear: I am not a sociologist. In trying to provide some leads for those who might actually apply control theory to sociological problems, I will no doubt reveal my lack of expertise, and perhaps urge as new points of view some which are tiresomely familiar. That can't really be helped when one crosses disciplinary boundaries, so I hope that the reader will forgive any blunders. I shall merely try to develop some of the clearer implications of control theory, leaving it to others to devise methods for seeing whether these implications can be turned into something of value. My main aim here is to bring out some differences between at least one control theorist's world view and that which prevails in a number of the life sciences.

Scientific method

Control theory offers an alternative to the common method of manipulating variables and searching for correlated changes in behavior, and for some researchers that may be enough to deal with. That alternative, however, is only one of the implications of control theory that might affect the way sociologists and other life scientists go about their business. Another implication of control theory may be much more important, the test for the controlled variable being merely its justification.

That more important implication, in my opinion, is that we are now justified in restoring the concept of *organism as purposive actor* to all the sciences of behavior. Since behaviorism arrived on the scene three quarters of a century ago, large numbers of psychologists and social scientists have been working, in effect, with one hand tied behind their backs. Faced with obvious instances of goal-directed intentional processes, scientists have been forced to ignore what seems the most

natural way of describing human events and recast their observations in conformity with the old cause-effect concept. Given an example such as that of Beethoven seated at the piano writing a symphony from within his deafness, the behaviorist was forced to explain the symphony as nothing more than a complex response to unnamed and unnamable complex stimuli. This had to be embarrassing, especially considering the degree of creativity attributed to the piano.

The only recourse has been to develop a thick skin and a kind of bravado with the aid of which one can speak of *any* human activity—problem solving, loving, resting, thinking, or even writing symphonies—as consequences of external forces. Often there is evident glee at the shattering of superstitions and ignorant beliefs; you believe that you have intentions, but that belief is really just an epiphenomenon, and how do you like *that*?

Needless to say, it is impossible to maintain a perfect record as a behaviorist; the moment a behaviorist mentions controlling people or other organisms, he is talking about something that his model of behavioral organization cannot do. One cannot even talk about doing a scientific experiment without admitting that *some* organisms can preselect environmental conditions and then bring them about.

I mention embarrassment because I think embarrassment is probably the greatest obstacle to scientific revolutions. Clark McPhail (1977) said "Mead contends that one of the hallmarks of modern science is a readiness to re-examine existing theories when the individual experiences empirical facts discrepant with those theories." (p. 4) That may be true, but it is not my experience that this readiness extends so easily to what one terms a "fact" in the first place. There are some changes of theory which cause one to look back over previous writings and teachings and experience a definite threat of what can only be called embarrassment. It is very uncomfortable to admit a flaw in something to which one has given wholehearted loyalty, expecially if this loyalty has demanded spirited criticism of those who did not share it, and stern insistence that hundreds of students learn the correct view or flunk. Mead's contention expresses an ideal, not a practice.

Of course embarrassment alone could not be the only reason

why a scientist might hesitate to adopt control theory; in fact I think that the proper degree of embarrassment occurs only after one has found compelling reasons to accept the principles of control as viable. Those who get over that hump will do so not *just* because they have resigned themselves to a period of discomfort, but because they are satisfied with the logic of this model. I certainly don't mean the above remarks to be taken as an *ad hominem* argument in favor of this theory. I'm really talking about embarrassment out of a wry sort of sympathy and a remembrance of things past. Despite dispassionate statements of ideals, science remains a most human enterprise.

To get more specific: control theory provides a model that shows how a physical system can specify goals for its own experiences and act to bring them about. It may not, right now, provide exactly the *right* model, but all that is needed is *a* model that can behave in this way without violating any principle of physics. That is enough to remove all objections ever offered by scientists to investigations of human intentions. The same theory, as we have seen, provides a method for testing any hypothetical intention, through ordinary experimental methods that satisfy the usual requirements of public accessibility of data and analysis. Thus control theory may go against certain beliefs about how organisms work, but it complies with basic principles that all scientists think are important.

Beyond the re-introduction of purpose to scientific studies of human nature, control theory provides one more thing that I think is of exceeding importance in the life sciences: a true model, of the same nature as the models on which physics and chemistry depend. I must explain this at some length, because most life scientists are likely to miss my meaning.

A model, not a metaphor

The early behaviorists, who set the tone for scientific studies of organisms even by anti-behaviorists, *thought* they were emulating the methods of physics, but they were not. All they did was adopt a loose metaphor based on Newton's laws of motion. As forces act on masses to make them move, so do stimuli act on organisms to make them behave. In fact, no

physicist even in the 18th Century would have applied such a simplistic view to a *complex* object, especially not to an object that contained its own store of energy and its own motive power. Physics deals with simple homogeneous objects, all complexities existing in the relationships among the objects. Any complex object, composed of many different parts having different properties, has to be analyzed into simple homogeneous components before most laws of physics can be applied. An organism is certainly not a simple homogeneous component.

The early behaviorists missed the central factor in the physical sciences, and perhaps they cannot be blamed: physical scientists did not become clearly aware of it until well into this century. *What holds physics together are its models.* Physics does not progress by generalizing from observations of the behavior of natural phenomena, but through proposing models of underlying forms and structures which, if they existed as proposed, would *require* observable events to take place as they do. Such statistical generalizing as does go on in physics represents only a preliminary fumbling around for an idea, or the final refinements of a measurement. I exclude quantum mechanics here, because I still consider it to be based on magic.

Consider one of Newton's *other* laws, the law of universal gravitational attraction. Newton proposed that every bit of matter of any kind attracted every other bit of matter, with a force inversely proportional to the distance between the bits and proportional to the product of their masses. He proposed, furthermore, that the amount of force on any one bit was the *vector sum* of all such forces calculated between that bit and all other bits of matter in the universe.

We do not, of course, *observe* any such sum of forces. We observe only objects, total forces on restrained objects, or trajectories of freely moving objects. Newton did not cast this law in terms of the typical behavior of objects under stated circumstances. The law does not mention behavior at all. He did not carry out innumerable experiments, throwing objects of different kinds into the air, recording their trajectories, and determining the average trajectory; he did not look at the tendencies of objects to move under particular circumstances (indeed, that is what the pre-Newtonians had done for 2000

years, without getting anywhere, and that is *why* they didn't get anywhere). Newton looked for some underlying structure or form which, if it really existed, would account for *all possible* observations of this sort.

All physical models, from the kinetic theory of heat to the Feynman diagrams of particle physics, are of this nature. They are not statistical averages revealing tendencies to behave under fixed conditions. They are *proposals* concerning something we cannot and never will be able to experience, the entities of the real universe, that lies on the other side of human sensory nerves. No physicist has ever observed an electron, a meson, a quark, and none ever will. What the physicist observes are *perceptual consequences* of the existence and the behavior of these proposed entities. Even the strictest methodologist in physics would hesitate to say that a cosmic ray is really an organized set of water droplets in a cloud chamber. A cosmic ray is proposed to be the *reason* for the water droplets showing up just where they do, and for many other related observable phenomena. It is never, itself, observed. It is part of a model.

The only model that the life sciences have had which is of this kind is the cause-effect model, the SR model. Only this model proposes an underlying form or structure that accounts for all particular observations one might make. Being unrecognized as a model, being qualitative, and being incorrect to boot, it was never made to pass the kinds of tests that physicists demand their models pass: *all possible tests, all of the time.* It could not have passed such tests, if they had been made.

The control system model is also a model of the physical type—not because it is about the entities of physics, but because it proposes an underlying form or structure that demands that we observe what we do observe. It is not a generalization; it does not summarize a mass of observations.

The nearest approach in psychology to a model of this sort, other than the basic SR model, consists of a series of proposals about *intervening variables* and *intrinsic forces* (such as drives). Scientific psychologists have uniformly been adamant in rejecting such proposals, recognizing that they have been not only *ad hoc*, but even worse, *single-purpose*. If human beings show behaviors we would call aggressive, we learn nothing

from proposing that this is a result of a trait called aggressiveness. We are not told what aggressiveness *is* by this proposal; the proposal says only that there is something that has just the properties needed to cause observable aggressiveness, whatever those properties might be. That is the general nature of intervening variable or intrinsic cause models. They aren't really models, because they don't really propose anything, except that some sort of cause exists, appropriate to the observation. If Newton had taken this approach, he would have said that the tendency of objects to move toward each other is caused by move-towardity. That is how theories were made *before* Newton.

What most life sciences call theories or generalizations, a physicist would call *observations*. To say, for example, that reinforcement increases the probability that a preceding response will occur again is to propose that this phenomenon, this relationship, can be observed under stated circumstances. If it can, fine. An observation has been made and verified, given the definitions of terms. To the physicist, and to the control theorist (me), the observation merely exposes the need for a theory: *why* does this relationship hold? To assert that it holds because of the laws of physics and chemistry is to skip everything interesting that we might learn about organization. If you asked why a radio produces more sound when its antenna is extended, a true observation, would you be satisfied with the answer, "Because of the laws of physics and chemistry?" That may be, in the long run, a *true* answer (or it may not), but it is a *useless* answer as well. It is nothing but arm-waving, and would not lead to building any more radios, or fixing one that stopped working.

This science of sociology in which my readers are engaged is long on observation and short on theory. In fact, in the light of what we have just been through, I can propose that sociology does not yet have any theories, unless you count Freudian theories adhered to by some (oddly enough, Freudian theory *is* a theory of the physical type, although unfortunately not a verifiable one since there are no independent tests for the existence of Id, Ego, and Superego). What sociologists offer to each other as theories are really proposals to the effect that certain phenomena and relationships among phenomena

occur. They are *observations*, which lack an explanation.

Maybe I am overstating the case; indeed, I can think of a possible exception, which stands for a fairly large class. Consider the "law of supply and demand." Here is what seems to be a proposal concerning an underlying form or structure that requires that we observe what we do observe (or did, before this law was apparently repealed). I have to admit that this *could* be considered a real theory. But I have a prejudice about such theories, which seems justifiable. The problem with the law of supply and demand, as with many other such global laws, is that there is nobody to administer it. It is a law without mechanisms to embody it. This bothers me, because I pursue the ideal of achieving consistency among *all* our models of reality, not just one little chunk at a time. I can see continuity among the control-system model, direct experience models, the neurological model, and the models of physics and chemistry, with no gaps that cannot in principle be filled. I cannot see any proposals, in economics, for how the law of supply and demand is made to work. It seems to float, along with a host of other such laws, in the air *between* people, and I do not think that air has sufficient complexity to support such a phenomenon.

So I would rather treat such laws as observations, not theories. If a supply-demand relationship is observed, then since it involves people, a proper theory of human organization should be able to say *why* it is observed (and also, when it is *not* observed to hold true, why something else happens instead). I might even guess that human reference levels have something to do with the matter. So perhaps I am not overstating the case after all. Perhaps control theory *is* the first rival to SR theory in sociology, and perhaps there are no other true theories in this field. I am quite curious to know what my readers think of this proposition.

Concluding remarks

If I were really a sociologist, I would at this point come up with some reinterpretations of currently-interesting topics. Unfortunately, I cannot do anything so useful. All I can do is is suggest how sociologists might go about such interesting

tasks.

The most important reinterpretations amount to re-reinterpretations. The world-view of sociology has been strongly influenced by behaviorism; even Mead's was. Behaviorism has devoted much of its time to restating common observations about human behavior and human nature to remove all taint of purposive language, and now all that labor must be undone. This does not mean returning to informal common-sense notions of purpose. The common-sense notion is that a purpose is what is accomplished outside an organism by its behavior. The control-theory view is that a purpose is an intended state of a perception, and that *actions* simply vary, having no purposes of their own. Neither does this mean opening the door to some of the loathsome practices to which behaviorists so rightly (in my view) have objected: wild guesses about entities living in the head, for instance, or introspection that turns into semantic Onanism (although I would not throw out introspection altogether). Control theory demands just as much rigor and public observability as does physics, although it admits to phenomena that were rejected on principle by behaviorism (and without proof).

This re-reinterpreting will not be as hard to do as it might seem, nor will it necessarily spoil as much work as one might fear. For example, wherever one has written about events or relationships which appear to cause changes in behavior or in relationships among people, there is the possibility of recasting the same observations in terms of disturbances and actions which counter those disturbances. The moment that is done, it becomes clear that somewhere there is a controlled variable waiting to be discovered. Its nature might be obvious, or —even better—hidden, requiring a new look at the situation, experimentally. In effect, control theory will often say not "Your study didn't prove anything of importance," but "Look further—you are approaching the paydirt."

Finally, despite my earnest evangelism on behalf of control theory, I am too old and experienced to be disappointed if my new friends in sociology do not rush to board this bandwagon, especially if they jumped off it the first time it came wheezing around, in the 1950s. With time, I think that the logic of control theory will prove compelling. I have not, however, learned

how to compel time, and so can only watch with interest to see how others make their own peace with this idea.

References

Jones, R. W.; *Principles of biological regulation*. New York: Academic Press (1973).

McPhail, C. and Rexroat, C.; The methodological perspectives of George H. Mead and Herbert Blumer: divergence and rapprochement. Department of Sociology, University of Illinois, Urbana-Champaign (1977).

Powers, W. T.; *Behavior: The control of perception*. Chicago: Aldine (1973).

—; Quantitative analysis of purposive systems. *Psych. Rev. 85*, No. 5, 1978, 417-437.

[1980]

CT Psychology and Social Organizations

This paper was written as theoretical background for D. T. Campbell's investigations of the uses of social indicators in judging social programs. The orientation is that of "CT psychology," CT standing for control theory; familiarity with the basics of this approach is assumed, and the theoretical underpinnings of CT psychology will be used here without justification.

The following consists of a systematic development of a number of themes without regard to their obviousness (or for that matter, their factual correctness). The object is to see what kinds of theoretical statements applying to social interactions appear to follow from a belief that all organisms, human beings especially, are organized as hierarchies of negative feedback control systems. In lieu of a review of the principles of CT psychology, we will begin with some statements that follow deductively from these principles and will be important to later arguments.

Some generalizations from the principles of CT psychology

Human beings are assumed to have certain fundamental properties:

1. The world with which human beings knowingly interact consists of a hierarchy of representations in the human mind (for which read "brain" or any other acceptable substitute). These representations consist of information-carrying signals

Working paper, March 1980, distributed by W. T. Powers and D. T. Campbell. Supported in part by NSF Grant BNS 7925577.

called, for short, *perceptions*. Perceptions of the lowest level are signals that depend directly on sensory stimulation from outside the organism, creating real-time analogues of stimuli in the dimension of intensity alone. Higher-level perceptions are signals computed continually or iteratively on the basis of sets of lower-level perceptions; their behavior depends in complex ways on the behavior of sets of stimuli, and they represent *abstractions from* or *aspects of* the world of intensity signals. Since we have, however, no other direct access to the external world, these perceptions are experienced as the world which exists and is acted upon outside us. Successive levels of rerepresentation can be identified approximately as sensations, configurations, transitions, sequences or events, relationships, categories, programs, principles, and system concepts, all of which types of representation we project to some degree into an external world. Whatever effects human beings have on the outside reality in the course of controlling perceptions remain unknowable by direct means. Since all control actions take place, ultimately, through output effects on the external world, constraints due to natural laws are reflected in the organization of the control systems that are acquired, these constraints providing the raw material for theories about the nature of the outside reality. All CT analyses of human actions, however, must explicitly take into account that each individual controls perceptions, not realities.

2. The appearance that human beings react to stimuli or situations is created by the universal disturbance-opposing property of control systems. Each control organization (and hundreds may be simultaneously active in a given person) controls one perceptual variable, which is a function of external events even if not a copy of them. A disturbance is defined as any independent external variable capable of influencing the state of a controlled perception. When such disturbances are present, they tend to alter the state of one or more controlled perceptions away from their *reference states*, states selected by means of reference signals from higher-level systems. The resulting *error*, which depends both on the current state of the controlled perception and on the given reference or goal-state, drives behavior. In a properly-organized control system, the behavior is always such as to diminish or limit

the error, keeping it small. This error-limiting action is seen from outside the organism as a "reaction" to the disturbance as if the disturbance were the cause, the action the effect, and the organism a mediating transfer function. This is the traditional view. Under it, if different events are followed or accompanied by the same behavior, it is assumed that the different events share some property in common to account for the common effect on behavior. In CT psychology no such common factor is assumed: all that such events have in common is the fact that they disturb the same perception in the organism. External events, in CT psychology, have no psychological properties such as "stimulus value."

3. Distinct levels of control exist, corresponding to distinct levels of perceptual representation of reality. Many independent control systems exist at each level; they are coordinated only by receiving reference signals from many higher-level systems. They act, in turn, by sensing contributions to the reference settings for many lower-level systems (only first-level systems act by producing physical forces via the muscles). The overall arrangement is hierarchical but it is not a single "tree." Instead, at each level there may be a large number of control systems each of which is at the peak of a hierarchical tree, and many or all of which act by contributing to the net reference state of many or all systems of the next lower level. One system may thus be a member of many hierarchies. As a result, the organism must adopt control systems at a given level which can operate independently, a design problem that reduces to the solution of a set of simultaneous equations. Each higher-level controlled perception must be an independent function of the set of lower-level perceptions, and the number of simultaneously-active higher-level systems can't exceed the number of independently adjustable lower-level reference inputs. To put the same statement differently, the number of higher-level goals that the organism can adjust independently of each other may not exceed the available degrees of freedom of the lower-level set of systems, regardless of the kind of perceptual transformations linking one level of perception to the next. This principle proves to be of fundamental importance in understanding social systems in which each member is a hierarchy of control systems.

4. All control systems *control* their inputs by *varying* their outputs. The purpose of any action is defined as the reference setting for the perception which represents the inputs and is the actual controlled variable. Intentions or goals must therefore be seen not as intended actions, or as intended states of the physical outside world, but as intended states of *perceptions*. In a hierarchical structure, the means of achieving a given intended state of a perception is to vary the intended or goal states of lower-level perceptions until the derived perception matches its internally-specified goal state. Lower-level goals are adjusted in the service of higher-level goals. Since many higher-level systems use a given set of lower-level systems at the same time, however, the specific action taken by a given higher-level system will depend both on external disturbances that may be acting and on the actions being taken by other systems of the same level. This apparently complex requirement is met automatically by a proper control-system organization (a very similar arrangement of feedback systems is commonly used in analogue computing for the apparently magical solution of sets of simultaneous algebraic, differential, or integral equations). A given higher-level system does not *determine* the goal-states for lower-level perceptions; it can only add its own influence to that from other systems of similar level, tending to increase or decrease or otherwise alter the goal-settings of many lower-level systems at once. By continually raising or lowering its contribution to the reference settings of many lower-level systems while continually or repetitively sampling its own error in an ongoing process of feedback control, it can control its own perception without being able to specify a unique state for any single lower-level goal. Thus a set of lower-level goals may show adjustments that have no obvious relationship to any one higher-level goal, while the complete set of lower-level goals contributes simultaneously to the achievement of all higher-level goals. This principle has direct applications to the design and analysis of social systems.

5. Finally, the human hierarchy comes to be organized during a process of development and learning. There is some set of fundamental requirements of the human organism which is at the focus of this process. These requirements act like refer-

ence settings for crucial variables inside the organism, and specify basic and unalterable intentions; intentions which are achieved by eating, drinking, breathing, procreating, keeping warm, avoiding illness and injury, and perhaps, for human beings, such activities as intellectual or aesthetic creativity, problem-solving, and the seeking of novelty. In line with the general tenets of CT psychology we can't say that these behaviors themselves are inherited, but only that the need to control whatever they affect is inherited.

The unknown set of variables that heredity demands be kept under control is called, collectively, the intrinsic state of the organism; there must be a corresponding set of intrinsic reference levels that specify the particular state demanded of each variable. When intrinsic error occurs, reorganization commences; the process of adding to or altering the organization of the hierarchy of control systems that operates through motor behavior. The aim of reorganization is not to achieve any particular kind of control over any particular selection of perceptions, but simply to minimize intrinsic error. The set of intrinsic reference signals, therefore, represents the highest goals of the organism, and is the ultimate explanation for the settings of the highest level of reference-settings in the acquired hierarchy of control systems.

This overall picture leads to the conclusion that the goals of an organism are always, in the final analysis, adjusted for the benefit of the organism and for no other reason. External events can disturb the perceptions in a human being's hierarchy, thus calling for adjustments of goals lower than the level at which the disturbance is first effective, but external events have no direct effect on a person's internal goals; external events can affect only the perceptions being compared against those goals. The reaction of any organism to disturbance of one of its perceptions is to make adjustments in many lower-level goals, adjustments which take into account *all* the higher-level goals, not just the one pertaining to the disturbed perception. The result of these adjustments is to minimize or cancel the effects of the disturbance on the variables with which the organism is most concerned. An organism can be made to *behave* in specified ways if its entire goal structure is understood and if carefully selected disturbances are applied,

but its whole goal structure and its ultimate intentions make it basically autonomous. Non-living systems bend before a disturbance, but living systems push back.

Degrees of freedom, conflict, and competition

When two individuals share an environment, each acts to control some set of perceptions by acting on that shared environment. It is likely that each person's actions will constitute a disturbance of some of the other's perceptions. Most often this interaction will have no noticeable consequences; each person will adjust various details of action to oppose the disturbance caused by the other, and both will succeed in controlling whatever perceptions are involved. Control systems have the ability to continue to act as if they were independent even when there is fairly strong coupling between them.

A direct conflict can arise if (a) two perceptions derived by "linearly dependent" functions of identical external variables are being independently controlled, or (b) the same perception derived from the same variables is being controlled relative to a different reference value in each person. In either case, successful control by both systems at once requires some physical situation to be in two different states at the same time. This being apparently impossible, neither system can eliminate the difference between perception and reference value: the error each detects and tries to eliminate.

The efficiency of a control system is measured by the amount of effort it will create for a given amount of error. The greater this amount of effort is, the smaller the error required to bring the effect of action into balance with the effect of disturbances on perception, and thus the more precise control will be. As a result, the deleterious consequences of unresolved conflict become worse in direct proportion to the efficiency of the control systems that are in conflict, and of course in proportion also to the disparity between the goals of the conflicting systems. If two control systems are precise enough to keep their perceptions matching their reference settings within one percent over all disturbances within their physical power to oppose, a disparity of goals of only one percent will drive both systems to pit their maximum possible outputs against each other.

Whichever control system is the weaker will have all of its possible output cancelled by the opposition of the other system. The other system will have only the difference between its maximum possible output and the maximum possible output of the first system to use in counteracting disturbances that aid the first system. Both control systems will therefore become much less effective in controlling their respective perceptions. Many detailed situations can be analyzed in this way, and they all arrive at the same result: unresolved conflict destroys some or all of the ability of the conflicting systems to exert control.

Direct two-person conflicts, outside games such as arm-wrestling, seldom last very long because they are easy to resolve. The more important perceptions to control are the higher-level ones, which are functions of many external variables. Normally a direct conflict is accidental and momentary, and is resolved simply by a change to alternative means of accomplishing the same higher-order goal. Analysis of a sustained direct conflict, however, shows clearly why conflict must be avoided by persons constituted as hierarchies of efficient control systems: some part of the skills of both persons is lost for the duration of the conflict. If we consider, furthermore, that control systems are acquired as an organism's means of maintaining its physiological and organizational status in many critical regards, loss of an acquired control ability is dangerous to the organism, control being its principal means of surviving.

A direct two-person conflict is a special case, the simplest, of a far more general situation, the problem of degrees of freedom. Direct two-way conflict involves just one variable with one degree of freedom; conflict arises because the two systems try to specify two values for the variable at one time. In the more general case a number of variables is involved, and a number of control systems whose goals each imply a different pattern of these variables. If the perceptions being controlled are independent functions of the variables, there is no problem, because the interactions of the control systems as each counteracts the disturbances due to the control actions of the others will lead (or can lead) to just that distribution of output actions by each system that will solve the simultaneous

equations. Consider a situation involving two variables, with one control system controlling their sum and the other controlling their difference. No matter what reference level each system sets for its own perceptual representation, sum or difference, both systems can act to bring the sensed error to zero. But if a third control system enters the picture, sensing and controlling, for example, the product of the two variables, there is no way that all three of the systems could achieve the state of zero error. Two variables define only two degrees of freedom; it is not possible to specify the values of three functions of those variables *independently*.

The general degrees-of-freedom problem is that in which n control systems attempt to control n functions of m variables, and n is greater than m. A number of people share an environment, so many that there are more variables under independent control than there are available degrees of freedom in the environment. When this situation occurs, the result can only be some loss of control by all persons involved. The severity of the loss of control is determined by the excess of goals over degrees of freedom, and by any failure of the individuals to control truly independent aspects of the same environment.

The lower organisms in general have solved the degrees-of-freedom problem (to some extent) by finding *niches*. Organisms that employ differing means to obtain differing necessities may disturb each other's controlled perceptions by their actions, but a solution to the simultaneous equations exists and they find it; that is probably what "niche" means. Within a niche, however, the degrees-of-freedom problem is solved in the most direct way possible. The number of organisms increases until each new organism comes into conflict with all the rest sufficiently to prevent further increases in numbers.

That is, indeed, the way in which the degrees-of-freedom problem makes itself known. If all the degrees of freedom have just been used up, and a new organism arrives intent on controlling—independently—some variable that is an arbitrary representation of the common environment, its control actions will tend to cause slight disturbances in the perceptions of all the other systems. Each of the other systems will resist slightly, and the sum of all those resistances will add up to massive opposition to the control actions of the new system. If the

new system succeeds in gaining a foothold, all the systems will begin to experience increased error. The set of simultaneous equations now can only approach, not reach, the solution that would permit each system to have independent control of its own perceptions.

It can be seen that there is a quantitative aspect to this problem. In a large collection of organisms in an environment with many degrees of freedom, a small excess of independent goals over the maximum mathematically permitted will result in only a small increase in overall error. The greater the excess, the more error there will be, and the greater will be the proportion of efforts that are exerted only to cancel the efforts of other control systems. As the excess grows, the stress in the whole system will grow, even though all systems can retain a reasonable degree of control. Control will deteriorate with each new excess goal that is added, but the whole assemblage can continue at some level of skillful control for some time yet. The assemblage becomes theoretically overloaded the moment the first excess goal is brought in, but it is in serious danger only when the excess has become much larger.

The quantitative aspect extends in the other direction, too, for different reasons. We are speaking here of real organisms, not mathematicians or computers. While there is a strong tendency for a collection of interacting control systems to seek the minimum-effort configuration and thus a solution to the simultaneous equations, this seeking is for the most part carried out by living organisms in a trial and error way, involving repeated reorganization, or "blind variation and selective retention." This means that the actual solution to a set of thousands of equations is not likely to be found or even approached. In order for a collection of organisms to share a common environment while avoiding the penalties of mutually-cancelling efforts, the environment must be abundant in degrees of freedom, more so than mathematics would imply. The collection of organisms will begin to lose control when finding an independent variable to control becomes somewhat difficult, long before it becomes entirely impossible. The practical limit on independently-seekable goals must be far lower than the mathematical limit.

The necessity for finding independent goals is probably

more important at intermediate levels of individual organization than at the highest levels, for the reason that intermediate goals are *means* relative to higher levels. Higher-level goals pertain more directly to states within the organism itself, so there is less interaction between organisms at those levels. Intermediate goals, however, represent ways in which the organism tries to achieve those higher goals, by selecting intended effects on the outside world. There is nothing in principle to prevent everyone's achieving a degree of self-respect, a high-level perception, but the more standardized the accepted means for achieving it become, the more likely a degrees-of-freedom problem becomes. In general, the more structured interactions become, the fewer independent goals can be chosen. If the *only* accepted way to achieve self-respect were to drive an automobile, then the maximum number of people who could achieve self-respect at the same time would be limited to the number of automobiles available to be driven at the same time.

Presented with a degrees-of-freedom problem for the first time, members of a naive group of people will most probably react to disturbances from others simply by trying harder to achieve their individual goals. If such a group could be called a "society" at all, it would be a purely competitive society, the survivors being those whose physical resources are greatest, the rest being overwhelmed. Each individual would head directly toward his or her own goals, exerting whatever amount of effort is required to overcome error. Since the degrees of freedom are effectively exhausted, what one person gains would necessarily be at someone else's expense. The competitive society is a society in which social intelligence has not yet developed: it is merely a collection of independent entities interfering with each other.

In the following sections, I will propose some stages through which it is plausible that social intelligence grows.

(Note: in this discussion any reference signal at any level may be called a "goal." A goal is a specification for the state of some perception at some level.)

Adaptation and the alignment of goals

When human beings live in proximity, especially when necessities are scarce, certain kinds of goals probably exceed the available degrees of freedom rather easily. Since prolonged degradation of control harms all individuals involved, some solution to the degrees-of-freedom problem has to be found. The simplest solution is to cut down the number of independent goals by sharing them.

An example: Suppose each person in a community independently selects goals for how to obtain enough food—what animals to hunt and when and how, what grains to plant and when and where, what means of preparation to use and with what utensils over what fire. The community would not have to be very large for this amount of freedom of goal-setting to exceed the numbers of animals available, the number of arable fields, the available weapons, and the number of utensils and fireplaces. Each individual would have to spend most of the time stealing what he needed from a neighbor, scaring animals away from other hunters, uprooting someone else's grain to plant some he likes better, and guarding his cooking pot. Everyone would be trying to cancel everyone else's efforts, and all would risk starvation.

The only solution, other than scattering the tribe or getting rid of some people (and both have undoubtedly been tried), is to limit the freedom to choose independent goals. This is done by arriving at common goals.

This result does not necessarily have to be arrived at by communication. When there is a degrees-of-freedom problem, goals which caused the maximum number to be exceeded will prove difficult to achieve, since the other individuals in the system will all experience disturbances and will resist their effects. "Difficulty" implies that seeking one personal goal violates another—for example, constantly wrangling with neighbors over a few ears of corn might violate a reference level for maximum acceptable effort. To achieve control over one variable entails loss of control over another, so internal conflict is experienced. The person experiencing it will attempt to resolve it through finding a more suitable combination of means and ends at lower levels, and in this case the only such

combination available is one that involves sharing of goals. The alignment of goals comes about as a natural consequence of each person's seeking to minimize all errors and resolve internal conflicts when no more degrees of freedom are easily available.

Thus even before people were able to say, "Let's go hunting together," they were probably heading in groups in the same general direction, throwing their spears at the same time at the same animal, and resolving the simple conflict over the result by tearing it into independently manipulable pieces.

Adaptation, as the term is used here, implies only that each individual makes adjustments in choices of lower-order goals so as to make his or her own control processes work better. Other people are simply aspects of the environment, and each adapting person operates essentially independently. The alignment of goals is not formally planned, but comes about as a side-effect of each system's seeking a state of minimum internal error. Adaptation to the properties of the natural environment probably follows for similar reasons, but our concern here is with social adaptations, which are forced only when there is a problem with degrees of freedom. Adaptation is the first step toward the appearance of social intelligence.

Specialization: The next step

Adaptation will result in the sharing of goals, but the sharing of goals is not as easy as it may sound. One can imagine two blacksmiths sharing the goal of striking a hot piece of metal with the same hammer, but even though their combined strength would be improved, their combined aim would be much worse. Control theory would predict, by the following reasoning, that a task requiring precision of control will eventually become the responsibility of a single controller. The more precise a control system, the more sensitive it is to small errors, and the more effort it will exert for a given amount of error. Error is the difference between what is being perceived and the particular state of that perception selected as the reference level. In order for two control systems of high quality to share the same goal successfully, they would have to have closely matched reference signals and perceptual representa-

tions, a most unlikely state of affairs in living systems. If matching is not precise, the control systems will come into conflict, for what one system considers a state of zero error will appear to the other as error requiring energetic corrective action ("social parallax"). The only alternative to conflict is to reduce the quality of control, so that the sharing control systems respond less to a given error, allowing greater differences between the systems without calling up large mutually-cancelling efforts. This alternative, of course, entails some, and often a significant, loss of control. Sharing of a goal is thus practical only when the goal is unimportant (i.e., the corresponding perception does not require precise control).

By definition, a skill is the ability to control something precisely (or something complex at all). Conflict and precise (or complex) control being incompatible, we would expect that in any long-term grouping of people, specialists would appear who carry out their work one person to a task. The degrees of freedom are partitioned into essentially independent subsets, so instead of everyone participating to some degree in every major activity, we would expect the most skillful people to take over particular tasks and not participate at all in others.

Within a given specialty, there may be more than one specialist. The only way for this to work, however, is to partition or duplicate the task itself, so that we still have one skilled task per person. If there are several flint-flakers, each works on a different arrowhead or axhead; they do not all work on the same piece of flint.

Specialization, according to these deductions from control theory, arises when skilled control arises, because the only alternative is conflict. The specialist does not become one because of any predisposition to specialization, but because the only way to carry out a control task demanding precision is alone.

Cooperation: The first social variable

An elementary example of cooperation is standing aside to let someone else pass through a door, and to avoid a collision that would also prevent one's own passage. By deliberately controlling a relationship between oneself and another, conflict

is prevented and a higher-level goal is achieved.

To cooperate means to operate together. Sometimes it means taking turns (controlling a sequence of which one's own act is an element), sometimes it means adding one's efforts to those of another, and sometimes it means contributing an action that completes a pattern made of the actions of many people. Whatever adjustments are made, their immediate purpose is to create a social unit, a two or more person entity that acts, for the moment, as one. More practically, the result is to create at least a temporary social unit that can act without dissipating its efforts in internal conflict.

In order to cooperate, a person must deliberately choose cooperation as a goal, which means, of course, that the person must be able to perceive in terms of degree of cooperation. To cooperate excessively is to abandon other important goals; to cooperate insufficiently generally leads to the same result. Cooperation is a perceptual variable that can be controlled by altering one's own actions in relationship to those of others. One's reference level for cooperation specifies just how far one is willing to go.

For members of a group to cooperate, they must perceive the relationship between a goal they are each trying to maintain and the degree of organization of their various efforts. This is why I call cooperation the first social concept: before it can happen, the members of the group must develop a perception that is specifically about the group, rather than about the results of its actions. Adaptation and specialization may result in such awareness, but do not depend on it. Cooperation depends on perception of cooperation, a variable that does not exist for one person alone.

Cooperation is not easy, for the same reason that sharing any goal is not easy. Cooperation requires each person adjusting his actions to make a pattern with the actions of others; pulling on a rope in a common rhythm, picking up the other end of a log too heavy for one person to carry, serving as lookout while others stalk, cooking the food that others brought home and eating last. Rather than treating disturbances from others as obstacles to be overcome by increased effort, one must see conflict itself as a mistake, and make reasonable adjustments and postponements to lessen it. Cooperation

requires higher levels of perception than do adaptation and specialization. It requires social intelligence, since it demands analysis of tasks into parts that can be carried out without conflict, and understanding of the functions that others play in the overall task. The most complex kinds of cooperation require constant communication, to inform others of one's own intentions and to find out what they are going to do.

There will always be conflict between those who have achieved some stage of social intelligence and those who have not. Consider a group that is pursuing some common goal such as laying up food for the winter. Those in the group who do not share the goal of cooperation act as disturbances to the rest of the group, since their independent actions (bringing food to the village but running home with it) reduce the degree of cooperation perceived by the others, even if the others have no immediate need of the food. Those who are cooperating will each resist this disturbance, resulting in a conflict between the cooperating group and the non-cooperating others. Since the others who are not cooperating are not cooperating with each other, either, they will find themselves individually in conflict with an apparently organized group and a scattering of other individuals. The only way to escape this conflict, which a single individual cannot win, is to separate from the group (or become part of it by adopting cooperation as a personal goal). Thus when cooperation appears in a society, so do outcasts and criminals. The conflict between those who cooperate and those who don't is inevitable.

Cooperation is a social perception that must exist in all those who cooperate. It is a shared goal, sharable if it is not sharply and quantitatively defined (when it is too sharply and quantitatively defined, as in some fundamentalist religious groups, conflict eventually disrupts the group). It appears because it is a way of dealing with limitations on degrees of freedom, but also because it creates power to control that is greater than that of any individual. It is usually temporary because to work at all it must be loosely defined. Cooperating individuals make their own judgements about what adjustments are appropriate and needed. Nobody is in charge, because cooperation is an internal attitude.

Coordination: The appearance of a social hierarchy

I am defining coordination, to distinguish it from cooperation, as the control by one person of a set of tasks being carried out by others. The coordinator must, somehow, adjust the reference signals in each "worker" (as we will call those whose actions become coordinated) so that when each worker acts to bring his own perception to his given reference level, a higher-level variable in the perceptions of the coordinator is thereby maintained under control. As disturbances appear, the coordinator alters the goal-perception being maintained by each worker in the direction that will oppose the effects of the disturbances on the net result. The coordinator is thus acting as a control system, using the workers as part of his output equipment.

This familiar result cannot be carried out literally as just described. Human beings have sensory inputs on which their hierarchies of perception are built, but there are no reference inputs from the outside world. All reference signals in a worker have to be generated by his own higher-level systems. A person who is to act as a coordinator must accomplish his task in an indirect way.

The first and most obvious requirement is that the workers must be willing and able to set their own reference signals according to instructions from the coordinator. This implies a different organization from that needed for cooperation, in which each person generates his own reference signals independently, according to an independently conceived idea of what is needed. Coordination demands that the workers make a commitment to the principle of following orders, at least within a certain range of orders and in a particular context. The mechanisms that maintain this principle in effect can be analyzed in some detail.

The crucial question is, how can one person transmit a reference signal to another? This cannot literally be done, but what could have the same effect?

A step toward understanding how reference levels are transmitted can be seen in the following printed instructions. "Put tab A in slot B" has no meaning unless one can read English. Even if one can read English, one must be able to translate

"tab" and "slot" into appropriate visual images, so that one can *find* the tab and slot in the first place, not as words but as visual configurations. Then one must understand "put... in" as the establishment of a certain set of transitions leading to a particular visual relationship; the sequence in which A and B occur in the sentence has to be translated into that visual relationship rather than other possible relationships. Only after all the necessary translations into non-verbal perceptions have been made is it possible to commence an action that will produce a result. The outcome, of course, is a set of perceptions that match the reference-images one has made from the verbal communication, and it may or may not produce the overall result that was expected.

Broadly, the coordinator must use some form of communication to convey the reference signal to the worker. But whether this communication be in words, through gestures, or in the form of direct demonstration, it must enter the worker's hierarchy through his lowest-level systems as first-level perceptions of intensity. At each level, a perception has potential *meanings*, which in CT psychology are taken to be stored perceptions evoked through the associative property of memory (taken as a given). The meanings of a perception are other perceptions of the same level, drawn from past experience. Words are recognized at many levels in the hierarchy. At each level they can evoke corresponding levels of perceptions from memory. The evoked perceptions in which we are interested here are non-verbal.

A simple spoken order such as "hammer that nail all the way in" enters the worker's hierarchy as sound-intensities. These are interpreted as phonemes, then as words, then as sequences, then as relationships among sequences (word-order being one relationship of importance). Within the verbal hierarchy, however, the words remain linguistic perceptions of various levels, having nothing in common with nails or hammering. A nail is a visual configuration which is *named* "nail"; hammering is a visual-kinesthetic sequence and a visual relationship *called* "hammering." It is clearly up to the worker to translate the coordinator's spoken order out of the realm of sound and into the realms of vision and effort, for the only reference signals that will produce hammering are those relat-

ing to the perceptions actually involved in that task. The perceptual operation involved must be something like finding the non-verbal perceptions which are equivalent, at a given level, to the verbal ones. The non-verbal perceptions found in memory then can become reference signals for guiding specific control actions. The basic principle being invoked here is that the reference signal must be of the same kind as the perception to be controlled.

The reference signal that becomes effective in the worker thus depends on the way the worker translates a communication into perceptual goals. The coordinator can only transmit what seems in his own perceptions to convey the wanted reference image; the rest is up to the worker. Since the coordinator has no way to see either the reference image or the private perception being compared with it in the worker's brain, the only way to judge the success of the communication is to observe what the worker does. In an important sense, therefore, the coordinator can never know just how a worker has translated his orders into perceptual images, because the coordinator must always interpret the results through his own hierarchy of perceptions, not the worker's.

This is not a serious limitation—no more serious, that is, than our basic ignorance of what we actually do to the outside world when we act on it to alter our own experiences. The coordinator can adjust his communications, see the effect on the final result, and arrive at effective communications without ever having to see inside any worker's head. The communications may work for reasons entirely different from those the coordinator may be postulating, but that is of little practical importance.

A person acting as a coordinator is controlling a perception of a higher level than the perceptions under control by a worker. The worker, of course, has just as many levels of perception as the coordinator has, but by agreeing to translate the coordinator's communications into reference signals at some level, the worker has provided a way for a higher level in the coordinator to operate through lower-level control systems in the worker. This is the effect; in reality, all the worker's actions are under his own control, the most important ones being the active translation of words into reference signals, and higher-

level goal, such as that of keeping a salary coming in, that requires doing this.

Coordination may come about for the same reason that specialization comes about; indeed, it may be classified as a form of specialization, in terms of the level of perception being controlled rather than the task at a given level. Cooperation has its limits, since every cooperating person is trying to control a variable, among others, at a level higher than the particular individual task at hand. If a group is undertaking a task that requires a particular way of acting together, a particular relationship among individual sequences of actions, each person trying to cooperate is operating on the basis of his own conception of that relationship. Despite all good intentions, each person is going to have a slightly different (or greatly different!) image of that relationship, and will see the error in a different way. The potential for conflict is great.

This is where the coordinator comes in. If a particular relationship among individual tasks has to be maintained, and if there are not very many equally effective arrangements of individual actions, the potential for conflict and its attendant likelihood of failure would be greatly reduced if a single person took over the task of coordination. Then there would be just one perception of the required relationship, and one reference image relative to which it is being controlled. The coordinator becomes a specialist at the relationship level, specifying sequences which the other individuals are to control. If the result is reasonably in accord with their individual conceptions of what was intended, the others will continue to agree to let the coordinator set their reference levels—by the indirect means already discussed. (For obvious reasons, there is likely to be only one coordinator for a given group task.)

The coordinator might or might not also be a worker. It is not, however, the role as a worker that distinguishes a coordinator, because anyone willing to carry out the orders that the coordinator, in effect, gives himself would be a suitable substitute. What makes the coordinator different from the workers is the fact that the perception he controls is a perception of the relationship of the results being accomplished by many others, a variable that is not apparent to any one worker who is specializing in a particular task. To control coordination one must

perceive coordination, and to perceive coordination one must observe all the tasks in relationship to the pattern they are supposed to make. This is not merely a matter of monitoring the final product, such as an output of an assembly line. That alone would not reveal how to alter orders given to workers in the way that would improve the final product. It is essential to perceive *coordination itself* as a variable, because that is what has to be controlled in order to have any coherent effect on anything else being done.

A coordinator, therefore, must inevitably occupy a special place in a group, because the value of his work lies in the abstract variable he is controlling, not in his skill as a worker-specialist. The first appearance of a coordinator in a group must be taken as the first appearance of a hierarchy in the group.

Management: Expansion of the social hierarchy

I have tried to define coordination as the control of relationships among the sequences of actions carried out by workers. It is possible to conceive of specialists at higher levels of perception, who would act by specifying reference levels for coordinators, and so on up a hierarchy corresponding to the levels of human perception assumed in CT psychology. For example, a specialist at the category level would be a person who classifies tasks according to some convention: now it is time to plant, now time to harvest, and now time to make sure the Sun doesn't sink any further into the southern skies. The classifier would tell the coordinators what to coordinate. A specialist at the program level might be a general who plots strategies, a scientist who directs a research operation, a governor who sets up a program to levy taxes. At the principle level we might find policy-makers, long-range planners, priests and preachers. At the system concept level we would have law-makers, boards of directors, and in general those who create and maintain the entities we know as social organizations.

It is highly likely that in social systems we would find specialists at levels corresponding to any perceptual levels that may be shared among all human beings. Beginning with coordinators, who specialize in variables that have meaning only

for a group, we can see how a whole branch of specialists would develop, dealing in higher and higher levels of perception all built on perceptions of relationships among people—social perceptions. As soon as simple coordination is discovered, a hierarchy of management seems to follow as a natural result. The distinction between manager and worker is not one of intelligence or complexity, but of *the kind of perceptual variable that is controlled*. And for the management hierarchy even to exist, the active functioning of workers is essential, not merely in doing their jobs but in interpreting symbolic instructions as appropriate perceptual reference levels, and even more important in agreeing to do this in service of their own higher-level goals. That agreement is the *sine qua non* of management.

The highest level of specialization

Human societies have experimented with many varieties of social hierarchies. Some like the Third Reich have attempted to invest the highest level of management in a single person; others have tempered this approach with advisers or formal governing groups; still others, for brief periods, have tried to do away with the vicarious setting of goals at the strategy, principle, or system concept level by any formal group. The variety has been impressive, but all social systems share two obvious characteristics: they have all been limited in size, and they have all broken up (except the ones currently in existence).

The upper limit of size on a management hierarchy is at most the size of a country; there are no larger groups that accept goals from a higher level. The fact that we have so many sovereign nations indicates that no system concept has so far proven acceptable to all human beings. But the actual upper limit in size of a social hierarchy is probably much smaller than a nation. It is probably more on the order of one division of General Motors. Social organizations of a size larger than that are largely illusions; they are not organizations, but groups of organizations that expend major parts of their available effort on mutual conflict. They are constantly on the verge of breaking up; they are composed of shifting

factions; they labor mightily even to have a small temporary effect. At best, social entities as large as or larger than General Motors function at the level of competition, and perhaps occasionally adaptation. There is no human society as a whole; there are many societies, each committed to some system concept. If we take each organized hierarchy that does exist as being one individual, it is clear that this hyper-society exists in a state of severe internal conflict.

Why is this the case? Why have human beings been unable to agree on a system concept under which they all can live peaceably? I suspect that the answer is unpleasant, for the reason that we can do little about it now: system concepts may not be a high enough level of perception to do the job. The creation and preservation of social entities such as governments, corporations, teams, religions, cultures, and the like certainly can bind people together, but not very many at a time under one system concept. Furthermore, since an individual can pledge allegiance to more than one social hierarchy, and since the system concepts and substructures of these hierarchies are usually in conflict with each other in small and large ways, most people find difficulty in avoiding internal conflict, because of the conflicting requirements of the various groups to which they belong.

If conflict among social entities is to be resolved, we clearly need a level of perception from which social entities can be seen as means rather than as ends. This is unlikely to be a matter of finding a better system concept; more likely, it is a matter of seeing social entities themselves from a new point of view, as conveniences to be manipulated as necessary in order to achieve—what? Vague terms like "the dignity of Man" don't seem to get us far, but perhaps they point in the right direction.

CT psychology and the mechanics of management

We turn now to an exploration of what CT psychology has to say about the way management hierarchies work. In such a hierarchy, an individual at any given level is placed relative to the others in terms of whose orders are accepted by the manager and translated into reference perceptions, and who

will in turn accept orders (or instructions, suggestions, directives, guidance... reference levels). In a large organization there will generally be more than one manager at a given level (that is why higher levels arise), and within a given level there is therefore the need for adaptation, specialization, and cooperation in order to avoid conflict. There are also, of course, conflict and competition to a degree that depends on the available degrees of freedom and the success of conflict-avoiding reorganizations.

It is very unlikely that a manager at any level would accept reference levels at just one level, or transmit them to just one lower level. To work in that way the manager would have to be essentially ignorant of all higher-level and lower-level aspects of his organization, and would have little motivation for being part of it. It is impossible, furthermore, for a manager to anticipate all the detailed goals that ought to be spelled out in order to affect the actions of those at lower levels in the direction that is needed; those at lower levels must understand not only their own immediate tasks, but the reasons for which they are done, which is to say the goals that the manager has accepted from higher levels. Having just as much basic equipment as the higher-level manager at all human levels of organization, the lower-level managers are capable of judging, for example, what orders would have been given in the face of some higher-level kind of disturbance. Anticipation of orders can, naturally, lead to mistakes, but failure to anticipate orders can be even more disastrous.

Establishing a smoothly functioning management hierarchy is therefore not a matter of issuing clearcut directives and sticking strictly to one's specialized level. It is a matter of all the managers at all the levels sharing as much as possible the same understanding of what the whole system is, what its principles are, how its methods and strategies are structured, what categories of control exist, and how the different actions taken within the organization relate to and affect each other. In this sense all the managers are on equal footing, simply because they are all human.

The formal structure of the organization is defined by the control tasks within it and how they are related. Each manager agrees to occupy a specific place in this structure, which means

accepting reference levels of certain types, and carrying them out by specifying reference levels for others. This invisible network exists only because of these agreements; it is not embodied in any one place. The structure is not real in the sense that it could be associated with a single physical entity; it exists in the imaginations of those who share the same system concept. If all the managers of General Motors ceased to accept the existence of General Motors, General Motors would cease to exist.

The hierarchy in a social organization has some important resemblances to the organization of an individual, but a social organization is not a person. It is a consequence of some group of individuals adopting similar system concepts and accepting and giving directives accordingly. The hierarchy is functional only, not factual; no person in the hierarchy is superior or inferior to any other (or to put that differently, comparative skill or intelligence or any other such factor is not what makes the difference of level in a management hierarchy: *function* is what makes the difference. There can be wise flunkies and stupid policymakers). With that understood, it is still possible to speak of a management hierarchy in some of the same terms used to describe an individual's organization, leaving in the background the knowledge that none of this would work if the individuals involved did not function properly at *all* human levels of organization.

Perceptions

All control systems control perceptions, not actions. Reference levels are specifications for perceptions, the action needed to match perception to the reference level being determined by the control system itself and by disturbances tending to cause error. An order passing from one level of management to another, therefore, is always a directive to perceive something, not do something, even if it is not always understood that way by the person issuing the directive. "Increase production!" is not and cannot be an order to produce more: it is an order to perceive that more is being produced. A manager who does not understand the difference is in for trouble.

Where do a manager's perceptions come from? From his

entire hierarchy, of course, beginning with intensity inputs. But the source of those inputs makes a difference.

In the human hierarchy, a given level of perception is derived strictly from the perceptions of lower levels; the sequence level of perceptions, for instance, cannot be constructed until the lower levels have created a world of objects and transitions. This is not quite the situation in a social organization. A production manager has options about where to get the information from which he derives a perception of production. He could stand at the end of the assembly line and count the nuts and bolts as they pass, and make his own tallies using nothing but his own hierarchy of perceptions. Or he could delegate various foremen on the floor to do this for him, telling them to pass the information on in the form of a production report. Telling them, that is, to *perceive* that they had passed on this information.

The latter is obviously the only practical course in a large organization. The production manager cannot normally see the ends of all the production lines, or stay awake 24 hours a day. Nor can he afford to mix in with the specialties of others in the organization at other levels, because to do so would certainly result in conflict and loss of production. A manager almost has to construct perceptions of the variable he controls out of the combined perceptions of many other people of the variables *they* control.

In the human hierarchy, one level of organization controls one class of variable, and does not deal at all in other classes; it is not equipped to deal with any other class. This is not true in a social organization. The foremen who fill out production reports are quite aware of the uses to which those reports can be put, at least in terms of their higher-order perceptions of the whole system. One effect of filling out a production report is almost certain to be an effect on the person filling it out, and if control systems do anything, they control the effects their own actions have on themselves. Higher levels of management do not simply receive unedited copies of the perceptions of those at lower levels; the information received by higher levels is at the same time an action taken by lower levels for the purpose of affecting what happens to those lower levels. At the same time that people at lower levels are acting as if they were

controlled by goals sent from higher levels, they are acting to control perceptions of what is happening to them, and one means of doing this is to vary the information they supply to higher levels. This situation results in complex interactions among the persons who make up the management hierarchy; it is worth exploring these complexities.

The effects of altered perceptions

Each level in a management hierarchy controls perceptions constructed primarily on the basis of perceptions reported by people carrying out lower-level functions, and only occasionally on the basis of first-hand observation. The perceptions that are reported are compared with their respective reference perceptions, and on the basis of any error that exists, changes are made in the reference levels transmitted to those from whom the reports come. A higher level of management controls, therefore, the interpreted meaning of reports, and not the reality which those reports are taken to represent.

Those who receive these "reference levels" actually receive perceptions of words, and must translate those words into meanings at many levels of their own perceptual hierarchies. The meanings of the directives are evaluated in terms of the lower-level person's entire hierarchy of goals, not just the aspects of it pertaining to the business of the organization. An order implying that someone should closely monitor the activities of his department, personally, for seven days and nights, would be disobeyed because (among many possible reasons) to obey it would violate personal reference levels concerning fatigue and ability to function. In many ways, therefore, both large and small, orders given by higher management levels constitute disturbances of the private perceptions of a lower-level functionary.

In a smoothly-running organization, such disturbances are of the simplest form: the recipient, on translating the order into its corresponding set of reference perceptions, finds that there is a difference between what is being perceived and the new specification of what is to be perceived. Because the agreed-to goal is to maintain a match between perception and the interpreted reference-perception, the change of specifica-

tion creates an error. The change qualifies as a disturbance in the technical sense of CT psychology because it enters the recipient's hierarchy in the form of an independently-caused change in a perception. Normally, the effect of the disturbance is cancelled by altering the orders given to still lower levels in the social hierarchy, and when those lower levels alter their own perceptions accordingly (and report the new state of affairs), the error at the higher level is diminished or erased. Following the order—doing what is necessary to make perceptions change to match the newly specified reference level—cancels the effect of the disturbance. This is how control is successfully carried out when a higher level of management detects an error and acts to correct it (or to keep it from becoming significant).

Organizations function on the supposition that this simplest form of action to oppose disturbance will be carried out in a straightforward way—that each level in the hierarchy will report its actual perceptions, and act to make them match the given reference-perception. It is unlikely that this supposition is always justifiable.

As already hinted, there can be strong motivations to make reports that deviate from actual perceptions. A manager who for any reason decides not to alter the orders he is passing down the chain can simply close what is known in the CT model of the brain as the "imagination connection." He can take the verbal order he has received, alter its syntax appropriately, and pass it back upward as a report of what has been done. Since this results in zero error at the next level up (or can), the result, ideally, is that no further changes in orders from above will occur. The manager operating in the imagination mode thus deals with his own disturbance from the higher-level person, and effectively cuts higher levels of the social organization off from what is actually being perceived at lower levels. People at higher levels, of course, are pleased about the alacrity with which their orders are made into realities, and may reward the lower-level person accordingly. That possibility, indeed, can be the effect on himself that the lower-level person intends to have.

The complete short-circuiting of higher-level directives through the imagination connection is, undoubtedly, an ex-

treme form of perceptual distortion, and even when it is tried is not likely to be successful unless there is collusion at an entire level of management, combined with a remarkable degree of insulation of higher management levels from personal knowledge of what goes on at lower levels. That should not, however, be taken to mean that in many forms short of the extremes this is a rare phenomenon. Shaded, colored, puffed-up, projected, wishful, and deluded reports are probably commonplace. I have never belonged to or heard of any organization in which reports to higher levels are not altered, as a commonplace procedure, from what even the parties involved, at the time or later, would agree privately is the true state of affairs.

The immediate effect of distortions of reports must be to make management think itself more successful than it is. Policies and strategies appear to have been carried out as ordered even when they have not been (or for that matter could not possibly be) carried out. It is not likely that a lower-level person would distort a report in the direction of causing greater error at higher levels, because as already mentioned this would probably result in undesirable effects on the person issuing the report. There must be a strong bias toward reporting that the specified goal has in fact been accomplished, as opposed to a bias toward exaggerating failures. Thus management in general must tend to receive the impression that its demands are being met to a greater extent than they are. This inevitably results in failure to take corrective steps as strong as are needed to maintain higher-level goals, and assuming that the higher-level goals have been established in ways that are good for the organization, the result can only be a loss of control of the perceptions important to the organization.

Motives for distorting perceptions

Is there any *general* reason we can identify for which a person in the management hierarchy might make reports that deviate from what he or she knows as the apparent truth? I think there is, and that it traces to the same theme that has held this discussion together in earlier parts: conflict.

The critical conflict is, of course, between the goals of the or-

ganization (either as accepted by its members at all levels or as conceived by the persons at the highest levels) and the goals of the individual. The individual wants or needs something for himself, and the consequence of making a truthful report is that this want or need will be frustrated. All that is required is that this conflict exist at the same level, whatever it may be called, at which the agreement to follow orders was made, or at a higher level inside the person. The agreement to follow orders would normally be made as a means of controlling higher-level perceptions of importance to the person making this internal commitment. If following orders causes errors at those higher levels, the agreement will be violated or at least poorly carried out because of the conflict.

Such conflicts are built into most organizations. Consider a high-level goal which seems important to most people: maintenance of whatever perception of self one experiences as good. For many people, a perception of self-esteem is derived as much from the expressed opinions of others as from self-observation and independent evaluation. One commonly expressed judgement of the worth of a person concerns his status, which boils down to how many people the person has to take orders from in order to survive. Status symbols are symbols of independence and power, of one's freedom to establish and achieve goals without external direction. Thus in an organization that is structured in terms of receiving and setting reference levels, status and self-esteem (defined in this way) increase as one functions at higher and higher levels in the hierarchy, being undisturbed only when one occupies the top level. This urge to function at higher levels is strengthened by the prerequisites which managers acquire, both monetary rewards and enforceable demands for expressions of respect and admiration.

Taking orders at all, for the status seeker, has to create an error in self-esteem. Many people in this bind falsify reports; the principle is to "keep them happy upstairs" while doing "what actually needs to be done." This bolstering of self-esteem can be heard at every level. To claim that one does what actually needs to be done implies that the person doing it understands what is needed at the highest level; to keep them happy upstairs by telling them what they want to hear implies that those upstairs are on the wrong track, and that the organi-

zation will suffer (and trouble will be caused at lower levels) unless they are placated by telling them that their directives are being carried out when something else is really being done. Conflict is created by the custom of confusing personal worth with position in the structure of a hierarchical organization. The giving and receiving of directives, which is essential to the functioning of such an organization, become variables which are too few for the number of people trying to control them, reducing the organization to the level of competition.

There are many less abstract ways in which a person may suffer internal conflict as a consequence of agreeing to achieve reference levels defined by others. Managers at the same level may issue directives that entail contradictions in lower-level tasks ("increase production" and "conserve energy," for example, which may entail operating a lathe in the dark). A management directive may cause physical and psychological suffering, through overwork or incessant criticism and accusation; unions arose from this sort of conflict. Individuals might be required to work in ways they find impossible; going through motions without ever being able to see what they are affecting, for example, which destroys the ability to control anything interesting. Or a manager may try to specify not only what goal a person is to achieve, but the actions by which it must be achieved: this places the manager's control in parallel with that of the worker, and as explained previously is likely to result in conflict.

Rewards and punishments

Rewards and punishments for performance are based on the same idea that underlies utility theory in economics: the idea that performance is a monotonic function of reward (and of the opposite sign for punishment), so that no amount of the reward will ever be viewed as "enough." In principle, according to this view, the more a person is rewarded for good performance, the better the person will perform; the greater the cost of poor performance, the less likely the person is to perform poorly. Within limits, this apparent relationship may hold, but not for the reasons usually assumed. The limits, furthermore, are such that the relationship expected is likely

to hold true only under conditions of extreme deprivation.

A reward is anything that will tend to correct errors in a person's control systems; it is associated with something the person wants. It follows that a person will not make efforts to obtain a rewarding thing or situation unless that person already has a reference level that is set higher than the presently perceived state of that thing or situation. Likewise, a punishment is anything which, if received, will cause an error to be experienced. Its effectiveness depends on disturbing some controlled perception. Both rewards and punishments depend for their effectiveness on the existence of some specific intact control system.

Punishment has already been recognized as counterproductive, since while it may result in avoidance of a given action, it cannot be used to steer behavior in any specific direction; the result is unpredictable. It is not generally recognized that reward is also counterproductive (that is, reward that is introduced outside the context of the person's ongoing actions and goals). I have said elsewhere that the power to reward depends on the ability to withhold the reward, and requires the one who proposes to give the reward to *create* error (or take advantage of an existing one) before offering to reduce it. Establishing the conditions required before a person can be rewarded is therefore likely to be taken as a hostile act.

But there is another effect of using rewards that is disruptive to a social organization, even when the participants have learned to accept the idea of reward as a normal feature of life. A reward is simply something visible that pertains to a person's goal structure, something that is involved in correcting errors. We can, in fact, drop the term reward and substitute the more general and fundamental term *disturbance*. To reward someone is to provide something that the person is trying to get, or is already maintaining, through his own efforts. A reward is thus a disturbance acting to move a controlled perception toward its reference level. Such disturbances are opposed by a control system through relaxation of the control system's efforts (owing to the decrease in error).

Consider the case of money rewards for good performance. If a person's goal for money is behind his performing a given function, then performing that function is a means toward the

end of receiving money. Presumably, the person is working at his job just hard enough to balance the need for money against the many other needs he has, such as the need for leisure time, for socializing, for doing well whatever the task, and so on. All errors are as small as that person knows how to make them; his individual hierarchy is in balance even if not in a state of absolutely zero error.

Now an external agency decides to reward this person by increasing his salary. Will this person then start working harder? That is unlikely; it is more likely that he will relax his efforts to make more money, since that error has been reduced, and redistribute his efforts to control other variables which are now (relatively) out of balance. Control systems oppose disturbances, even when those disturbances act to lessen error. A disturbance which aids a person's efforts reduces error, which results in a reduction of error-correcting effort. To repeat an important principle: A reward is simply a disturbance that aids the efforts a person is already making to achieve or maintain some goal-perception. Thus the effect of a reward that we are led by control theory to expect is the opposite of the effect that would be predicted from the concepts of incentive, reward, and conditioning. This important conclusion is at variance with commonplace assumptions, and for good reason: commonplace assumptions are based on a special case which we will now examine in detail.

The conditions for effective reward

There are cases in which providing a chance for a reward will spur people into action; these are special cases from which the concept of incentive has erroneously been generalized. It is true that under certain special circumstances, the lessening of an error will cause an increase rather than a decrease in the efforts being made to correct it. A quantitative analysis is needed to show how in detail this works, but a qualitative treatment can show at least in general terms how this situation can come about.

Suppose that associated with the effort to maintain control of a variable A there is a cost proportional to the effort. This cost can be expressed as a disturbance that tends to move vari-

able B below its reference level (where variable B is the variable in terms of which one would define cost). The error between A and A* (A's reference level) calls for an increase of effort, while the error between B and B* (B's reference level) calls for a reduction in the same effort. There is thus a potential for conflict between control of A and control of B.

Now consider the effect of a given increase of effort. The result will be to move A closer to A*, but to move B farther from B*. The question is, then, which effect is the more important to the organism? This is tantamount to proposing a higher-level control system which senses both kinds of errors and weights them, adjusting the reference levels for B and for A to maintain the combination in some reference state. Such a higher-level system might monitor the total energy reserves of the organism; if variable A involves eating, and variable B involves amount of effort expended in general, then A and B will have opposite effects on the total energy reserves.

If obtaining food requires little effort and the organism has adequate energy reserves, the reference level for food intake (A) can be set to some modest value, while the reference level for total activity (B) can be set reasonably high. Since both control systems, A and B, can achieve their goal-states at the same time, no conflict exists. Assume, however, that food becomes harder to obtain, so that for a given food intake, greater effort is required. Now the expenditure reference level has to be set lower, which demands less activity (to conserve energy) and more activity (to obtain more food). Conflict starts to appear. However, as long as decreasing the reference level for activity causes a net increase in the higher-level error (i.e., doesn't pay off), the reference level for activity will remain reasonably high.

When food becomes just difficult enough to obtain, the cost of seeking it will just exceed the benefit of getting it. At this point the total error in the energy-control system (the higher-level system) will begin to decrease as the reference level for activity is decreased. This may result in a still-higher setting for the food-getting control system, but this will not lead to greater activity; the decrease in the activity reference level will override it. (That is an approximate explanation; the quantitative explanation of these effects depends on nonlinearities of various functions.)

The result is a positive feedback situation in which a decrease of activity, which makes food still harder to get, leads to a further decrease in activity. The organism as a whole gives up, and makes a transition to a state of very low activity. This is a state of *very large* error, far larger than would ever be allowed by a normally operating control system. It is a state of extreme deprivation.

In this state, the normal relationship between disturbance and behavior is reversed. A disturbance tending to aid behavior in reducing the error now results in an increase in behavior, not a decrease. If a small amount of food is given to a starving animal (one which has learned and has available its own means, however ineffective, of producing food for itself), the animal will begin to increase its own food-getting behavior. Since positive feedback is now involved, this situation will tend to be unstable, and behavior will be very sensitive to small disturbances, the effects of disturbances now being amplified instead of opposed.

These are the conditions under which conditioning experiments are normally run. The reason is clear: under more ordinary conditions, where control is easy, disturbances are opposed and may have little visible effect on behavior. Under conditions of extremely large error, however, positive feedback exists instead of control, and behavior is very sensitive to disturbances. Since classical experiments are done under the assumption that external forces drive behavior, experimental conditions have been chosen under which the expected effects occur, other more normal conditions being rejected because under them "nothing happens."

Now we can return to the social situation in which rewards are used as incentives with the expected effects. We can see now that these expected effects will occur when the cost of getting a reward by one's own efforts is greater than the benefits. This will hold true only when the thing used as a reward is related to a controlled variable that is very far from its reference level, and when the person's own efforts are only weakly related to the controlled variable. Incentives or rewards will have the expected effects, therefore, only under conditions of extreme deprivation.

A definition of affluence

Affluence is usually treated as a situation in which people have nearly everything they want. This way of putting it, however, leaves out the most important consideration from the standpoint of CT psychology: how much does it cost them to get all this? Affluence is probably better defined as a situation in which a person can achieve his goals without excessive costs in terms of excessive effort, pain, or other unpleasantness.

From what we have just been considering, then, affluence can be defined as the state in which people can act as successful control systems, meaning that they can keep all their perceptions reasonably close to the state they desire. In order to do this, their behavior must be effective enough in controlling their perceptions to keep the benefit of action well in excess of the cost. In fact, an absolute threshold can be defined, the point at which the net effect of an increase or decrease in effort is just zero; where the cost of an increase of effort exactly matches the benefit.

If the total benefits of action decrease slightly from this point, or if the total costs of action increase, the result will be a rapid decline of effort, a giving-up, and a transition to the positive-feedback region in which reward and incentive cause an increase in effort. If the total benefits of action increase, there will be a transition into the control region, where a disturbance tending to reduce error leads to a relaxation of effort.

This threshold effect, this transition from positive to negative feedback, may be a far more useful way of distinguishing the poor from the rich, the deprived from the affluent, than a mere measurement of possessions, health, and so on. The poor and deprived are those who cannot try harder to correct their errors because doing so will only lead to a net increase of error. In this condition, people respond as classical theories of human nature would predict. Windfalls, opportunities, rewards, and outside help will encourage them to be more active in their own behalf, while adversity and punishment will drive them toward complete apathy. Unless their own efforts become effective enough, however, removal of these outside influences will result in their dropping back to the former con-

ditions. The question is not only whether an individual's goals are being satisfied—that could occur simply because of fortuitous disturbances that happen to be acting in a helpful direction. The important question is whether the individual has the means of maintaining all his perceptions at those goal-states even in the face of disturbances which *oppose* his efforts. The question is whether the person *has control of what happens to him.*

The most dramatic, and perhaps the most useful, applications of this analysis will probably be in the area of dealing with undeveloped countries. Undeveloped countries are and remain undeveloped because the people in them do not have enough influence on what happens to themselves; they do not yet have control. Simply supplying them with aid can make them seem affluent, but cannot increase the coupling between their actions and the consequences of those actions that matter. Therefore what aid is given should be in forms that will increase the effectiveness of each individual's efforts to control his experiences. Incentives and rewards will not have this effect. The only change that will bring people from the "reactive regime" into the "control regime" is an increase in the effectiveness of a given amount of effort. Paradoxically, making work easier will lead to harder work. And this is no more than good common sense: a well-fed farmer can and will plow more land than a starving man can or will prepare with a sharp stick.

Our subject here, however, is social organizations, and considering previous remarks, social organizations small enough to have a chance of stability. While "affluence" and "poverty" may seem strange terms to apply to persons within a managerial hierarchy, analogous states probably hold, for the same reasons that they hold in backward and developed countries.

Control and reaction in a social organization

The controllers in a social organization can be identified as those whose actions have strong effects on their own perceptions, in areas relating to the structure of the organization. The reactors can be identified as those whose goals relative to the organization are so far from being achievable that rewards and

incentives are effective.

This distinction does not necessarily correspond to a distinction between high and low levels in the social structure. Any person at any level can find his efforts to control the variables he thinks important so ineffective that the cost of making efforts is greater than their benefits. A vice-president who desperately wants a promotion, but finds that his efforts to get one have led to alcoholism, the threat of divorce, and avoidance by his peers may well give up those efforts and lose control of his life. He may find himself completely under the control of anyone in a position to reward him, even the janitor willing to give him a little respect. The janitor, on the other hand, may be affluent in the same terms, knowing exactly how to achieve the goals he has set for himself, being unmoved by the gift of a Thanksgiving turkey from the company.

I should point out explicitly that when a person has fallen into the positive feedback state, the reactive state of very large error, that person has ceased to behave like an effective control system relative to the variables involved in the large error. If an organization is set up so that any group of people (and I admit that this is likely to mean people low in the hierarchy) is seriously thwarted, those people will cease to be effective controllers. As a result the whole hierarchy will lose effectiveness, since the higher-level managers must, to perform well, leave the details of control to those at lower levels. Good control is required at the lower levels to prevent the effects of low-level disturbance from being passed on to higher levels, where neither the perceptions nor the means of control are adequate for dealing with them. The sign of a person who has passed from the control mode to the reactive mode is *apathy*: in the absence of external forces, the apathetic person acts as little as possible. A social organization built on refusal to allow lower-level people control of what happens to them will find those people turning apathetic, and that can't be good for any organization.

[1983]

A Bucket of Beans

Applications of Control Theory to the Understanding of Human Nature and the Design of Social Systems

At this meeting you will quite probably hear proposals for solving the difficulties that face our society and our world. These proposals will have to do with applying cybernetic ideas to the design of large systems; industrial systems, economic systems, and political systems. Some will just be high-tech proposals, but others will discuss how people ought to interact with each other and organize their society, whatever machines they employ in the process. If past meetings of our cybernetics society are any guide, you will hear a great deal about how a system ought to work, but little about the people who are supposed to make it work and who have to be persuaded to live under it.

This is a general shortcoming of what is called social systems thinking. One famous example of a social system design is expressed in the principle, "From each according to his abilities, to each according to his needs." The unspoken assumption about human nature on which this design depends is that human beings are willing to give all they are able to give, and will demand only what they do in fact need. When human nature proves to be different from what was assumed, the tendency seems to be to fall back on force as a means of making the theoretical system seem to work in practice. Another example, just to show I am not taking sides, is the principle of competition in a free marketplace. Theoretically, competition is supposed to drive quality up and prices down. In practice, what it produces just as often is either market-

Tutorial presented at the American Society for Cybernetics Meeting, October 7, 1983, at Palo Alto, California.

cornering collusion or ruinous conflict in which the parties involved use up their resources and everyone else's just to cancel each others' efforts. As a result the very same interests may find themselves lobbying for both removal of restrictions and protection against "unfair" competition. In the end, force is again needed to make the system appear to work.

At the core of any design for a social system is a model of human nature. Usually, however, the model is never formally stated, but exists only in the form of the designer's unconscious assumptions and prejudices. Equally often, the model is drawn, intentionally or not, from whatever current sciences of behavior are saying, right or wrong, about human nature.

Until Norbert Wiener came on the scene, there was really only one scientific model of human nature, a model that for the most part followed common sense. It can be expressed very simply: what people do is caused by what happens to them. We are attracted to things that are attractive, repelled by things that are repulsive. We can be enticed to work by the promise of rewards, and frightened out of evildoing by the threat of punishment. Incentives elicit our labor and remuneration reinforces our industriousness. Through approval and disapproval, satiety and hunger, pleasure and pain we can be made into members of a coherent society: we can, as is said, be "socialized."

The oddest aspect of this picture of human nature is its implication, very much in the background, concerning how these supposed properties of human beings are to be used in the ordering of a society. If proper and useful behavior is to be brought about by suitable application of rewards, incentives, disincentives, and punishments, somebody is going to have to apply these causative factors. Somebody is going to have to be in control.

But what does that mean, to be "in control"? In order to control something, be it a person, a society, or merely the path of an automobile, the controller must be able to preselect some desired outcome, judge the difference between what is happening and what should be happening, and select actions that will tend to make the "is" more like the "should be." The controller, in other words, must be able to work according to principles that do not apply to the people or things being

controlled. The model we would need to explain how the controller can control is clearly different from the model that is assumed to apply to the members of the human society that is to be controlled. Are we to use two different models of human nature, one for those in control and a different one for the hoi polloi? That is exactly what happens, although the model of the controllers is seldom made explicit.

It was Norbert Wiener who called our attention to the model that applies to a human being as a controller. But he did more than that. He introduced into our concept of human nature a new way of seeing how people work, whether we view them as agents or objects of control. He gave us, in short, the nucleus of a model that applies to all organisms, whatever they are doing. This model has slowly been gaining in importance for some thirty years, as we have seen more and more deeply into its significance. It has now reached the stage of development where it is ready to displace the old model completely.

The new model is based not on lineal cause and effect but on control theory. Control theory in this sense is the theory of systems that control, not the theory of how to control other systems. As we come to see human behavior more and more as the behavior of a controlling system, all our concepts about human societies will undergo a far-reaching change. We will have to start taking into account the controlling nature of the people we formerly thought of only as objects to be controlled. The new psychology that is growing out of fundamental cybernetic concepts will lead to a new sociology, and to new attitudes toward the problem of collective survival and well-being.

This tutorial will teach you a small amount about control theory and the way it applies to human behavior. There are many aspects of human nature that are perfectly real, such as consciousness and creativity, that it won't mention. They're important and we will have to understand them some day, but I don't know how to analyze them and won't even try. You may see, however, that control theory will lead us to some new ideas in these fields once we know enough.

Norbert Wiener put control first in the subtitle of his 1948 book, saying that cybernetics was concerned with control in, not of, the animal and the machine. That's basically what we'll

be talking about here: control in and by organisms. Let's start with a very simple question, one that seems to have been more or less overlooked in the rush to get on with more important matters. In fact this will be the one question I'll try to cover today, with the help of a bucket of beans.

What is behavior?

Behavior, of course, is what people do. In this approach, however, we're not going to be concerned with what behaviors people produce under various circumstances, but with the nature of behavior itself. Since we don't care about any particular behavior, we can choose any handy example. Let's start with a demonstration, so we'll all be thinking about the same example.

Here's a toy bucket. The handle, you can see, is held on by rather long rubber bands. Here's a bag of dried beans, which I hope to eat later as part of my sister-in-law's excellent cooking. And here is a volunteer from the audience, who will now receive a secret message.

Now the volunteer holds the bucket up by its handle. Watch the behavior of the volunteer carefully as I slowly pour the beans into the bucket.

As you can see, the weight of the beans stretches the rubber bands that hold the handle to the bucket. But the wrong end of the rubber bands moves. Instead of the bucket descending, the handle rises. The bucket remains more or less even with the back of the chair, which is what the secret message asked the volunteer to accomplish. Nobody would object, I think, if we said that this volunteer was controlling the position of the bucket.

I'll pour the beans back into the bag now and ask the volunteer to repeat the experiment. This time I have another bag of beans. When the subject is ready, I once again apply the same weight of beans. Only this time the whole bag drops into the bucket at once.

The subject still showed behavior, and in the end brought the handle of the bucket to the same position. But surely you noticed something different. When the beans fell in the form of a single sudden stimulus, we saw a complicated response that

went on for some time before it came to equilibrium.

The way we see the subject's behavior depends on the model we are using. During the first experiment, it was natural to see the subject as producing whatever action was needed to keep the bucket stationary. The influx of beans caused a continually growing disturbance, and the behavior changed continually to counteract the effect that the disturbance would have had if nobody was controlling the bucket's position. We could see that the bucket was under control becoause the subject's actions were obviously counteracting the effects of the added weight, and keeping the bucket nearly stationary.

But when we made the disturbance large and sudden, the control of the bucket receded into the background. Now the most dramatic events were the thud of the weight, the immediate descent of the bucket, and the rapid movements of the volunteer's hand. We saw an initial event followed by another event, a response to the first event. If the subject of control were left out, we might even say that the initial event *caused* the response. A simple experiment with continuous control turned into a complex stimulus-response experiment. Now we could forget about all quantitative measurements, and speak in terms of classes of events. We could call my action "bean-dropping," and the subject's response "handle-lifting." The appearance was that bean-dropping caused handle-lifting.

This is generally the relationship between control theory and the older cause-effect theory. Under control theory one would apply disturbances slowly and in moderate degrees, so as to be able to see easily what is being controlled. Under the older theory, one would apply large and even irresistable disturbances, and do so suddenly to create the picture of a sudden event. Then one would observe the action that followed this event, to see what the event caused the person to do. Under control theory we would see a continuous relationship between disturbance and action. Under the old theory we would be looking for a temporal sequence of discrete artificially-created events. Control theory explains what happened during the sudden stimulus as the action of a control system presented with a disturbance too large and sudden to be resisted while it was happening. In fact, as I will show Sunday in a demonstration session, using control theory one can predict

from the continuous case what the response to the sudden case will be. Under the older theory, the observation of the response to a sudden stimulus-event is useless for predicting what will happen under the continuously variable disturbance.

Now let's change the subject. We'll get back to behavior in the end, but first we must discuss a fundamental aspect of control that will help us understand behavior better.

Intention

A great deal of nonsense has been written on the subject of intention, the most nonsensical idea being that there isn't any such thing. The old cause-effect model had no place in it for intention (except perhaps as an illusion), because the same behavior that supposedly had been explained in terms of external causes didn't need to be explained again in terms of redundant inner intentions. Rejection of the idea of intention was a theoretical necessity, not the result of experimental disproof. The old model often requires discarding inconvenient observations.

The idea of an intention is absolutely essential for anyone who wants to understand how a control system works. In control theory proper, as handed down to us by control engineers, there is no term "intention," but there is an exactly equivalent entity that must appear somewhere in the hardware if the control system is to do anything. The entity is called the reference signal.

This subject is a little easier to introduce using hardware instead of organisms. Let's talk for a moment about that familiar old friend, the thermostat. What does a thermostat and its output device, the furnace, control? You say "temperature," and for many purposes that is a perfectly adequate reply. But we need a more exact analysis here.

Suppose the living room is at a comfortable 72 degrees. The host, getting ready for a party, lights the candles on a low bookcase, explaining to his wife, "Always use candles at a winter party; they keep the room from getting too hot."

What's this little gem of empirical wisdom? But look where he puts one of the candles. It is almost under the thermostat control unit on the wall. Sure enough, the furnace goes off and

the room cools to 65 degrees, and stays there. This shows us that the thermostat is not controlling the room temperature, because the room temperature changed and the thermostat did nothing about it. But if we could measure the temperature right at the sensor of the control box, what do you think it would be? Seventy-two degrees, just like before. The furnace simply stayed off until the cooling of the room counteracted the heating from the candle, in terms of the temperature *at the sensor*.

This shows that we have to be specific, saying not only what is under control but where it is under control. The thermostat, in the semifinal analysis, is controlling the temperature of its own sensor, turning the furnace on or off as required by thermal disturbances from elsewhere. It couldn't care less about the temperature out in the room.

But we have to be even more specific than that. Temperature is an abstraction that can't have any physical effects. Actually air temperature is a measure of the agitation of air molecules, but even agitated air molecules can't turn a furnace on or off. Before the agitation of air molecules can have any effects on the thermostat, it must be converted into some physical process inside the thermostat that is capable of affecting the furnace's electrical actuator. In this case, the converter is a bimetallic spiral carrying a contact at its free end. In the thermostat, air temperature is represented by an analogous physical variable, the spatial position of a contact.

Suppose this contact is very nearly touching another stationary electrical contact. If the room cools off, the bimetallic spiral starts to curl up and the contacts touch allowing electric current to flow that turns the furnace on. When the air warms up, the spiral relaxes and the contacts part, turning off the furnace. On the average the contacts remain nearly touching. So now we can say what the thermostat is actually controlling: it is controlling the spatial position of a contact, its internal representation of what we call temperature. We can verify this by hanging a small weight from the spiral, opening the contacts a small amount; we find that the temperature falls so as to bring the contacts back to the nearly-touching condition, and remains at the new lower temperature. The thermostat resists disturbances of its movable contact.

We have now found the actual controlled variable, although there might be several other ways of describing it. We have found where it is: it is inside the thermostat. We can think of the bimetallic spiral as the thermostat's sensor. A sensor converts the state of some physical input variable into an output variable that the system can handle internally. In the thermostat, the actual controlled variable is the output of the sensor, the position—not the input, the temperature. So in the final analysis, what the control system here is controlling is the position of a movable contact. We know, although the thermostat does not, that this position is an internal representation of the external variable. We say that the system is controlling temperature because we usually see temperature being stabilized. But we should remember that it is really controlling something else that varies as a function of temperature—and as a function of other factors, too.

If you understand just this much about control systems you will never make the kind of mistake that was made at Three Mile Island. A control system controls only and exactly its own sensory representation, nothing else. Thus if an indicator light is turned on by the command signal that also is supposed to close a valve, it doesn't matter at all whether the valve really closes. If the light says "closed," the control system is satisfied. That is what happened at Three Mile Island. Some control system designer didn't really understand what control is. That light should have been operated by a flow detector, shouldn't it? This, by the way, is also what is wrong with the "reafference principle" of von Holst and Mittelstaedt, if you happen to know what that is.

Finally we're ready to say what a reference signal is. In the thermostat, we discover it when the hostess, who understands control better than the host does, removes the candle and turns the control knob on the thermostat to 65 degrees. If we could see inside the box, we would see the stationary contact now moving, making the gap considerably larger. Now the sensor has to cool to 65 degrees before the moving contact catches up with the adjustable one, turning the furnace on and halting the cooling.

The position of the adjustable contact is the reference signal. Suspend your doubts about this use of "signal" for the

moment. The reference signal is measured in exactly the same units as the output of the sensor: spatial position. In effect, the reference signal shows the thermostat where it is to place the contact that represents actual temperature. Because the output of the sensor and reference signal are variables of the same physical nature, we can say with meaning that this apparatus is set up to *compare* the two positions. The outcome of the comparison is elementary in this case: no current flowing to the furnace actuator means too hot; some current means too cold. This crude control system can make only a binary judgement.

For the record, I will define "signal," as I use the term. A signal is a variable inside a system the magnitude of which represents the magnitude of some other variable, often an external variable. As the magnitude of the external variable changes, so does the magnitude of the internal signal. Under this usage I can call the position of the bimetallic sensor's contact the "temperature signal," or more generally, the "perceptual signal." Another useful convention: the reference *signal*, which is not normally observable, determines the reference *level* or *state* of the controlled variable, which we can observe. The reference level or reference state of the bucket was an observable position in space.

Generally control systems and other kinds of complex systems handle signals, not the variables for which the signals stand. The brain, for example, can't process information in the form of light or force or sound; it can process only signals representing such variables. So before the brain can deal with any phenomenon, each significant variable attribute of that phenomenon must be converted into a corresponding variable neural signal. Then the brain deals with the signals, not the phenomenon or its attributes.

You will notice that in the thermostat the perceptual signal standing for temperature is not a temperature, but a position. The reference signal standing for the temperature to be maintained is not a temperature, but a position. The internal representation is not a copy or a duplicate of the external variable, nor is it isomorphic to the external variable in any way but in magnitude. But the physical situation is set up so that normally when the internal signals are controlled, so is some external

variable closely associated with a signal.

If you disassemble any hardware control system you will always find a sensor and a signal that is the analogue of the external variable to be controlled. Sometimes there are many sensors and a computer that combines sensor outputs to produce the perceptual signal, which then represents an abstract external quantity such as relative humidity or density.

You will also always find something equivalent to a reference signal and a comparator that produces an output based on the difference between the perceptual and reference signals. It is this difference signal, or error signal, that causes action by the system.

In control engineering you will often find the reference signal referred to as the "input." That is because human users can usually adjust the reference signal to tell the artificial control system what level of the controlled variable to sense. I hope you will forget that usage, because it's best in a model of behavior to use "input" to mean strictly sensory input. Human control systems have no knobs allowing an external user to adjust their reference signals. Reference signals are adjusted by higher-level systems, just as the setting of the thermostat was adjusted by the hostess—but we won't get into hierarchies of control today.

Let's leave the thermostat now by stages, and go back to our human control system controlling the position of the bucket. As we return, let's note the parallels. First, what was the person actually controlling? Not the bucket, of course, or even its relationship to the chair. I'll give you a hint: what would have happened if I had asked the volunteer to do the same thing, but blindfolded?

As a first approximation, we can say that the person was controlling a retinal image. Any apparatus that could create the same relationship between action and the retinal image would have been controlled in the same way by this person: for example a computer simulation of a bucket, operated by a handle on a spring. But an image on the retina is an optical phenomenon; light cannot cause muscles to operate. Furthermore, that image is not exactly what was controlled, because it was interpreted in terms of an imaginary horizontal plane passing through the top of the chair-back. The person was

computing the actual position of the bucket relative to this plane, perceiving the bucket as even with the chair when it actually appeared to be below it, in terms of the visual image. The vertical distance was adjusted for parallax, and the *adjusted* perception was controlled.

So what was really under control was a computed relationship based on, but not the same as, the relationships in the visual image. This computed relationship must have been in the physical form of a neural signal or set of signals. The volunteer was controlling a perception of a relationship called "even with." This perception was a representation in the volunteer's brain: a perceptual signal.

The relationship of bucket to chair is simply a variable, as is the perceptual signal representing it. Perception reports the apparent state of affairs; it makes no recommendations about the state that is to be maintained. The reference signal, in short, is not to be found in the bucket or the chair or the actual relationship between them that we can see from outside. There is no reference image on the retina. The reference signal can be found only in the brain of the controlling person.

If a perception of relationship can be modelled as a neural signal, then so can we introduce another neural signal, just like the first one, that signifies the state of the perception that is to be maintained. Comparing two neural signals and producing an appropriate error signal on which to base motor behavior is no trick at all, requiring us to put nothing into the model that the brain cannot do. The other signal is the reference signal, as you guessed.

It is also the physical form that an intention takes. An intention is a signal in the brain that specifies the state to which a perceptual signal is to be brought, and in which it is to be maintained regardless of disturbances. Evidently we can become aware, to some degree and in variable detail, of such signals in our brains, for we are often aware of intentions before they are carried out. It is no more preposterous to propose that experiencing an intention is experiencing a reference signal than it is to propose that experiencing an external state of affairs is experiencing a perceptual signal. And there is certainly nothing metaphysical in this way of thinking either of perceptions or of intentions.

Experiments with control give us the means to infer intentions. We could infer the intention of a thermostat by watching the relationship between its actions and various forms of disturbance. The only kinds of disturbance to which it normally responds are thermal ones, and when it does respond it always opposes the tendency of a disturbance to alter the temperature of its sensor. We can thus deduce that the thermostat-furnace system has the fixed intention of maintaining a specific temperature. It's actually an adjustable intention, since we can see it change from time to time. It's not a conscious intention, but awareness is not what defines the principal function of an intention. Most human intentions, such as the intention that the body temperature be sensed at a certain level or that the body feel vertical, ordinarily operate without benefit of awareness.

In the bucket experiment, anyone who understood control theory could have seen immediately what the subject was controlling, and could have inferred the intention, which was evident in the stability of the bucket against disturbance. Such inferences don't support the neural model of intention I have proposed, but look at it the other way around. The neural model makes sense of the idea of intention. Sometimes, to accept a scientific idea all we have to know is that there is at least one model that can make sense of it.

We can now see that speaking of an intentional *action* is a mistake. Actions are hardly ever intentional. It is not the thermostat's intention to turn the furnace either on or off. We vary our actions in any way required in order to make our perceptions match the reference signals, which means that action must be adjusted according to the effects that disturbances of the controlled variable are having. The human thermostat, the hostess, did not intend for the furnace to be on or off; adjusting the thermostat was intended only to make the room feel cooler. To intend is to intend to perceive.

Only if it is specifically one of our own actions that we intend to perceive, as in the case of a dancer, can we call that action intentional. When an action is produced merely to maintain some other perception at its reference level, the action is neither intentional nor controlled. It can't be, because in order to use it for control, we must be prepared to vary it in any way required by external disturbances. If we intend to

move the steering wheel of a car in a certain preselected way, we will shortly drive the car off the road. To stay on the road it is necessary to let disturbances such as crosswinds, road camber, and soft tires determine the action. We keep the road where we intend to see it in the frame of the windshield. How we turn the wheel to do this can't simultaneously be an intention. Try it and see, but not when I'm in the car.

Control theory and social systems

Having persuaded our volunteer to hold the handle, and having obtained agreement that the volunteer would try to control the bucket in a particular way, I was then able to control the volunteer's behavior. By pouring the right amount of beans into the bucket, anyone could have made the handle rise to any preselected level. But this isn't really controlling behavior, is it? It's just using a disturbance to control an action that the volunteer was perfectly willing to vary in any way needed. The volunteer didn't come into the experiment with any particular goal for hand-position. That is why I was able to control it, as long as controlling it didn't prevent the volunteer from achieving the primary goal.

Let's suppose, however, that there are two handles for the bucket and two people trying at the same time to control its position. Let's suppose, too, that this control becomes very important: the people are offered $100 for each full second they can keep the bucket within one inch of the right place. Furthermore, one person is told that the bucket's top is to be held level with the top of one chair, and the other is told to hold it level with the top of a different chair. Knowing what you do about control theory, can you now predict where the two handles will end up?

When two control systems try to control the same variable, one likely outcome is conflict and failure of control. Even if the variables are independent, as they would be if two people were each controlling a bucket, conflict can occur if the controlling actions of one person interfere with the other's controlling actions. Tying one person's hand to the other person's hand with two feet of slack in the cord would, sooner or later, produce direct physical conflict.

Suppose now that we agree: all organisms are control systems, and all of their actions are aimed at controlling some perceived variable. They are always controlling something relative to internally-selected reference levels, as long as they are up and about. How do you design a social system in which human beings of this nature can coexist? It isn't easy, particularly if you try to make it work through controlling what people do.

One reason it isn't easy is that there is no way for an external agency to reach inside an organism and alter its reference signals. For the organism itself, altering reference signals is easy, but it's impossible for anyone or anything else. So there aren't many choices left if the goal is control. One choice is to become so powerful that resistance is meaningless: you simply crush the opposition and have your way. There is, unfortunately, a segment of the human race or at least a closely related race that considers this approach perfectly satisfactory. Another choice is to get control of some commodity such as food that everyone needs, and withhold it until other people or organisms voluntarily decide to do things your way. There are some people who would claim that this approach is even scientific, although they seldom mention the police force needed to keep it working. A third approach is to lie. If you convince people that acting in a certain way will get them what they want, they will often act that way, at least long enough for you to escape with the loot.

But presumably a designer of social systems is searching for a long-term solution to social problems, one that could in principle continue to operate forever without creating forseeable disasters. If that is what is wanted, none of the above solutions will work on all of the people all of the time, although any of them will work on some of the people some of the time. In order for any social system to work over the long run, the designer must take into account that people control their own perceptions relative to reference levels they choose themselves at many levels, and the designer must understand what control is and how it works. In particular the designer must understand that people are not just control systems, they are the best, most precise, most complex, multilevelled control organizations ever to appear on Earth, and as a result when

they come into conflict with each other the results can quickly become deadly. The designer of social systems who understands this will begin by rooting out of the design anything that depends on controlling other people. This isn't moralizing, and it has nothing to do with idealism. It is simply facing up to the fact that people are control systems, and therefore cannot be moved in the same way one would move a rock. A new concept of human nature requires new concepts of human interaction, too.

[1983]

Deriving Closed-Loop Transfer Functions for a Behavioral Model, and Vice Versa

Transfer functions have been used in representations of human behavioral organization since World War II. In general, a transfer function is a concise mathematical statement of the way the action or output of a behaving system depends on a stimulus or input to that system. Given a valid transfer function, one can predict the output of a system for any input, static or varying, provided the external setting remains the same. The purpose of this paper is to show both how a transfer function can be derived without the usual reliance on an analytical model, and how the method of derivation itself came from a relatively new approach to modelling behavior: a closed-loop, rather than open-loop, approach. The method and the approach may have applications in fields other than behavioral modelling: weather prediction, for example, or possibly economic modelling. The basic method is not new, but perhaps my simple-minded way of arriving at it is.

Transfer functions and mediation theory

There is a family resemblance between a transfer function and other kinds of principles that have been used to express theories of behavior. In nearly all experimental studies of behavior, there is one underlying model, based on what can be called *mediation theory*. Under mediation theory, what an organism does is strictly a function of what happens to it. The

Presented at the American Society for Cybernetics Meeting, October 1983, at Palo Alto, California.

form of that function, of course, depends on the internal constitution of the organism: a human being and a bacterium have different transfer functions. The form depends also on which inputs and outputs are studied and how they are measured. If the inputs are words and the outputs are more words, the transfer function might be stated in terms of percentage of correct responses to questions, or perhaps in terms of linguistic rules that transform statements like "Mary made the bed" into responses appropriate to one or the other of the possible meanings, such as "How neat she is!" or "Could she make me a chair?" If the inputs and outputs are stated in terms of occurrences of events, the transfer function will be expressed as a logical function: if this occurs, that behavior will occur, while if something else occurs a different behavior, or none, will occur. And by far the largest class of transfer functions studied under mediation theory is measured statistically, as a "tendency" for certain behaviors to be generated by an organism under some class of input conditions. In this latter application, the implicit transfer function "tends" to be rather simple: a constant of proportionality, or at best a weighted sum of the inputs, with suitable conversions from input units to output units.

Most of behavioral science is conducted under the assumptions of mediation theory. Even scientific method itself as most commonly taught assumes that if altering certain stimulus or situational variables has a systematic relationship to concurrent or subsequent changes in behavior, the form of the observed relationship must be describing something about the intervening, or mediating, organism. In fact, it is possible that few scientists recognize mediation theory as a theory demanding experimental test and vulnerable to refutation: it seems to be accepted by most not as a theory but as a fundamental principle of nature: all behavior, in the final analysis, must be traceable to external events. Indeed, the methodologies of many experimental sciences of behavior appear to depend crucially on the validity of this hypothesis.

A relationship discovered under the assumptions of mediation theory is, by definition, a transfer function. But discovery of a transfer function does not, reciprocally, support mediation theory, for transfer functions can be either real or apparent. The measurements, of course, can be taken as real, which

means only that they weren't fudged: that appearances were honestly recorded. But the transfer function that is derived to describe the behaving system depends on the model under which those measurements are combined to yield a result. The model expresses one's assumptions about causality, which determines, in turn, the causal pathways one attempts to describe. If the causal pathways assumed do not in fact exist, the transfer function will be spurious, although it will contain no hint of that fact.

Lineal action vs. simultaneous interaction

The original application of transfer functions to human behavior occurred in a study of human beings acting as parts of control systems: the term itself comes from its use in servomechanism engineering and filter design. When this mathematical approach was adopted for the analysis of control behavior (by workers who came to be known as engineering psychologists), it was applied not to human behavior as such, but to human behavior as part of a man-machine system. In effect, mediation theory was carried over into these control analyses, for in all such studies, extending into the present, the "man" part of the system was represented by a single box having one input and one output. The equations of control theory were applied to the whole system, not to the human being alone.

Whatever the shortcomings of this way of representing the "human operator," and we will come to them, this kind of experimental approach had several features that had been lacking in behavioral science until then. One was the use of continuous measures of inputs and outputs, as appropriate in tracking-experiments. The input variable was not simply an event that was counted: it was a physical variable, the position of a moving target on a screen or the angle of a control handle, which was always present and always had some measurable state. Thus there was no interval between stimuli or between responses. Instead, stimuli and responses were recorded as continually and smoothly varying states of physical variables.

The other feature was the recognition that physical variables interact simultaneously rather than sequentially. Under media-

tion theory, especially when applied in terms of events, there is a strong tendency to treat various input-output relationships as if each one took place in turn while all other relationships were put on hold. Thus a stimulus causes a response, after which the response may alter the environment so as to affect the next stimulus. Thinking of behavior in this way justifies concepts such as "response chaining," but also entails an untruth: the idea that a given stimulus is unaffected by the response it is producing.

There is no such thing as an instantaneous event in the macroscopic world. All real events have a beginning, a middle, and an end: during all events, physical variables undergo continual changes that take time. The effects of these changes are not held off until the final change, the end of the event, occurs. Creating artificial events by setting thresholds or marking arbitrary positions in a continuous movement does not affect the physical situation, which develops and changes smoothly as always. No rule of behavior based on such artificially marked event boundaries can have any force in nature, because the processes themselves undergo no special changes at artifical boundaries set at the whim of the observer. Thus we are always dealing with continuous variables in studying behavior, whether we choose to recognize that fact or not.

As a response begins, in most cases its effects propagate at the speed of light or of sound in the affected medium. Those effects begin to alter the effects on sense-organs, therefore, within microseconds of the start of a response, either by altering the external world or by altering the relationship of the sense organs to that world. In most cases of interest, the stimulus situation alters as the response develops, with no pure time-delay that is of any importance (it would make no difference if the propagation of effects were instantaneous). This means that if we wish to preserve physical accuracy in defining variables, we cannot employ the device of temporal sequence in analyzing relationships. The response does not affect the "next" stimulus: it is part of the ongoing current stimulus.

This simultaneity was recognized in the early transfer-function studies, and is the main distinction between them and all other kinds of experimentation with behavior before them.

Generally there are two simultaneous and independent relationships that hold between a stimulus input to a controlling person, and the person's output measured as the position of a control handle. One relationship is visible: it consists of the effects of the handle on the state of (most often) a visual display, combined generally with the effects of an ongoing disturbance that also affects the display. This relationship can be described quantitatively and accurately through application of simple physical principles and the fitting of mathematical forms to the known properties of the apparatus. This part of the representation describes how the state of the visual display depends on the position of the handle and the amount and direction of disturbance, at every instant. If we call the state of the display v, the position of the handle h, and the momentary magnitude of the disturbing variable d, this relationship can be expressed as $v = g(h,d)$. The form of g is known.

The other relationship describes, presumably, the person embedded in the control loop. The movements of the control handle depend continually on the state of the visual display, according to the internal organization of the controlling person. This is the relationship we would want to represent as a transfer function. With h being the handle position and v some measure of the visual display, we want to know the form of the function f in the equation $h = f(v)$.

We already have an equation showing how the state of the visual display, v, depends on h and on a disturbance d: $v = g(h,d)$. The variables h and v refer, in the two equations, to exactly the same physical variables, so they can have only one value each at one instant. It follows that the correct form of f, the transfer function of the behaving person, must be that which allows the observed values of h, v, and d to satisfy both equations at the same time, and at all times. There may be more than one valid form, in which case we want to find at least one of them, or even better the general form—but the latter would be very difficult.

The behavior of this person-environment system can be neither properly described nor properly understood in terms of sequences of events. We cannot, for example, imagine making an arbitrary change in the visual display variable v, and from that predicting the next motion of the handle, measured

as h. First, to imagine such an arbitrary change is to violate the model: making a change in v must be done by some means, and if we use that means we must incorporate it into the model and solve all the equations again to predict the effect of using it. Second, if the change is truly arbitrary—independent of all the other variables—the model no longer applies since we have destroyed the original relationships. And third, if the model remains intact, as we change v the effect of h and d on it will also change, so whatever our means of making the change, v will not vary in the way we would expect if the rest of the system were not there. For example, we might imagine using the disturbance to induce a change in v, since d is an independent variable and we can directly observe its link to v. But v is not a function of d alone; as v begins to change, the output h will begin to change, altering its contribution to the state of v. In fact in a control situation, the change in handle position will essentially cancel the change we would expect to be induced by varying d, and v might not change perceptibly. Thus while we may have intended to alter v in order to predict the behavior, our attempt fails at the point of altering v: it does not change as expected. Thus the behavioral changes resulting from changes in the disturbance d do not signify what we might have thought—or worse, what we conclude, if we fail to notice that v doesn't follow the changes in d that we might have predicted from knowing the nature of the physical link from d to v.

The most important change of concept introduced by engineering psychologists as they applied control theory was the almost complete abandonment of the ordinary cause-effect picture of behavior—even when they didn't realize they were abandoning it. Normally, since the visual display is clearly a stimulus input and the handle position depends on the response or output, the state of the display, v, would be considered the immediate cause of behavior, and the handle position h the effect. But in solving the equations for any specific instance of control behavior, we find that the only really independent variable is the disturbance d, which is not a function of either h or v. Thus the solution gives us, in terms of d, the states of *both* the input and the output variables. Furthermore, the visual variable, which would ordinarily be thought of as

mediating between the disturbance and the person (being the only means by which the disturbance can affect the person) is found to depend simultaneously on the disturbance and the response, and uniquely on neither one alone. Thus the immediate cause of behavior, the visual stimulus, cannot be predicted without first knowing the behavior. Under mediation theory, which involves sequential analysis, this would amount to a paradox or an impasse. The only way to handle it correctly is to employ simultaneous equations, and give up any simple notions of cause and effect. If space permitted, and a sidetrack could be tolerated, I could at this point insert some pungent remarks about current views on the effects of drugs and neurotransmitters, but I will leave such issues to the reader's own extrapolations. It is time to get down to cases.

The superposition principle and convolution

Suppose we have a black box with one input and one output. The input variable, v, and the output variable, h, are continuous with continuous time derivatives. We have recordings of the time-course of both variables, and from these data wish to represent the function f in the relationship h = f(v), where f will be called the transfer function.

The standard approach is to assume some analytic form for f, either a differential equation or a transform of one, having a few adjustable parameters, and determine the values of the parameters that best predict the output on the basis of the input. The obvious drawback of this approach is the assumption of an analytical form, which limits the degree to which real data can be predicted. The method we will see here does not involve assuming any analytical form, but it does entail retaining one standard assumption: linearity of response. As will be seen, that assumption does not seem to cause any serious problems.

In a linear system, the output can be conceived of as the superposition of responses to a train of instantaneous input impulses, where each impulse delivers an amount of input that is proportional to the actual input amplitude at each instant, and is of the same sign.

Each input impulse would, by itself, produce some output

waveform; generally we would expect a rapid rise of output followed by a slower decline back to zero. A "startle response" would be a good image. At any moment, the actual output amplitude would be made up of whatever is left of the responses to a series of impulses that have already occurred, from an instant ago to some time well into the past. The contributions of the impulses that have just occurred, still being on the rise, are small, as are contributions from impulses that occurred long ago and have almost died out. The maximum contribution would be from the impulse that occurred just long enough ago for its effect on the output to have reached its peak. Of course the direction, positive or negative, of each contribution would depend on the sign of the input variable at the time when each impulse occurred, and its amplitude on the input amplitude at that time.

If the input waveform is expressed as v(t), where t signifies time, and the form of the transfer function as f(tau), where tau measures a time also, the predicted value of the output variable, h(t), would be given by the "convolution integral" converted to a sum: the summing process conceals the assumption of linearity.

$$h(t) = \text{sum}[v(t - tau)\ f(tau)]$$

The maximum value of tau is selected so that no significant remnant of the impulse response is left after that many seconds (the units of t and tau are the same). Tau begins at zero. For each value of t, the above sum is calculated over all values of tau from zero to the maximum. Thus this sum at each time t represents the accumulated effects of the input variable at times in the past from t (the present time) to t - tau, each value being multiplied by the value of tau that corresponds to that time-delay. The function f(tau) itself, therefore, has the shape of the response that the system would show after one single impulse of unit magnitude. For that reason the form of the transfer function is also called the "impulse response" of the system. In graphical form it is exactly a picture of the system's response to a single standard impulse input.

Solving the above equation for f(tau) is clearly no trivial task. That is why the normal approach involves the assump-

tion of analytical forms that do permit inverse functions to be found; indeed that is the main criterion for selecting analytical forms. The solution we will see here, however, does not involve analytical forms; it is, in effect, a graphical solution, based on the general "method of steep descent" which permits, sometimes, solving large sets of equations by successive approximation. The main requirement for any method of steep descent is that it somehow guarantee that each trial solution will be a little nearer to the actual solution—preferably a lot nearer, although that is often just a hope.

I must confess immediately that my method will probably be more a source of amusement than enlightenment for a real mathematician (which I am definitely not). It has the virtue, however, that it works, and works within the computational and storage limits of a 64K microcomputer. Before developing the method further, we must examine the experiment from which the data are obtained.

The experiment

A subject sits before a CRT screen on which can be seen two objects: one pair of horizontal lines that always move up and down together, if they move, and a single horizontal line that just fits between the other two and also moves up and down, independently. The vertical screen resolution is 192 lines.

The subject grasps a control handle that swings in a horizontal arc; the arrangement is not optimal since the pivot is away from the subject, so the arm movement is rather complex. Subjects have no difficulty relating the horizontal movements to their vertical effects, with a little practice.

There are two versions of the experiment, both traditional. In one, the "pursuit tracking" form, the two target lines move up and down under the influence of a stored random disturbance, while the center line moves in exact correspondence to handle movements. The subject is instructed to keep the central line exactly between the pair of moving target lines.

In the other version, the target lines remain stationary at the center of the screen. The position of the central line is determined by the sum of handle deviations from center and the same stored random disturbance. This is called "compensatory

tracking," a name that has little to recommend it since what is going on is neither "compensation," an open-loop process, nor "tracking," which implies a visible moving target. The subject must now hold the central cursor still while the stored disturbance tends to make it move. The visible movements thus represent only the difference between the disturbing effects and the opposing handle effects. Data resolution is eight bits, signed.

There are 20 disturbance tables to choose from, each obtained by very rapidly sampling the audio output of a television set tuned between channels. The bandwidth is about 0.2 hertz, but was not spectrally analyzed. The data presented here should thus not be taken as benchmarks.

Each experimental run lasts 84 seconds, during which 2560 samples are taken and stored, so the time resolution is 32 milliseconds. Since the disturbances are stored and known, only the cursor (central line) position need be stored: the handle position can be exactly reconstructed. Also, storing the cursor position saves a computation step, since cursor position must be computed in any event. Each iteration involves (2560-256) * 256 or 589,824 computations, so saving a step is important.

The computer uses a Z80 microprocessor with 64K of memory running at 4 MHz. An Intel 8231 arithmetic chip is used to speed multiplications. The program has been written in assembly language, Pascal, and C: the results here are from the C version. The two main inner loops are in assembly-language subroutines: in other words, don't expect a BASIC version to run very fast.

The data, on reconstruction, show handle movements that are generally opposed to the effects of the disturbances; in the compensatory-tracking case, the handle and disturbance traces show a mirror symmetry while the cursor trace wanders slightly about the midpoint. For a practiced subject, the RMS cursor deviations are about 0.3 of the RMS disturbance deviations on the same scale. This results in a "stability factor" of -10 (Powers, 1978), which reflects about the optimum degree of difficulty for getting good transfer functions.

The method

In general the method involves assuming a trial transfer function, applying it to the input data to obtain a predicted table of handle positions, and then using the error, the difference between actual and computed handle positions, as the basis for making a correction to the assumed transfer function. This process is repeated until the RMS difference between actual and computed handle positions reaches an asymptotic minimum. The way of converting the error into a correction of the transfer function is the crucial part of this method; the rest is relatively straightforward. The initial transfer function assumed, incidentally, consists of all zeros.

Each experimental point, h(t), can be expressed as the value of a polynomial in which the unknowns correspond to the values of f(tau) and the coefficients are the corresponding experimental values of the input, the table of cursor positions. This table, however, actually consists of a table of values of h(t) + d(t), where d(t) is the disturbance table: the cursor position or the tracking error is given by exactly this sum in either version. Thus the convolution integral is actually

$$h(t) = \text{sum}\{[h(t - \text{tau}) + d(t - \text{tau})]\ f(\text{tau})\}$$

The real input variable is thus the disturbance, the cursor position being a dependent variable. However, since the stored data represent h(t) + d(t), we can use it directly in the summing, calling the argument c(t) for "cursor." It is well to remember, incidentally, that this is a closed-loop transfer function that we are deriving here, and it will not in general be the same as the open-loop function that would be observed without the feedback effects, if such is ever possible.

The working equation for computing trial values is thus

$$h(t) = \text{sum}[c(t - \text{tau})\ f(\text{tau})]$$

The cursor position is just c(t) = h(t) + d(t), so the above equation is equivalent to

$$c(t) = d(t) + \text{sum}[c(t - \text{tau})\ f(\text{tau})]$$

which is the working equation actually used for generating the predicted handle—now cursor—positions. The actual handle positions are also converted to actual cursor positions by c*(t) = h*(t) + d(t), so the error for each data point is just

$$e(t) = c^*(t) - c(t)$$

yielding the polynomial

$$c(t) = d(t) + c(t)\, f(0) + c(t - 1)\, f(1) + \ldots c(t - n)\, f(n)$$

where n is the maximum value of tau.

There are 256 values of f(tau), and 2560 values of the data points. This leaves 2560 - 256 = 2304 usable data points (we must start at least 256 time units into the data). This means that we have, in effect, 2304 equations in 256 unknowns to solve. The coefficients, f(tau) are the same in all equations, but we do not initially know what they are. The method proceeds by slowly altering these coefficients, recomputing the "predicted" cursor position after each change.

The system is highly overdetermined, and there is no exact solution, so we must find the best approximation possible. The best approximation can be taken as that which minimizes the summed squared difference h*(t) - h(t), where h*(t) is the actual handle positions and h(t) the reconstructed positions.

The method does not rely on minimizing, although it produces a minimum (if not "the" minimum). Instead it depends on a behavioral model in which the "equation solver" is conceived of as a set of 2304 analogue control systems, each acting to increase or decrease the values of the table of transfer function entries independently of the rest. The reference level for each control system is h*, the actual handle position for the control system's location along the time-scale. The controlled variable is h(t), computed as sum[c(t - tau) f(tau)] for that one value of t. The error, h*(t) - h(t), is then used to increase or decrease *all values of the transfer function entries by the same amount*. The *sign* of each change, however, is determined by the sign of the corresponding value of c(t - tau). For positive values of c(t - tau), the error is added to f(tau); for negative values it is subtracted. The result is that tracing any path

around this control loop (via any entry in the f(tau) table and its corresponding value of c(t - tau)), one finds that the feedback is negative. Thus for any given control system it is guaranteed that the net effect of the correction during one iteration is to move the value of h toward h*, decreasing the error.

To sum up the correction method, the correction for each value of f(tau) is given by

$$\Delta f(tau) = sum\{sgn[c^*(t - tau)\, e(t)\}/N$$

where sgn is the "signum" function, N is 2304, the number of usable data points, and the sum is taken over t from 255 to 2559, with tau constant.

To help avoid edge effects, the values of c(t) are set equal to c*(t) in the range t = 0 to t = 255.

These multiple control systems all act at once, an effect that is achieved by completing each iteration before changing the data on which the next is based. The net change of value for each entry of the transfer function table is the sum of all changes for that value of tau from all 2304 control systems. Even though each control system simply selects between adding and subtracting its error into the f(tau) table, there are so many systems acting at once that the result is the best net change that can be computed (or, apparently, very nearly so).

The remaining consideration is the stability of these control systems, which are strongly coupled through their output effects on the relatively small set of values of f(tau). Since 256 values of f(tau) can exactly satisfy at most 256 of the control systems at once, the systems are necessarily in some degree of conflict, and the final result will be a compromise among them all. As it happens the large number of control systems seems to get rid of the opportunities for mutual oscillations that one might in general anticipate. One suspects that the more systems there are the fewer problems there will be with this sort of oscillation. Such oscillations have been observed, but are removed by limiting the amount of final correction attempted.

There is, however, another source of oscillation, which is simply the possibility of a given control system's overcorrecting. In the computation this is taken care of by a slowing factor

applied indiscriminately to all systems, in much the way I have described (Powers, 1978) for stabilizing control systems that operate in discrete steps. The gain of each control system, whatever its inherent value, is made integrative by allowing only a fraction of the computed correction to be applied on each iteration. The fraction is simply reduced until the system with the highest loop gain remains stable. This means that other systems are unnecessarily slowed, but since the mode of control is essentially integrative, all systems will eventually reach the minimum error possible—with some unknown distribution of weightings. A more elaborate way of limiting loop gain would involve scaling each system's gain by the sum of absolute values of c(t - tau) for that system; then the error would be equalized in all systems at the end—at the expense of more computation time. I have not done that yet, nor does it seem necessary.

One aspect of the method for executing the convolution integral should be mentioned. Since c(t) appears on both sides of the equation (delayed on the right), the equation really represents a closed-loop system. Before each iteration, the values of c(t) are set to zero except for the first 256 entries, which are never altered from c*(t). In effect this computation is executing a model of the subject each time it is run, and in fact when it is used in place of a subject after the solution is complete, the behavior is indistinguishable from that of a real person.

Results

A transfer function was obtained from the data in 25 iterations of the above process. The total time required was 16 minutes (40 seconds per complete two-phase iteration). There are various ways of describing the final result; the simplest is to estimate the percentage of the variance in the cursor and the handle positions that is accounted for by the final transfer function in its closed-loop model. For the handle position this percent variance accounted for is generally between 0.97 and 0.98; for the cursor position, between 0.75 and 0.85 (well-practiced subjects who maintain consistent behavior during the whole run). For a test transfer function computed as the

difference between two negative exponentials (giving a fair approximation of the observed curve) and used to generate an artificial "experimental" run, the analysis recaptured the same transfer function and yielded a percent variance of handle position accounted for of 0.9995. In other words, without the presence of random behavioral variations the solution reached is essentially perfect.

It is interesting that nearly all of the correction is made on the first iteration: 90 to 94 percent of the variance is accounted for. The shape of the transfer function, however, changes greatly as the next few percent of difference is eliminated.

Discussion

This method arose, as might be suspected, from attempts to build models of the nervous system. In a hierarchical model, one possible organization involves higher-order control systems acting through varying reference signals reaching lower-order control systems, all in parallel. Such a model has many interesting features, the most important being that the organization of behavior is determined primarily by the perceptual side of the systems, and very little by the output organization. Basically, the outputs need not be weighted; they must contribute to lower-order reference signals with the correct sign (excitatory or inhibitory) to maintain negative feedback, but the amount of contribution is of very minor importance. I published a two-level model (Powers, 1979) showing how three control systems could independently control a force in the X-direction, a force in the Y-direction, and total force using three muscles lying at abitrary angles—without any weighting of the outputs. While developing that exposition I realized how the same principle could be used to solve the transfer-function problem, on which I had been working since the early 1960s (in the guise of a method for sharpening optical images!).

And now the development has come full circle, for the fact that this method works is proof that the basic principle of the hierarchical model can be extended to very complex systems, in which many independent control systems, each controlling just one aspect of a common world of lower-order systems, can

share control while remaining stable. Proof by demonstration, of course, is hardly general, but it is certainly encouraging.

References

W. T. Powers, 1978, Quantitative analysis of purposive systems: Some spadework at the foundations of scientific psychology, *Psychological Review 85*, 417-435.

W. T. Powers, 1979, The nature of robots: Part 3: A closer look at human behavior, *BYTE 4*(8), August, 94-96, 98, 100, 102-104, 106-108, 110-112, 114, 116.

[1983]

Learning and Evolution

In the present context, "learning" has a special restricted meaning: the process by which an organism acquires control of what happens to it. Much of what is studied under the label of learning should, I think, be thought of as performance. The capacity to reproduce past experiences is a natural feature of all control systems; studying that capacity is a way of characterizing the abilities of those systems, but does not explain their development. The principal requirement we must impose on any learning theory that is to accompany the present model is that it must be able to explain how new control systems come into being—especially the first new one.

Control by bias of random output

The basic problem that any theory of learning must solve is what Petrie (1981) refers to (in the context of applying control theory to education) as the "Meno Paradox." Plato had one of "Socrates' yes-men" (Petrie's phrase) wonder how we can enquire about any new subject: if it is new, we do not know what to enquire about, and if we do know what to enquire about, it is not new. To paraphrase: if an organism can neither sense nor control some new variable, it has no means of altering its behavior in the ways required to acquire control. If it can alter its behavior appropriately for bringing the new variable under control, it already has the required control system, or most of it. How could a really new control system come into

Excerpt from unpublished manuscript "CT Psychology—A New View of Organisms."

being, perceiving, comparing, and acting in relationship to a new kind of variable? How could a new level of control be added?

Both in my 1960 paper with Clark and McFarland and in my 1973 book were the beginnings of an attempt to resolve this paradox, patterned closely after Ashby's idea of superstability. This approach was anticipated by Campbell, in his principle of "blind variation and selective retention" (1956). As in most theories of learning offered since Thorndike's Law of Effect, it is assumed that the needs of an organism, when not met, somehow motivate it to learn, and that learning ceases when they are met. The concepts of need and motivation are easily cast in terms of control theory.

We can suppose that needs are associated with variables that are especially important indicators of the state of an organism's "life support systems" (although in a general theory there is no requirement that all needs be physiological). To say that a need is met is to say that one of these variables is in a particular state. When all of these "intrinsic variables" are in such states, changes in the organization of behavior cease. Motivation to learn, then, can be associated with a departure of these intrinsic variables from some corresponding set of intrinsic reference levels. Presumably, the intrinsic reference levels, at least those with which an organism begins life, must be specified genetically. "Need" can thus be described in terms of a difference between a perceived state of an organism and some reference state specified by inherited information (the state, perhaps, that we recognize in ourselves as contentment).

All that remains is to posit a comparison process that turns this difference into an intrinsic error signal, and an output process that will in turn convert the intrinsic error signal into action that reduces the intrinsic error to zero. That action is called learning, or to use a more general term, reorganization.

I have chosen a basically random reorganizing process, just as Ashby did. What I have added is a continuously variable rate of reorganization that depends on the absolute magnitude of intrinsic error. An intrinsic error does not cause any specific kind of change of organization in the behavioral systems: it simply alters the rate at which reorganizing changes are ran-

domly applied to those systems (or the parts of the organism that will become those systems). Changes of organization that increase intrinsic error lead to faster reorganization, and changes that decrease intrinsic error slow the rate of reorganization. Thus changes in the right direction tend to persist longer than changes in the wrong direction. When behavioral systems acquire a form that comes close enough to eliminating intrinsic error—meeting the needs of the organism—reorganization ceases.

This picture of learning is quite different from the one that underlies reinforcement theory. Under reinforcement theory, some unobservable property of special stimuli causes specific randomly-produced responses to stimuli to become permanent. Learning is imposed on a passive organism by an active and suspiciously intelligent environment. The organism learns specific responses to specific stimuli.

Under the theory described here, the random behaviors represent attempts by a reorganizing system in the organism to bring critical variables under control and keep them in specific states. The control systems brought into being and modified by the reorganizing process control external variables that affect the critical intrinsic variables. When a behavioral organization appears that prevents disturbance of the intrinsic variables, reorganization ceases and that form of control behavior persists. Organisms do not learn specific responses; they learn control systems. In a particular experimental situation, of course, that would look like acquiring a specific reaction to a specific disturbance. Under control theory, however, "generalization" is to be expected.

The reason for making the reorganizing process random rather than systematic was to avoid the "little man in the head" problem. Human beings, and to a lesser extent lower animals, begin life in ignorance. There is certainly some preorganization, but most of the variables that bear on survival are not initially under behavioral control. We must somehow acquire that control before we have developed any capacities for systematic or symbolic analysis, before we can recognize most features of the environment, and before we have developed any habitual strategies for learning. I have always thought of this as the problem of acquiring the first new con-

trol system. The chief question about this theory is whether a random reorganizing process could be effective quickly enough to explain the kind and amount of learning that we observe taking place in the lifetime of a single organism. I have not published much on learning because I have not known the answer to that question.

A few years ago, however, I came across the work of Daniel Koshland (1980), who has been exploring the control behavior of microorganisms (although he does not yet call it that). Koshland, through a series of mind-bogglingly ingenious experiments, has found over 20 substances that are sensed by, among other organisms, *E. coli*. Furthermore, he has discovered the internal chemical signals that represent these external variables inside the bacterium, and has determined the relationship of these signals to the behavior of the organism. He has, serendipitously, supplied a direct demonstration of the kind of control system required by my theory of reorganization.

E. coli can produce only two behaviors: it can spin its flagellae clockwise, or spin them counterclockwise. When spinning one way the flagellae unite into a single spiral that propels the bacterium straight ahead at a constant speed. There is no directional control. When spinning the other way, the flagellae fly apart like a shaken mop and the body of the bacterium tumbles in space. When forward motion recommences, it is in a direction demonstrably unrelated to the previous direction. With this primitive motive machinery, *E. coli* can make its way up a gradient of chemical attractant or down a gradient of repellent nearly as efficiently as if it could steer in the appropriate direction.

Koshland found by direct perfusion experiments that the basic output variable was the time interval between tumbles. In a uniform concentration of attractant, the delay between tumbles was constant at some baseline interval. When the concentration was increasing with time, the delay lengthened, and when the concentration was decreasing, the delay shortened. The quantitative relationships were determined with the bacterium restrained (!).

The swimming motion of the bacterium converts a spatial concentration gradient into a time rate of change of concentration (as experienced by the bacterium) that depends on direc-

tion of swimming. The bacterium has no ability to sense spatial location or direction; it senses only chemical concentrations. If the direction is down the gradient of attractant, another tumble occurs very soon; if up the gradient, the next tumble is delayed. The overall effect is that the bacterium spends more time swimming in the right direction than the wrong direction: the random walk is strongly biased up the gradient of an attractant or down the gradient of a repellent. I have modelled this sort of control system on a computer, and have found that if the sensitivity to rates of change is high enough, hardly any time is spent swimming in the wrong direction: almost as little as if the organism could actually steer. Control through varying the rate of application of a basically random process is far more efficient than I had dreamed it could be.

In order for this kind of control to work, the random output must have effects that are systematically and continuously related to the variable that is under control (in the case of *E. coli*, the time rate of change of concentrations). If a small change in direction of swimming by *E. coli* produced an unpredictable change in the time derivative of concentration, this sort of control would fail. The general implication is that the variables involved in control must exist in some continuum, such that small changes have small effects. An example of a system in which this condition is absent is a digital computer. The failure of "self-programming programs" that work directly with machine codes can be traced to the fact that the effect of changing one instruction in a program is not proportional to the amount of change: a given change from one instruction to another might produce any effect from none to catastrophic failure. No gradual convergence to a systematic final effect is possible in such a universe.

The kinds of changes that my proposed reorganizing system must make in the behavioral systems must obey a law of small effects: a small change must have a small effect. This implies some pre-existing organization. I do not yet know what conditions must be put on that preorganization. There is clearly room here for a large amount of experimental and theoretical work. This statement applies as well to the proposals in the next, the last, section.

Control theory and evolution

The final subject we will examine here is evolution. The reader may not be surprised by this time to find that I propose evolution to be an active behavior of a living system, not just an effect of an environment on it (not an entirely novel proposal: see Hitching, 1982). Of course we are now talking about many generations of organisms, not individuals, and the only possible location for a control system that can operate continuously through many generations is in the genetic machinery. The replication of organisms entails replication of the genetic material, and, I propose, passing along the states of any variables involved in control as well.

I should say first that natural selection will remain in effect whatever more detailed theory might be proposed. Organisms so poorly organized that they cannot control the variables on which life depends will be eliminated by the crude criterion of survival versus extinction. We scarcely need be concerned with them; what I am concerned with is the stability of species over long periods of time, and the strange appropriateness of those evolutionary changes we have been able to observe since Darwin's original discoveries.

To avoid going too far outside my field, I will confine most of my speculative remarks here to the basic process of control that is proposed, leaving ramifications (and proofs!) to other times and other workers. There is a basic relationship between control phenomena and stability of replication, which must have been at work even when the complex molecules that were to become living organisms first appeared. In this relationship we can find a precursor of the reorganizing process just discussed.

Replication of a complex molecule is affected by many variables, such as concentrations of substrate molecules and extraneous molecules capable of altering the replication process. We can view the substrate molecules on which replication depends as the central variables, and the molecules that can change the process as disturbances.

We start with the conclusion: suppose that the complex molecule develops an interaction with its surroundings that reduces the effects of disturbances on the substances that are

involved in the process of replication. In order for this to occur the molecule would have to sense and affect the concentration of the critical substance. By "sense," I mean only that something in the molecule would be systematically altered by the concentration whether or not replication was in progress, and by "affect," that the alteration would change the interaction in future replications and thus tend to alter the concentration of the substrate substance.

I am proposing, in short, biochemical negative feedback between the generations of replicating molecules and the substrate. This is no longer a far-fetched idea; Michael Savageau (1975) has worked out the beginnings of a general feedback theory that applies to purely biochemical processes (including genetic processes), complete with analytical mathematics and practical demonstrations. He has not yet made the extensions of this analysis that would be appropriate for analyzing biochemical control systems (the role of the reference signal is not included), but I anticipate that this step will follow shortly. Jones (1973) also anticipated this idea.

The argument now gets a bit involved, but having the example of Koshland's bacteria and the chemical principles developed by Savageau before us should help.

Between a complex molecule and its chemical environment there will be many interactions. Some of them will amount to negative feedback interactions; some of these interactions will involve substances on which the stability of replication depends. As replication proceeds, over and over, the disturbing substances, those that can alter replication, cause continual random variations in the details of the replicated molecules, thus continually changing the feedback relationships between the molecules and their immediate environments.

What will be the effect when a negative feedback interaction comes momentarily into being, when it involves a variable on which replication depends? That variable will become resistant to disturbances; the disturbing molecules (and physical conditions) will have a lessened effect on it. Whatever aspect of replication depends on the stabilized variable will therefore tend to persist for a longer time, through more replications.

Thus negative feedback per se will be favored; it tends to protect the very process that brings it about from further

change. Of course in the beginning this protection must have been transient, not permanent. As random changes in replication continued, however, the relative population of molecules in the negative feedback relationship with the substances on which replication depends must inevitably have increased, just as Koshland's bacterium inevitably moves up the gradient despite its frequent movements in irrelevant or wrong directions.

The same process would most strongly favor forms of molecules that produce the largest stable negative feedback effects on the critical variables. In the end, most or all of the variables that affect replication would be stabilized against disturbance by the acquisition of control systems for each one. At that point we would have molecules that could continue replicating for long periods of time, environmental changes no longer being able to affect the variables that determine accuracy of replication. This is the same bootstrap process that takes *E. coli* upstream.

An interesting feature of the behavior of control systems is that they can create a pseudo-threshold effect. If a steadily increasing disturbance is applied to a controlled variable, the control system will produce a steadily increasing opposing output, preventing any significant change in the controlled variable. But all real systems have limits on the amount of output they can create. When the disturbance reaches the magnitude where no further opposing output can be produced, further increases in the disturbance will cause the controlled variable to change as if no opposition existed. The appearance is as though there were a threshold for the effect of a disturbance on the controlled variable. (This may have some bearing on the so-called "blood-brain barrier.")

If the controlled variable is a substance on which accurate replication depends, a steadily increasing disturbance caused by long-term changes in the environment will, over many generations, have no apparent effect on the molecule. The increasing disturbance will be cancelled by increasing feedback effects. But when those feedback effects reach their upper limit, further increases in the disturbance will begin to alter the critical variable as if no control existed. Replication will suddenly begin to be affected. The molecule, in short, will begin to

evolve, probably going through a period of rapid evolution that will end only with extinction or the acquisition of a more effective control system. The effect is very much like the one proposed by Prigogine, et al. (1972); there is even a family resemblance of mechanisms, although my proposal has nothing to do with thermodynamics.

This model of evolution can easily be extended to co-evolution (Durham, 1982), for any interaction of organism and environment that affects the accuracy of replication will tend to favor control systems over any other kind. The overall picture is one of a new mode of evolution, one that centers on selection criteria that are much more useful to the species than is the bludgeon of ordinary natural selection. We can begin to see that life itself may be intimately related to control phenomena, and that the principles of control apply to organisms at every level of organization.

References

Campbell, D. T. (1956). Adaptive behavior from random response. Behavioral Science 1, 105-110.

Durham, W. H. (1982). Coevolution and law: The New Yam festivals of West Africa. In Gruter, M., and Bohannon, P. J. (Eds.); Biocultural foundations of law (in press). I am indebted to Harvey Wheeler and Paul Bohannon for bringing a manuscript of this remarkable paper to my mistakenly reluctant attention. The author is in the Department of Anthropology at Stanford University, Stanford, CA 94305.

Hitching, F. (1982). The neck of the giraffe. New Haven: Ticknor & Fields.

Jones, R. W. (1973). Principles of biological regulation. New York and London: Academic Press.

Koshland, D. E. (1980). Bacterial chemotaxis as a model behavioral system. New York: Raven.

Petrie, H. G. (1981). The dilemma of enquiry and learning.

Chicago and London: University of Chicago Press.

Powers, W. T., Clark, R. K., and McFarland, R. L. (1960). A general feedback theory of human behavior. Perceptual and Motor Skills 11, 71-88 (Part 1) and 309-323 (Part II).

Powers, W. T. (1973). Behavior: The control of perception. Chicago: Aldine. Now published by Walter de Gruyter.

Prigogine, I., Nicolis, G., and Babloyanz, A. (1972). Thermodynamics of evolution. Physics Today, 25, Nov., pp. 23-28 (Part I) and Dec., pp. 38-44 (Part II).

Savageau, M. A. (1976). Biochemical systems analysis. Reading, Massachusetts: Addison-Wesley.

[1986]

An Agenda for the Control Theory Group

As I was looking forward to this second annual meeting of the Control Theory Group, I found myself getting strangely depressed. It took a while to figure out why. Our numbers have grown—at least the mailing list has grown—to somewhere between sixty and seventy, and there are indications that it will grow much more in the year to come. The thought occurred to me that we're approaching a time when we will no longer have to concentrate on overcoming opposition. I wondered if I might have lost the knack for doing anything but showing what's wrong with conventional ideas and writing introductions to control theory. That unpleasant idea led to wondering just what I would do if someone handed me a blank check, and said, "Go to it." And that wondering has led to this talk.

It's strange how difficult it is to find applications for the first new conception of behavior in over three centuries. We've all had this trouble, but those of us who have had the most trouble are those who know the most about the life sciences. I think that's a clue. Think of clinical psychologists (we have a few of them here). What could clinical psychologists do with control theory? Well, they could play how-and-why, they could try to find out what the client is controlling for. But how would they use these techniques? Probably to try to help people get over anxieties, phobias, compulsions, depression, or stress. And that's where I think the difficulty starts.

No matter what discipline in the life sciences you are at-

Distributed at the Control Systems Group Meeting, August 20-24, 1986, at Kenosha, Wisconsin.

tached to as scientist or layman, when you look around for something to study, you find that the human pie has already been sliced. Everything has a name. People study personality, intelligence, aptitudes, attitudes, and preferences. They study mental illness. They study customs. They study marital problems. They study crime. They aren't trying to find out if such things really exist: if they have names, they must exist. The categories in which we think about human behavior have already been established, and that is really why it's so hard to wedge control theory into the structure.

Control theory really doesn't fit any of the established categories, any of the official objects of study. If you discover that someone has difficulty in controlling for principles, what category of human behavior is that about? This difficulty could involve stress, criminal tendencies, depression, confusion, ineptitude, and voting preferences. From the standpoint of someone whose primary interest is in one of those established categories, control theory might seem a tool for finding out more about such behaviors and conditions. But isn't that a way to pull the teeth of control theory?

I'm trying to say that control theory treats the human pie in a new way. It doesn't slice it into clusters of symptoms: it slices it into layers. And these are layers of perceptions and functions, not of specific behaviors. The control theorist doesn't need to think in terms of typical named classes of symptoms. In fact the control theorist doesn't expect any kind of behavior to be "typical." The control theorist tries to understand how people do what they do. The question of why they do it turns into how they do what they do at the next level up. In studying how people control what they control, the control theorist discovers problems with perception, comparison, and choices of action, problems having to do with insufficient learning, with conflict, with actions that don't work. Teaching control theory is sometimes a way to alleviate these problems, because many of them come from a person's trying to behave like an imaginary animal, a stimulus-response device. Sometimes the way is to help a person understand his or her map of control systems and their goals. Most of the time, I suppose, the way to help is made up on the spot, depending on what isn't working right. But whether the motive is to help or simply to

understand, the theory is the same, the model is a model of the same creature.

I'm speaking, of course, of the ideal control theorist who is free from all entanglements with existing categories of thought about human and animal behavior. I could be wrong, but I don't think there are any of those paragons here. We are all more or less stuck with one foot still in the old world, the world we grew up in. For most of us, applying control theory is still done in the context that we inherited from others, our peers, our parents, our teachers, our colleagues. All of us are still in the habit of thinking of control theory as something to apply to something else, some category of behavior that is already recognized.

What we all have to realize is that control theory is not a tool to be used in sociology, psychology, linguistics, economics, and the rest. It is the basis for a direct study of human and animal nature that makes no disciplinary distinctions and recognizes no traditional boundaries. A sociologist who begins studying human nature as a control theorist and an economist who does likewise will end up doing the same kinds of experiments, asking the same kinds of questions of nature, and using the same model of a human being. No matter what field the control theorist is employed in, the subject matter is the same: how organisms work. Through control theory we can arrive at a model of human nature that will be the same no matter who uses it. The scientist who is studying the relationships between brainwaves and behavior will be asking the same question as the clinician who is trying to help people who bring their problems in to be solved: the question is, "How does this person work?" The answer to that question, in whatever context, is the answer to whatever problem one has posed: a paradox in economics, a chronic misery of the mind, a phenomenon of mass psychology, a puzzling linguistic pattern. The question is always, "How does this person work?" and by extension, other people too and people in groups large and small.

The point of this talk is more to raise questions than provide answers. I'm saying that to realize the potential of control theory, we must begin to apply it directly and not think of it as a means of finding new answers to old questions. The old questions are loaded with assumptions, and their answers are

probably meaningless. We need new questions, and the basic question, I suggest, is how people and animals actually work. We have a believable model to start with.

If you still have a problem thinking up something you want to know about human nature, you might begin by going up a couple of levels. There are certain phenomena that remain unexplained, and we have a rather large stake in understanding them. War. Poverty. Crime. Drug addiction. Corruption. Misuse of power. Ethnic hatreds. And violence everywhere. Do you understand what is causing these problems? Have you considered that control theory may well be the tool that will finally allow us to solve them? And if you haven't, why not?

An old-time mystic said, "Think in other categories." That is what we must do to arrive at an agenda for the Control Theory Group. I am through talking: now I want to listen.

[1987]

Control Theory, Constructivism, and Autopoiesis

The objective of this essay is not to achieve a final synthesis of the three concepts in the title, but to explore similarities, differences, possibilities, and alternatives within the consensual domain these words suggest. My biases as a control theorist with a favorite axe to grind will probably be visible just as one can see biases in other approaches, reflecting their authors' primary orientations. My conclusion—for the benefit of those who seek certainty in science and thus can be excused early from this discussion—is that human consciousness and the world in which it lives are likely to remain a mystery even after we have applied our best intellectual tools to them. The most we can hope for while exploring waters this deep and murky is to find a few hints as to directions we can take.

One effect that is hoped for from this essay is a clarification of the relationship between the approach taken by Humberto Maturana and my own considerably simpler approach to an understanding of human nature. The term "simpler" is not intended to be approving: it can also be taken to mean "less advanced" if one is so inclined. The article "What Is It to See?" (*Cybernetic*, Fall-Winter 1985) is my working source for reference here to Maturana's ideas, in addition to understandings gleaned elsewhere, I hope correctly. Readers who are coming into the middle of this conversation can find my view, a little out of date, in *Behavior: The Control of Perception* (Powers, 1973).

Distributed at the American Society for Cybernetics European Conference, March 15-19, 1987, at St. Gallen, Switzerland.

Common ground

There are two striking similarities between Maturana's view and mine. The first is the ultimate intellectual ground, the bedrock on which we both stand. For both of us, the questions of experience and action, how they work and what they are, begin with an Observer and something that is observed. Putting aside all descriptions, conclusions, analyses, and other propositions about experience, we agree on the observation that for each sentient being, there is a world of experience that simply is, without explanation and implying nothing. This world is the reality we cannot doubt, because even doubting is itself part of it. This is the view to which the theorist must learn to return on those numerous occasions when words lead into a maze having no apparent goal or escape route.

The second similarity is of quite a different kind: Maturana and I both find it convenient and instructive to accept the neurological model constituting the scientific picture of the brain and central nervous system. We have both discovered —and no doubt there has been some mutual influence here— that the nervous system is conceptually closed. That is, we have both seen the main implication of the neurological model, which is that the brain and whatever observing function is associated with it (or vice versa) knows a world that is made entirely of activities inside the nervous system, having no other way of grasping what lies outside the nervous system, the "real" reality that Maturana calls the "medium" and I call simply the "environment." We are both a bit nervous about the implied solipsism of this view, even though we can hardly reject it while still accepting the neurological model and the implied fact that we have, or are, nervous systems ourselves. Since we both reject solipsism a priori, we are both looking for a way out of it, with incomplete success.

This agreement extends to the understanding that even our actions have consequences that are known to us only after they have been encoded by the nervous system. To ask how the objects and events of the real world are encoded into neural signals is to miss the point: objects and events are already in the form of neural "codes" by the time we become aware of them. This is how neural codes look, taste, feel, and so on. The

situation was put best by von Glasersfeld: the brain is not the Black Box: the environment is. Our actions seem to affect a real objective world, whether those actions be the informal ones of everyday living or the formalized ones of scientific experimentation. To recognize that the knowable effects of actions are really activities in the brain, already encoded and transformed, is to be forced into an unorthodox view of scientific objectivity. What can objectivity mean when it is applied to a world that is subjective from the start?

There is one final agreement, although Maturana's approach is somewhat divergent from mine. It is that the nervous system and the organism as a whole are "autopoietic." By the actions that grow out of its structure, it maintains its own structure. Maturana, with Varela, has tried to describe the essence of an autopoietic system in very general terms, using models that are largely mathematical and abstract, and demonstrations that are instantiations of the mathematics, not processes that might be physiologically observable. My approach has been to postpone the problem of development and system maintenance as long as possible, trying to characterize the structure that appears in the adult organism without asking how it got that way. My only contribution to a model that has autopoietic properties is the proposal that there is a "reorganizing system," the properties of which I have purposefully left vague since I do not know what they might be.

The control-system model, however, does help to make a distinction that appears missing from the Maturana/Varela approach. This is the distinction between what might be called in traditional terms "performance" and "learning." It is not, however, quite that naive a distinction.

In the control-system model there are two major classes of control hierarchies at work. One hierarchy can be called the "behavioral" hierarchy, the one that uses muscles and hormones to produce perceived effects (there are no muscles or hormones in Maturana's model, but we will get to that subject later). This system pursues a hierarchy of goals that relate to the perceived world, including what is normally considered to be the "environment" but also what is considered as the world of emotion, feeling, and proprioception in general.

The behavioral hierarchy attains its adult form as a result

of a long process of reorganization, which entails a control system of a different kind. The ultimate goals sought by this system are inherited, not learned. These built-in goals provide the ultimate criteria for retaining or dissolving any new organization in the behavioral hierarchy. In a human being, these built-in goals or intrinsic reference levels constitute a sketchy definition of what it is to be human. In a chimpanzee they define what it is to be a chimpanzee; in a bacterium, what it is to be a bacterium, and so on. All organisms are free to choose any hierarchy of behavioral goals whatsoever, as long as they are not inconsistent with the inherited definition of what it is to be that kind of organism—and of course, within the limits of "choosing" available to that organism. Evolution changes these definitions, but not fast enough to matter to an individual organism.

The behavioral system can be studied strictly in terms of performance: what goals are sought, by what means, and in what relation to other adjustable goals. It can be studied at a given time as if it were a static organization (static not in terms of behavior, but in terms of functions). The reorganizing system can be studied by learning how the behavioral system changes its functional characteristics as opposed to the specific activities being carried out by those functions (what Maturana calls, I believe, "structures").

This separation of performance and reorganization is very convenient, but it suggests that these processes alternate instead of coexisting. Thus Maturana's model, which does not make this distinction, is probably closer to the truth. It does not, however, seem to be any easier to apply to real behavior, so for the time being I shall maintain the distinction. In partial defense, I should point out that organisms are not actually autopoietic in any exact sense. The message in the genes translates into a structure that continually grows and changes form throughout life, reaching some peak of organization and then declining to the inevitable end. Autopoiesis is a relative concept: it expresses the resistance of the organism to the dissipative influences of inanimate nature, and also the resistance to disturbance of the path taken by the organism (and the species) as it develops in its characteristic ways through a variety of unpredictable circumstances. An autopoietic organ-

ism does not actually perpetuate a single organizational form forever, as do the mathematical instantiations of computational models. I do not mean to imply that this fact would be news to Maturana. The point is just that mine is not the only approach that departs somewhat from the case we suspect to be the truth. The truth is probably that what I call reorganization has influences on a time-scale that overlaps the one appropriate for studying performance, and that both reorganization and performance arise from one set of complex properties. But that conception is beyond my ability to analyze.

The areas of agreement between Maturana and me are large. The most important fact is that both of us stand estranged from conventional scientific understandings of behavior: it would not do for us to take our differences too seriously, for the same reason that the defenders of the Alamo had reason to put their smaller disagreements aside while considering how to deal with Santa Ana's army (not, we may hope, a prophetic simile).

There are, however, differences, and they are not trivial.

Cleaving the continuum

The only perceptual operation of the nervous system that Maturana describes is that of making a distinction: drawing a boundary that simultaneously creates two complementary classes. While I also accept the model of the nervous system as useful, I have reached essentially the opposite interpretation.

At the interface between the nervous system and "the medium," there are millions of tiny sensory receptors that report, as variable-frequency signals, their own states of stimulation and nothing else. Each sensory signal could in principle vary independently of the state of any other sensory signal. Thus the world of perception begins maximally divided: it cannot be subdivided further.

I have proposed a model in which perception (a generic term applying at all levels from raw signals to concepts) is synthesized in stages out of the behavior of these lowest-level "intensity" signals. A given perceptual organization, one of many at any level but the first, generates a perceptual signal that is a (learned) function of some set of perceptual signals

arising from a lower level. These perceptual functions are convergent, in that they all entail a many-to-one transformation. In this way the nervous system creates a world out of the first-order intensity signals, consisting successively of sensations, configurations, transitions, sequences, relationships, categories, programs, principles, and system concepts. I will not defend or elaborate on these definitions: the common-sense meanings will serve here.

The first six of these levels create an analogue world of perception, in which all perceptions can vary smoothly over their whole possible ranges. Many of the perceptions at all these levels are under skilful control, in a hierarchy that operates on continuous variables without benefit of symbol manipulations. This is the world that provides a continuum to cleave.

The pertinent level for this discussion is the seventh: categories. At this level, perceptual functions become organized to treat some set of incoming lower-level signals as instances of the "same thing." The continuous variations are lost: only the presence of a signal well above threshold matters. Some of the signals received stand for configurations or sequences of marks or sounds that are then used as the names of the categories. This is not to say that categories exist objectively and are detected at this level: once a particular category-perceiving function is formed, it begins to report the lower-order world in terms of that category, such as the category of things that are "mine." If none of the pertinent lower-level signals is present, the category-perceiver simply emits no perceptual signal. The nature of the category is perfectly arbitrary, although presumably we do not retain perceptual functions that do us no good.

The next level up, I have proposed, is the "program" level. This is the level where all functions, including perception, are rule-driven. Logic belongs at this level. In terms of logic, the absence of a signal is equivalent to presence of not-that-signal: if there is an input path where a signal might be received, then either state of that signal, present or absent, is explicitly represented. So in terms of logic, if there is no dog-perception, then that perception is reported as not-a-dog. At the program level, it stands to reason that creation of a category also creates another complementary category. That is only logical.

That is not, however, the way perception works at the category level. If there is no lower-order perception present that is in the set that the dog-category-recognizer senses, the result is not a report that there is no dog, but simply silence: no perception. To see that this is true, all one has to do is look around at the world. One might perceive grass, sky, cars, people, carving-knives, and rotation, but these perceptions are not recognized as not-dogs at the same time, any more than a dog would be seen as not-a-carving-knife. If anyone were so organized, there would be mental pandemonium, for no matter what one could name, it would simultaneously be *not* every other category in terms of which one is capable of perceiving.

Only at the program level, and even then only if the rules in effect are those of Boolean logic, can we recognize that the existence of a category automatically implies the complement of that category, the basis of the Spencer-Brown calculus.

In the process of defining these levels of perception (at an average rate of one third of a level per year), I have noticed the interesting and often dismaying fact that consciousness seems capable of operating as if from the standpoint of any of these levels. The dismaying part is that whatever the level in question, it becomes completely invisible *as* a level. That is why real levels of perception (I hope) are so hard to recognize. One can manipulate symbols for categories according to some set of rules such as differential equations or logic or grammar, and be totally conscious of what is going on except for the fact that one is using those rules. This leads to some strange effects, such as proving logically that there is no such thing as logic, or uttering words (that others can hear) meaning that mental phenomena can have no physical effects. One is simply unconscious of self-contradiction until one runs into a paradox—and even then, one can easily thrash around in the coils of reasoning without discovering that the way out is up.

Where the Maturana/Varela/Spencer-Brown approach fails, I suggest, is in using this program-level thinking to talk about categories, and logically, by the rules of the calculus, concluding that all we do is categorize. The Spencer-Brown calculus is itself a direct demonstration that we do other things, too. While consciously working from the level of programs, especially very difficult programs that require all one's conscious-

ness to operate (whatever that means—please don't ask), one sees the logical aspects of the categories not as interpretations at a higher level, but as properties of the categorizing process, right there in the categories, just as an artist might see the significance of his painting as residing right there on the canvas where any fool can see it.

This is not a judgement about the value of the *Laws of Form* approach. I am not competent to pass judgement on that. I am only trying to point out the contradiction inherent in perceiving complex logical relationships while deducing that the only function of perception is to do something else. In fact, I accept that the creating of categories is a real function of the brain: I have accepted this for some years, possibly because of hearing of Spencer-Brown's calculus (in the *Whole Earth Catalogue*? I think so). But I am also pointing out that logic, too, is something the brain does, operating on the names of categories created at a lower level.

The decision to employ Boolean logic is not a logical decision. It is a decision dictated by a principle, and at the same time is an instance of that principle. We use logic to satisfy a principle, not just to reach a conclusion. We could call the principle "being logical." And being logical is not the final goal, either: we are logical because that is one means of maintaining a concept called "rationality," one of many concepts that provide a context within which we order all of experience. I call these "system concepts," which is the highest level at which I have been able to catch myself being conscious.

Thus, I have some reasons to see the "calculus of distinctions" as a rather arbitrary game that seems to miss most of the activities that go on at most levels of the brain.

The nervous system and the physical environment

If we are going to analyze experience in terms of the neurological model of the brain and nervous system, it makes no sense to eschew the physical model of the world on which the neurological model depends. Maturana has interpreted the closed-loop nature of the relationship between the nervous system and its environment as if it could be wholly understood without a corresponding analysis, at the same level, of

the surroundings in which behavior and adaptation take place. I do not think this can be done.

In part I am objecting to the rather severe lack of clarity in Maturana's expositions. When he speaks, for example, of a "change in a sensory surface," it is not at all clear whether he means a change in the characteristics of the sensor, leading to a new relation between stimulation and sensory signal, or whether he means simply a change in the physical stimulation that results in a change of the sensory signal, with the characteristics of the sensory surface remaining the same. Likewise, when he speaks of an effector causing changes at the sensory surface, it is not clear whether he means simple sensing of an effector's immediate physical effects, or the long-term changes that might result as a sensor adapts to a new relationship of the effector to it.

Much of the ambiguity in Maturana's conceptions arises from the attempt to deal with the subject-matter at an extremely high level of abstraction, one that is far removed from the ideal of a "generative model" (his term). At this level of abstraction, one tends to overlook the fact that some valid instances of generalizations are false. For example, the statement that the nervous system is closed, so that the relationship between effector and sensor is like that occuring at any other synapse, appears simply to be false. The relationship is *not* like that at any internal synapse, because the sensor can generate signals, and its generated signals can change, without any action or change of action from any effector. This is our principal evidence that an external world exists. It is also the reason why organisms must produce actions in order to control what happens to them. Organisms are affected by external agencies that are indifferent to the organism's behavior, and that cannot be directly detected. Thus organisms must learn what actions to take that will stabilize their inputs despite these external disturbances. There is no mention of this problem in Maturana's discussions. The nervous system is declared to be closed without any evidence, indeed contrary to what evidence we have.

I suppose that because I have taken a lot of care to make the control-system model generative at least in principle and often quite literally, I object to any proposition that offers no hope of experimental test. If all one needs to show is a computer screen

with some dots making patterns on it to claim that one has modelled an organism, why have I had to spend so much time predicting exactly how a real person will move a real control handle?

My objection, I hope, is not really that pettish. I am saying that the Maturana/Varela approach is simply not experimentally oriented: it is not designed to make quantitative predictions of specific instances of behavior. It is not a scientific model, vulnerable to the awful scrutiny of other experimenters. It is a complex metaphor. In fact, it does not try to penetrate beneath the level of observation by proposing underlying mechanisms that will lead to the behavior we observe. It goes in the other direction, by proposing abstract concepts of which observed behaviors are only instances. That is not the direction that leads to a generative model. A generative model would propose a mechanism for making distinctions. Conceptual strokes cleaving an abstract continuum do not constitute mechanisms.

Practical epistemology

People who spend a great deal of their time building models, and I am certainly one of them, have a tendency to get their priorities reversed. This tendency is perhaps easiest to see in a physicist. Consider the theory of electronics. This theory is built on proposed relationships among entities such as electrons, voltages, and fields. When the physicist shows you a dot on the face of a cathode-ray tube, and moves it by holding a magnet near it, he will tell you that this dot is the place where flying electrons are hitting the phosphor. He will explain that the horseshoe-shaped object he is waving around doesn't have any influence on the dot itself: it really carries a magnetic field, and electrons are really deflected by this field on their way to making the dot.

The key word here is "really." When the physicist says "really" he should be saying "hypothetically." What is "really" happening is that waving the horseshoe-shaped object near the CRT reliably makes the dot move. The physicist *imagines* that the horseshoe carries an invisible magnetic field, and *imagines* that the dot is made by electrons, and *imagines* that

the electrons are deflected by the hypothetical field.

The problem is that the models of electronics work too well. They work a lot better than our informal expectations work, and they supply missing links in the cause-effect relationships we experience directly. The fact that we can't actually observe these links is a nuisance, but who can doubt that those links are "really" there?

As a result, it's easy for the physicist to get the idea that physical models are more real than the experiences they explain. More generally, reductionists who claim that all of behavior is really just a matter of physics and chemistry, and will some day be explained that way, are actually asserting that all of behavior is imaginary. This is what comes of adopting the point of view of a model, and forgetting that one is doing so.

Real reality is simply what we experience. That is it, that is all there is to experience. If we can tell our models to shut up for a moment, this will become obvious. Even the physicist has to admit it. No matter what the model says, the ultimate test of the model is to make it predict something in direct experience that we need no model to observe—a spot of light, a meter reading, a coincidence on a scale, an image in a microscope, a line in a spectroscope. If the effects of the model are just as hypothetical as the model, we don't have a model because we can't check it against direct experience. The ultimate authority is always direct experience, the real reality we are incapable of doubting, because before we can even begin to doubt it, it is there. The only point of a model is to provide a hypothetical link between one direct experience and another.

This brings us to models of brains and nervous systems. It certainly appears, from what we know of neurology and the physics of the world outside the brain, that the world we experience is a set of neural signals. Maturana has chosen to accept the neurological model and omit the physical model, but his conclusion about experience is nearly the same as mine. However, this conclusion has a serious problem.

The problem is this: the world of experience does not resemble what we think of as neural signals. There is no difficulty with understanding how, at least in principle, neural computing networks could create signals that covary with aspects of perception such as size, orientation, color, sequence, relation-

ships, and so on. The difficulty arises when we consider those hypothetical trains of impulses and compare them with the world we experience. Blips on an oscilloscope do not look like the objects of experience, even though the model says they should.

This disparity tells me that there is something wrong with the model, not that direct experience is mistaken. A correct model will tell me why it is that a cube looks the way it does. It is perfectly obvious to me that neither my model nor Maturana's can tell us that. There is still some major way in which we are misconstruing the process of perception.

The nature of the problem may be something like this:

When we observe the nervous system, we are of course observing someone else's nervous system: our own remains invisible. Furthermore, we are observing *with* a nervous system. When we insert a probe into another person's nervous system, the electrical fluctuations are carried into an instrument where they are displayed on a two-dimensional screen or chart, where by convention one dimension represents time and the other amplitude. We look at this display with our human eyes, which presumably represent the display to our brains in terms of exactly the same kinds of impulses that are depicted on the screen or chart. We think about the display by making more impulses race around.

Now spatial dimensions, time, and amplitude are clearly concepts. Concepts are one of the things that brains generate. They are not aspects of the actual brain, but of what the actual brain is doing. They do not represent the brain, but only its activities. So we are trying to represent the brain itself, the thing that supports and generates all these activities, in terms of the activities. It's as though we were trying to explain what makes a radio work as it does, using as elements of the explanation only music, commercials, news announcements, and weather reports. That clearly won't work. To explain how the radio works we need a theory like electronics, that deals in the relationships and functions of the parts of the radio without saying anything at all about what station will be tuned in or what will be received from that station. Once we understand the principle of the radio as it applies to *all* possible broadcasts, we then understand how it works for any specific broadcast.

But that kind of understanding cannot be found in the material that is broadcast.

This would seem to stymie us. What do we have to work with, in trying to understand how the activities of the brain come about, but those activities themselves? There has been vague talk about how "recursion" will rescue us from all dilemmas of this kind, but I have seen no resolution arising from that idea.

The truth is probably that we do not know enough to arrive at a model in which predicted perceptions look like real ones. It could be that neural signals themselves are a smokescreen: after all, the only function of neural signals is to make chemical concentrations in one place depend on chemical concentrations a long way away: perhaps we have been looking at a trivial aspect of brain function. But I think we will eventually find that the answer is much more interesting than that.

Needless to say I can't describe this more interesting answer. All I have to offer is some concept of what we should do in order to get to the point where an answer begins to look more findable.

I suggest that we accept what the current brain models seem to tell us about perception and action, and make the best use of them we can. I think the best use lies in control theory, because at least this theory attempts to reconcile neurological models and physical models, and control theory has one feature that is lacking in Maturana's conception. It is testable.

Control theory does explain a great deal about ordinary behavior, as experienced in other people and also from inside. It makes predictions about the effects of experimental manipulations, and in the simple cases we know how to set up, works better than any other theory of behavior ever has worked. It permits quantitative predictions in detail. If we stick with the time-honored experimental approach to science, always trying to verify our conjectures by predicting real experiences and seeing if we can make them happen as expected, we will ultimately come across something that utterly baffles us, perhaps even that destroys control theory. That is what we can hope for, in any case. When a theory that makes logical sense, that predicts behavior with laboratory precision, that fits the physical evidence of underlying processes that we know about, fails

—that is a major development. And if we are to judge from history, the manner of the failure will be a message telling us where to look next.

This is my bias: a theory that is not constantly checked against experiment—meaning against direct experience—will do nothing but entertain us, at best. We do not live in a hypothetical world, but in *this* world. In this world we will find the ultimate tests of our ideas. Either our ideas will lead us to expect experiences that do in fact happen, or they will not. That is the only way to judge the worth of any idea.

So my practical epistemology is really no epistemology at all.

[1987]

Control Theory: The Road to Utopia

When I first heard that this was to be yet another theme meeting I raised some churlish objections, muttering that people go to scientific meetings to report on their own work, not somebody else's. Naturally, my mind then insisted on returning to the topic of utopias and I finally realized that I no longer could think of anything else to talk about. The truth is that any theoretician who dwells on human behavior thinks that if everyone just believed in the right theory (guess whose), the result would be a vast improvement in the human condition. At least I don't think you will hear many theoreticians telling you that the road to utopia is the one we're on now.

Since the subject has been so conveniently raised, I would like to tell you why I think that control theory could take us a long way toward a more ideal society. But before I do that, there's a problem that has to be dealt with in this meeting of cybemeticists. The problem is that most of you don't know what I am talking about when I say the words "control theory."

Or perhaps I should say that many of you think you do know what control theory is about, but have something in mind that is very different from the conception I'm involved with. The biggest problem is this word control.

A few meetings ago I was asked by a rather bristling young woman why I thought that controlling people was a good thing to do, or words to that effect. I was too dumbfounded to give a coherent answer, and even if I hadn't been so shocked,

Distributed at the American Society for Cybernetics Meeting, December 1987, at Urbana, Illinois.

all I could have thought to do then would be to say again what I had been saying for 30 years, and that would have taken too long. Now I've had a few years to think about this question, and I think I can give the right answer.

What I should have said is that nobody would think that controlling people was a bad thing to do if they didn't believe that this process called control was a real phenomenon. You might think someone was foolish for trying to do something that can't be done, but you wouldn't object any more than you go around objecting to objects that fall upward, or hot ice. So even objecting to controlling people is an acknowledgement that people do this thing called controlling, which is different from just moving your arms and legs aimlessly.

Control theory gives us a systematic way to understand the process we call controlling. It was only a quirk of fate that control theory was developed by electrical engineers instead of by psychologists. If biologists and then psychologists hadn't decided that organisms are incapable of controlling anything, they could easily have derived the mathematical relationships of control theory. Then the idea that machines can control things would probably have been dismissed by servomechanism engineers as psychologizing. As matters stand, psychologists tend to reject control theory as a machine analogy, or mechanizing, but that is wrong, too. Control theory is just a way of dealing with a phenomenon.

The phenomenon of control is easy to define in the context of organisms. An organism selects some external condition that is to be brought about. It then acts in whatever way is required by the current state of the outside world, and brings that condition about. If the condition is to be a continuing or repeated one, the organism continues to vary its actions in any way needed to maintain or repeat it.

That's the external appearance of control. I don't think we can doubt that it happens. This hall where we are right now was conceived in a human mind, drawn on paper, and brought into being by human actions. This was done, no doubt, in spite of all sorts of unpredictable (or at least unpredicted) environmental disturbances such as strikes or illness or running out of nails. We can be sure that nobody was surprised to see this building taking shape, particularly not the

architect. It was put up on purpose.

The question that the control theorist tries to answer is "How does this process of doing things on purpose work?" This is why you've seen me, at previous meetings, playing with rubber bands and buckets of beans, and making people sit in front of computer screens doing tracking experiments. I'm not a rubber-band or bean fetishist, nor do I think that sitting in front of a computer wiggling a control handle is a very interesting or aesthetic occupation. The point has always been to show that these phenomena are both real and peculiar, because what seems to be causing the person to act is at the very same time an effect of the action, and that the effect, not the action, is under control. Cybemeticists have been trying to understand this circularity of causation for a long time, as long as I've been associated with them. For some reason they've been trying to understand it without using control theory, so I've been trying, somewhat hopelessly, to persuade them to start using control theory and do it the easy way. Control theory was designed to handle this very situation, and it was mostly designed before Norbert Wiener ever wrote his book. It may have been designed by people from the wrong side of the tracks, but that shouldn't bother cybemeticists who can understand abstract principles.

Now I can start working this introduction toward today's topic. Once you understand that controlling is something people actually do, you can begin asking how a person has to be organized inside in order to carry out this process. This leads to a model of human organization. The model may or may not be correct in every detail, and I would favor "may not," but there are some aspects of it that seem pretty firm. In order for control to happen there must be both perceptions of whatever is to be controlled, and internal reference signals that specify what states of those perceptions are to be achieved through action. I won't go into the details here. Concentrate on those internal specifications, which work very much as the blueprint of this building worked. Nobody would claim that a blueprint is a building, yet it was only through continual comparison of measurements and relationships seen in the actual building with measurements and relationships in the blueprint that the structure kept moving toward one fixed final form.

The blueprint wasn't a prediction or a probability: it was a specification showing, in symbolic form, the intended state of the piles of building materials. That is how a reference signal works: it is an inner specification, in the language of the brain, of a future state of the perceptual inputs, the inputs that tell us the current state of the external world. We achieve goals by acting continuously in the direction that brings current perception closer to the specified reference perception. We vary our actions as disturbances come and go, as needed to diminish the error and then keep it small. The heart of control theory is an understanding of how a fixed organization can accomplish this task, an organization that we call a control system.

Where do reference perceptions come from? Well, where did the blueprint come from? The answer is that the blueprint came out of the architect's imagination and training. That is where all reference signals come from: from internal processes that, at high enough levels, we call by names such as thought, reasoning, intuition, creativity, and so on. They come from higher-level systems.

I'll talk about those higher-level systems some other time. For now the important subject is not where reference signals come from or how high the levels go in a given species, but where reference signals do not come from. They do not come from outside the person. All that enters a person's nervous system is perceptual information, and this information tells the brain only how the world is, not how it is to be. Don't get side-tracked by epistemology; we don't care here what this word "is" means. Perceptions describe the current state of the world that is known to the brain. Reference signals describe the target state for at least some of those perceptions.

This means that nothing outside a person can directly determine that person's inner goals. I think we all know this from personal experience. This is one of the major discoveries of childhood, a discovery that we keep mostly to ourselves because we can see that adults don't seem to believe it. Nobody else can know what we're thinking, and nobody else can stop us from wanting what we want. Later we allow ourselves to be enticed into games in which it seems that people can manipulate each others' goals, and we may even become philosophers or psychologists or industrialists who think they have learned

the secret of manipulating other people in this way. But privately we know that we do it all ourselves even while we are blaming somebody else.

People who think they have learned the secret of manipulating other people's goals are the reason for the question asked by that bristling young woman. You see, control is the very essence of behavior; it is the process by which we do anything at all. But the human race knows how to control in the same way it knows how to digest: by just doing it. It wasn't possible to understand the process of control in any theoretical way until sometime in the 1930s, because until then the puzzle of circularity had not been solved. You can't blame people, therefore, for being a little confused about this process, and particularly you can't blame them for not realizing that there are some objects in the environment that can't be controlled: human beings and other organisms. You can't blame them, but that doesn't mean you have to like the results of their ignorance.

I say that organisms can't be controlled, but you all know of instances in which they can be controlled. Ask, however, what it takes to control an organism. You can control a cat; you can make it do anything you want it to. But only if it's dead. Cats are a good example because they are well-known for their militant autonomy. Cats are small creatures; even a child can overwhelm a cat. Yet we think of cats as uncontrollable. Why? Because when we speak of controlling another organism, we almost always mean doing this without cruelty and without the use of lethal force. We want the cat to want to do what we want it to do, so it ends up doing it all by itself. But a cat has a way of making it perfectly clear that even when it does what we want, that's only because what we want happens to agree with what the cat wants. The only reliable way to control another organism is through the use of physical force that the other organism can't resist. So that is in fact how it's done, which is why the young woman asked the question with a bit of hostility.

Dogs are different, because dogs are manipulators. Dogs do what we want them to do, because that gets them what they want. Among other things they want a lot of love, and many of them assume that we do, too. Of course there are fawning

cats and independent dogs, but we're talking stereotypes here, usefully. Dogs, stereotyped dogs, teach us a lesson that cats haven't learned. The lesson is that control systems basically don't care about their own behavior. When I throw a stick for a dog to chase, I do this because I haven't any better use for my arm or the stick, and I want to see a happy dog bounding back and forth giving me strokes. Nothing wrong with love, it's a basic relationship. We can control other people's behavior, as long as doing this doesn't interfere with anything important that the other people are accomplishing with their behavior. If I want you to open a door for me I can ask you to do it and you will, unless both of your arms are full of groceries. Then you say with a little tang in your voice, "Take your hand out of your pocket and open it yourself."

Don't confuse controlling behavior with controlling goals. Goals are what people care about. Goals are nothing but what people accomplish by their behavior, which includes trivial things like opening a door and important things like opening a business. Most of the time we're so good at achieving the little goals that we don't even separate the action from its effect: we just say we "do" things. How do you open a door? I just open it. How do you turn the knob? I just turn it. Those aren't really "how" answers, of course, because they describe the effect instead of the cause. People have an intuitive understanding that the effect is the doing, not the actions that bring it about. But people really don't understand control very well, even though they're good at it and do it all the time. If they had a better understanding of what control is and how it works, they would do many things differently.

Now at last I have finished the introduction and can give you my talk on the subject of control theory and utopia. Utopia is a society run without significant conflict among or inside the individuals in the society. Conflict is the inevitable consequence of control systems trying to control each other's goal-achievements. When all people understand how control works, they will see that controlling other people's goal-achievements is impossible, has always been impossible, and has always led to dangerous conflict when seriously attempted. They will then stop trying to do the impossible, and get on with whatever it is in life that they find pleasurable and useful

to accomplish, by whatever means they find efficient, graceful, and acceptable in terms of their whole structure of goals, which I think is pretty much their own business, and not something for a utopian theoretician to prescribe for them.

Thank you for your patience in sitting through this long talk on utopias and control theory, even though the introduction was somewhat longer than the talk. I hope that this time you have understood what control theorists are trying to accomplish, even if perhaps you still don't understand how they hope do it. Sometimes control theorists don't understand that either.

[1987]

On Violence and Aggression

A group Statement on Violence reprinted in the June 1987 issue of this Newsletter presents a case against the idea that violence is a scientifically-established property of human beings. Manson and Wrangham, in partial rebuttal, remind us of the evidence that at least some aggression may have genetic bases. Hinde, one of the original authors of the Statement, rejects the idea that behavior can be explained only in terms of self-interest (as suggested by Manson and Wrangham), and recommends, in the interest of world peace, more research. I recommend, in addition to both points of view, a new theory.

I would like to introduce the control-system theorist's view —an outsider's view without a doubt—to this discussion, by proposing some concepts that take neither side, but suggest some new possibilities. If Frans Plooij's interpretations of my theoretical notions are any indication, ethologists should have no difficulty in seeing the points I am making and improving on them.

Suppose, for the sake of argument and to bypass rhetoric, that the basic principle of living systems is found in their ability to achieve control of what happens to them. The lowliest organisms, single-celled creatures, do this modestly well, and adult human beings do it superbly (if not always successfully). The most successful organisms control the physical variables that affect them in the most immediate and important ways.

They do this in the manner of control systems: systems that

Unpublished comment on articles in the Newsletter of the Society for Human Ethology, Volume 5, Number 2, June 1987.

can vary their actions as required to make the world they experience match inner criteria at many levels of organization. The question of how systems have to be organized to behave this way is a technical one that need not concern us here, except to note that the study of such systems is now about 50 years old. We are not speaking of anything metaphorical or poorly understood (except, perhaps, in the behavioral sciences).

The ability to control is the ability to sense some state of affairs, and to act to bring it nearer to some internally-specified and adjustable reference state. Again, the mechanical details of how this can happen need not concern us. What matters is that the more skilful a control system, the more energetically it will resist disturbances that cause deviations of the world it senses from the world it intends to sense: also, and because of exactly the same organization, the more energetically it will act to alter the local environment when its inner criteria change.

The actions of control systems control what they perceive. But those actions also disturb the local environment in many ways. If there are other organisms in that environment, their control processes may be disturbed; being control systems, those other organisms will resist. They will alter their actions as energetically as required and as precisely as they can to maintain what they are experiencing close to their own inner criteria. The result of mutual disturbance between control systems is conflict, and all too easily, violence.

I am happy (as well as constrained by my knowledge) to leave the implications of these relationships for ethologists to work out and apply. It seems clear to me that there is a close relationship between many of the appearances we call "aggression" and the normal processes of control. Violence is an arbitrary point on a scale of action that ranges from mild corrections to all-out striving. It is not a characteristic of an organism, but a byproduct of control behavior in the presence of disturbances. I suggest that it would be worthwhile for ethologists to learn the properties of control systems. Aggression and violence may be the natural outcome of one organism treating another as if it were a movable obstacle rather than an independent goal-driven system. When we can see aggression and violence as skilled control processes combined with igno-

rance, the phenomena of conflict may prove more understandable, without requiring us to believe in inherited traits or to bemoan our fate while submitting to it.

[1988]

The Good, the True, and the Real

Cybernetics is suffering from a case of overapplied epistemology. To show what the problem is I will indulge in some exaggerations, but they do not go beyond the uttermost boundaries of actuality.

Two primary epistemologists of cybernetics, von Glasersfeld and Maturana, have informed us that experienced reality is constructed (Maturana adding, "... in an informationally closed nervous system"). I have added my share to the confusion. The result has been an enthusiastic reception by those who find the restrictions of mathematics, logic, and scientific experiment distasteful. These enthusiasts have taken the conjectures of the philosophers and made them into a new—but not really so new—legislated non-reality for cybernetics, a sort of inverted reality.

In this new reality, the old reality appears as a cartoon strip in which white-coated ex-nerds, now grown to terrifying stature, conspire to convert us into obedient cyborgs chained in front of television sets while our exposed brains are basted with the poisons of technology. Art, music, literature, and Bambi are compressed by a gigantic technological foot into data banks as the System fills our vacant minds with pi calculated to an infinite number of decimal places.

Against this threat, the New Epistemologists stand arrayed in aggressive permissiveness. The rules of this permissiveness are very strict, and often entail inexpertly-hidden political agendas. No longer may we say that some ideas are right and

Unpublished manuscript.

others are, if not wrong, noticeably less right. Adjectives like objective, real, and factual arouse derision and finger-pointing, which are met not with a hot retort but with apologies —excuse me, I forgot for a moment that we don't believe in such things. Consensual domains need no consensus, and the domains have no boundaries. There are, in fact, no domains at all, because anyone's construction is as good as anyone else's. Everyone has become a Watchbird watching Me watching the Watchbird, and the Central Committee is an eye that belongs to none of us floating over it all.

There's a certain haunting familiarity about this collectors' approach to realities, harking back to a time when colorful new universes were a dime a dozen, with a little help from your friends. More likely the ultimate source is the fairy tale and all the other unkept promises of childhood. We all weep for those unkept promises, those visions we had to leave behind. We are all disappointed; we want it all back. You will notice that I have started saying "we" instead of "they." We apply a soothing epistemology to our hurts, and pretend that there is not just the one tiresome, complicated, confusing, dangerous, deteriorating Boss Reality. We want the freedom to start over with "Once upon a time..." and make the time now.

There are, of course, good and human reasons for the desire to find something other than the Boss Reality. Technology, just like most other tools used in human affairs, is being handled badly. We still have war, poverty, and corruption: inequities and iniquities. We have an educational system apparently designed to discourage everyone from getting to the interesting parts. We still denigrate women, blacks, spics, greasers, wops, slopes, sheenies, polacks, bohunks, krauts, limeys, frogs, commies, capitalists, nerds, freaks, yuppies, fundamentalists, moral relativists, liberals, and right-wingers, in a frenzy of collective self-hatred. We arm for peace. We buy votes for democracy. We develop resources until they are gone. We do many foolish things.

The New Epistemologists claim rightly that science and technology aren't doing us much good, but they are wrong when they conclude that science and technology are doing us bad. Actually, science and technology are in rather good shape. That part of the Boss Reality, which the human race has

been patching together and polishing for several centuries of millennia, is now capable of doing almost anything we can agree on doing. We have, I think, the technical means to take care of everyone's needs and presently conceived pleasures and still have something left over for the future. If it were just a matter of developing and applying science and technology, the five billion people on earth would be living in peace on Easy Street.

The barrier in the way is our abysmal ignorance of human nature. We can send food to Africa, but political factions turn back or destroy the convoys. We can present hundreds of channels of information at four million bits per second simultaneously to billions of people at a time, but hardly anyone can think of what to say into that network: we can watch wrestling or learn that much wheat is grown in South Dakota. We stockpile food so it is safely out of reach of the hungry, and we berate the disheartened poor for their apathy, while withholding buying power from them lest they spend it foolishly and give it all back to us. We tell each other that we must become better, but nobody can say how we are to do this.

While art, literature, and music poignantly reflect the human condition, they have failed utterly to do anything to improve it. They depict and describe and remind us of all our sorrows, our moments of triumph, and our hopes, but the evidence is clear: they contain no remedies that have the kind of effect we need. Struggling and shouting, strumming and expounding, we move inexorably down the slope toward the Slough of Darwin. Individually we struggle upward, but the crowd is too thick and too blind, and while we try to move forward we lose ground.

The sciences of behavior have failed; the social sciences have failed; religion has failed; mysticism has failed; politics has failed; magic has failed; art, literature, and music have comforted us but they, too, have failed. We are doing something wrong, we're somehow missing the point, we can't seem to get organized to deal with human nature as we have learned to deal with inanimate nature.

Before enlightenment comes despair. It is time to despair. It is time to step back and look at all the ways that have been tried in our attempts to cope with each other and say to our-

selves "It's not working." We have to see that while the physical sciences have been trudging steadily ahead for 350 years, the sciences that deal with human nature have been moving in circles, unable to break out of the frames of mind of the Scholastics, or trying to mimic the moves of the physical scientists in the hope that something wonderful but undefined will result. We have to look at all our vast collections of knowledge about human behavior and admit that they consist mainly of disconnected anecdotes that are supposed to be illuminating but aren't.

The New Epistemologists have misunderstood. There is a Boss Reality. We discover it by being extremely careful and extremely honest as we investigate the regularities of experience. Our capacity to invent realities is only a tool, and when we apply that tool as it is applied in the physical sciences, we discover that something invisible is always acting to steer it back toward grooves that run through nature. To use this tool properly we must grasp it lightly and be sensitive to where it wants to go: if we seize it with both hands and push it where we think it ought to go, it will indeed obey, but then we will not feel the answering pressure, the reluctance of the Boss Reality to be forced into an unnatural shape. When we use the invention of realities in the right way, we find that the picture wants to fall into certain forms that convey to us beauty, orderliness, simplicity, regularity, joy. When we use it in the wrong way we discover that maintaining reality requires endless effort and awkwardness, and that the result is ugly.

When we use the creation of realities in the right way, we discover not the nature of the objective world and not the nature of human being, but the true outcome of being human in a real universe. It is our own nature that we find, but at the same time it is the human reflection of a different reality, one that we can never know directly.

The physical sciences began their modern era with simple investigations of simple things. They began when observers began to let natural processes go as they would, and allowed what they saw to shape their understanding instead of the other way around. All the great thoughts of all the great philosophers then counted for nothing; they were only distractions from the attempt to focus on how this little ball rolled,

flew, and fell. That little ball told us, in the end, about the great ball of the earth, and eventually about the great ball of space and the fireball in which, perhaps, it was born. Science did not come out of rhetoric and argument, reasoning and explanation. It came out of silent watching. I suggest that it is time to fall silent, and watch.

[1988]

A Manifesto for Control Theorists

I. Control theory was first offered as a model of behavior because artificial control systems seemed to act in some ways like organisms.

A. In the form of regulating devices they could maintain variables like temperature, pressure, chemical concentration and composition, flow, position, sound level, and almost anything else measurable in stable predetermined conditions, despite unpredictable disturbances.

B. In aircraft, autopilots could keep airplanes flying level and on course at constant altitude, despite changes in weight distribution, wind direction, engine efficiency, turbulence, and air density, automatically—meaning without outside direction of the details of their actions.

C. In homing devices they showed characteristics of goal-directed or purposive behavior, directing themselves toward distant moving or even evasive targets.

D. In general, they showed the properties that William James called characteristic of life: the persistent and repeated reaching of constant ends by variable means. Control systems behaved in ways that the traditional life sciences had proclaimed impossible for any entity in a world governed by physical laws. Yet they behaved like organisms.

Distributed at the Control Systems Group Meeting, September 28-October 2, 1988, at Kenosha, Wisconsin.

II. Investigation of human physiology and function has shown that the parallels extend to details of structure and action.

A. Structures playing the part of sensors, comparators, and effectors in living systems are connected in the same functional arrangement as in an artificial control system.

B. Feedback loops are found in all parts of living systems from the biochemical level to the brain. The feedback is always negative, and the set-points are always adjustable.

C. The methods used to analyze artificial control systems, when applied to human action, reveal characteristics that are highly reproducible and stable in a given person.

D. When control-system models are used for quantitative prediction of simple motor behaviors, they do so with an accuracy that is unprecedented in the behavioral sciences.

E. The same methods can be used to measure the properties of human control systems when the consequences of motor actions are as complex as we know how to devise. The same accuracy of prediction is found.

III. The control-system model brings a new concept of mechanism into the sciences of life.

A. All traditional models represent behavior as a function of antecedent events. When these models are expressed mathematically, they amount to curve-fitting, finding a mathematical form that will express the law of dependence. Thus a single equation is used to characterize behavior.

B. Among traditional approaches, only that of operant conditioning takes into account two connections between external events and behavior: the "reinforcing" effect of an external event, which modifies the behavior through action on the organism, and the "schedule of reinforcement," which makes the reinforcing effect a consequence of the manner of behavior. Other fields have considered closed-loop relationships (the

law of effect, response chaining, stimulus-stimulus or response-response laws, or "interactionism"), but not in the systematic way of the radical behaviorists who follow B. F. Skinner.

C. The control-system model provides the analysis that the Skinnerians did not discover, by introducing the simultaneous solution of equations representing both relationships between behavior and external events. One equation describes the organism's response to inputs, and the other describes the simultaneous dependence of those inputs on the response. Only when both relationships are specified (even if one is conjectural) can the variables be predicted and the model be tested.

D. Under control theory, both stimulus and response become dependent variables: neither "causes" the other. Both behave in ways determined by the entire organism-environment system of relationships.

E. The critical independent variable is now found inside the organism in the form of a variable reference signal. This signal determines the equilibrium point of the whole system, its goal or purpose. This kind of system controls its own inputs by varying its outputs. It is not controlled by the external world. Nevertheless, its operation is completely consistent with all known laws of physics and chemistry. It is a kind of mechanism unknown to those who formulated and still support the ideas of traditional behavioral science.

IV. The control-system model represents a complete revision of the traditional concept of the mechanisms of behavior, and of life processes in general.

A. By showing in quantitative detail how purposive or goal-directed systems work, control theory makes obsolete all previous schemes that were devised to substitute a cause-effect process for purposive processes.

B. By providing a quantitative analysis of the relationship

among stimuli, actions, and independent disturbances, control theory shows that apparent stimulus-response relationships result from actions that stabilize controlled variables against those disturbances. The apparent stimuli are not stimuli, but disturbances of controlled variables.

C. Relationships between actions and external events that seem only statistically detectable under the old paradigm become highly systematic and predictable under control theory.

D. Control theory shows that the methodology called "scientific method" presupposes its own conclusions, and discards all data relating to controlled variables. The causal relationships found in this way are almost entirely illusory.

V. Control theory represents a new beginning for the sciences of life. It is not an extension of traditional ideas, nor was it built on the base of older theories about behavior. It is a total replacement of those older ideas by a radically new concept. There can be no bridge from the old to the new: the new approach diverges from the old at its roots. The entire path of the life sciences from the time of Descartes to the present is an evolutionary track that must now end in extinction.

A. If they wish to be maximally effective, control theorists must not try to find suggestions of compatible ideas in older thinking. That generous impulse has the effect of slowing and obscuring the changes that must occur, and always encourages attempts to merge the new back into the old in order to justify the old. The result is always incompatible with control theory.

B. The motive for finding precursors of control theory in older ideas is to preserve the continuity of science, but the preservation occurs at the wrong level and in the end is a disservice to science. This implies that control theorists must break completely with older explanations of behavior: none of them contained the principles of control, and all of them, in the attempt to make sense of control processes without using control theory, propose explanations that are incompatible with control theory. Let others claim to have foreseen this

outcome if they wish. Let future analysis trace the evolution of control theory out of the grand course of science: they will not find its roots in the life sciences. For control theorists of today, none of that matters: there must be a new start.

C. Control theorists, if they are control theorists, must begin to pose and answer questions that relate to this new understanding of human and animal organization. The traditional categories of behavior are of no use. Under control theory there are no boundaries between disciplines: a control theorist is not a psychologist, a sociologist, a biologist, a cyberneticist, or any other such specialist. All areas of the life sciences come under control theory. The control theorist studies living systems in all their aspects. Out of this kind of study will come new disciplines that have no names today.

D. If it is agreed that control theory is a revolutionary idea, then let control theorists be revolutionaries. This does not mean being dogmatic or doctrinaire: it means speaking the truth as it appears to be but being ready to modify it for any good reason. The foundations of control theory are solid; anyone can recreate the basic ideas from fundamental principles at any time. To be a revolutionary means not compromising with ideas that are known to be wrong, and insisting that any challenge to control theory be as carefully formulated as control theory is.

E. Above all, being a revolutionary means undergoing those internal revolutions of thought that utterly change one's approach to the understanding of living systems; resolving the inner conflicts with older ideas that live on in every control theorist; re-examining every idea one brings to this revolution, with the conviction that no idea, however familiar or treasured, is worth preserving if it is wrong.

VI. *Today, Kenosha: Tomorrow, Racine!*

[1989]

After Galileo

My subject for this year concerns programs, principles, and system concepts. I will talk about programs of action that the CSG might want to pursue in the next ten years, the principles we might want to use to organize and justify the programs, and a system concept that might give us a reason for continuing to exist. I won't take these subjects in order, because as you all know that isn't how levels of organization work: they interact with each other and with the real world, and they reorganize in the process. Consider these ideas to be work in progress. You might even want to take notes when you hear something that suggests a better idea to you.

I look on the work of the CSG as being divided into two main parts: theory and practice. Of course our theoreticians are strongly inclined to experimentation, and our practitioners are hardly silent on the subject of theory, so this division is only functional, a matter of which hat you're wearing. I think we would agree that it is a theory that informs our practices, and our practices that give us observations that our theories must take into account, so everything is working as it should.

Theory and practice could keep one satisfied for a lifetime, but I am not satisfied. I want something more. I want what we do to lead to a revolution in the life sciences, a revolution that will change the way all of them go about understanding the nature of living systems. To be blunt, I want a real science of life to develop to replace what now exists, which is not science. I will have to amplify on that a little.

Distributed at the Control Systems Group Meeting, September 20-24, 1989, at Kenosha, Wisconsin.

Philip Runkel has written a book and told us about two methods that scientists use to study behavior: the method of relative frequencies and the method of specimens. While both methods have been used formally or informally since the beginning of history, the method of specimens has become workable as a way of understanding behavior only in the last 50 years. The method of specimens requires studying one individual at a time through making models of that individual's organization and testing them against real behavior. We're just learning how to use that method as successfully in the life sciences as the physical sciences have used it for 300 years or so. We'll get back to this method in a while.

Learning the rules

The method of relative frequencies pays almost no attention to individuals—either individual persons or individual instances of any natural phenomenon. It's the basis of empiricism; before Galileo it was essentially the only reliable basis for understanding most of nature. All it means is learning from experience. You watch what happens around you and remember it as best you can—or, starting about 10,000 years ago, you write it down. Then you try to figure out the rules.

If you plant a crop when the weather starts getting warmer, sometimes you get enough to eat and other times you have to clear out the frostbitten foliage and hope you can get another crop going and harvested before you starve. The times when you starve make you wonder: was something going on when I planted those crops that made the difference between success and failure? You entertain all kinds of theories, listen to the elders who have been planting crops for longer than you have, and try to remember. What was happening when you planted the crops, beside the warming that is too often misleading?

I don't need to recount all the kinds of theories that have been entertained. Not many of them worked particularly well. But through an incredible feat of noticing and record-keeping, the human race has discovered over and over that there are certain key events for a farmer to watch for: one of the most reliable is the rising of the sun over a certain place on the horizon, or the appearance of a star just before sunrise (which

is better because then it doesn't matter where the hills are). Somebody, let us say, had the theory that if you delay planting until you see a certain star just as dawn breaks, chances are that there will be no more killing frosts that year. That theory worked pretty well.

That's what most people mean by "a theory." It's a proposition that one thing leads to another. That's how you learn the rules when you haven't any idea what makes the world work as it does. You don't need science to learn the rules: all you need is accurate and honest observation, a good memory, and some way of sorting through the facts to find regularities.

That isn't science? No, I say, it isn't. We'll get to that.

After Galileo, the physical sciences went into a frenzy of development. The life sciences did not, except where they simply used physics and chemistry to study the insides of organisms. What we now know as the social and psychological sciences, or the behavioral sciences, kept right on trying to learn the rules in the same old way. The main developments in this process were the invention of statistics as a way of sorting out regularities, and the invention of systematic and controlled experiments to focus more efficiently on the relevant events. Now it became possible to show objectively whether or not one event led to another, without reliance on informal and fickle human judgement.

These developments did lead to an information explosion in the life sciences, or perhaps it would be better to say a fact explosion. Given formal experimental methods and formal methods of analyzing results, anyone could do an experiment and with a little luck find a provable relationship between causes and effects. The life sciences, insofar as they were not just transplanted applications of physics and chemistry, became a machine for the generation of facts. By and large that is what they are today.

Unfortunately, most of these facts are written down and immediately forgotten, because they are unimportant, uninformative, and worst of all, isolated. The average behavioral science paper probably isn't seriously cited even once by anyone but its author. This mindless accumulation of statistically significant but unimportant facts has become an end in itself, a sort of National Scientific Debt which, according to those who

justify this kind of effort, will be paid some time in the future when science finally is able to grasp the meanings of all these facts. In the meantime the facts accumulate in great drifts, reefs, and shoals.

I think we can say that the method of learning the rules —Runkel's method of relative frequencies—has had a fair trial. It has been developed about as far as it can go. And it has still not produced anything that will help the human race survive its own internal conflicts and its own ignorance of human nature. It could be carried on for another three hundred years without producing anything better than what we now have—except, of course, that if a real science of life is delayed that long, there may be nobody left in any condition to say that it's time to try something else.

Making models

Now, what is it that the physical sciences do differently? In Runkel's terms, the physical sciences use the method of specimens. This doesn't mean that physicists pay no attention to the rules; they start like anyone else by noticing phenomena and connections between them. But that's only the start, not the end-product. Once a relationship has been noticed, the physical scientist asks *why it is observed*.

The answer to this kind of why would probably be given by a physicist in terms of what he calls "properties of matter, space, and energy." To the physicist, the phenomena we can observe or measure with instruments are only the surface of what is really going on. Underlying gross observations is a whole world of interacting entities that we can only infer: mass, energy, momentum, entropy; electrons, neutrons, protons, and a huge family of finer particles; quanta and photons, electric and magnetic fields, curved spaces, influences that travel invisibly at the speed of light. The world of appearances rests on another world and can be explained only by reference to that underlying reality.

Until recently, physicists haven't been much given to epistemology. They generally assume that an entity like an electron really exists, because all their experimental tests for its existence are confirmed to the limits of measurement. In fact

most of them would say that the physical model tells us what is really happening in nature, whereas what our senses reveal is both superficial and misleading. If anyone suggested that it is the world of direct experience that is real, and the physical model that is imaginary, most physicists would tend to become rather hostile toward the messenger. As we've seen recently in following the "cold fusion" news, a hostile physicist can be a pretty nasty fellow (though most of them aren't —either hostile or nasty).

Students of human nature, who do tend to worry about epistemology, ought to find my view of what physicists do easy to understand. I say that they make models. Many physicists would say the same thing, especially modern particle physicists. A physicist tries to imagine what there might be underneath observation that would *necessarily* result in the phenomena we can see and feel. The key word here, and the key to the whole method of model building, is "necessarily."

When Newton was first trying to figure out gravity, there might well have been a time when he thought that the attraction between two particles might fall off as the inverse *cube* of the distance, instead of the inverse *square*. That's a model, because it says that the net forces and motions we can observe result from adding up the attractions between all the particles of a heavy object and all the particles that make up the Earth. How could Newton have tested such a model to see if it was right?

Obviously he couldn't measure the forces between every pair of particles, or prove by observations that those forces really exist. What he could do would be to deduce what would *necessarily* be observed if in fact the force fell off as the inverse third power of distance. He'd have to invent some methods of calculation (as he did), but then he could use them to deduce *strictly from the third-power property of this model* how an object would move in orbit if that were really the underlying property of nature. "Necessarily" means *necessarily in terms of the properties of the model*. It means that anyone, given that model and the publicly-defined methods of calculation, would have to arrive at exactly the same prediction.

That's not the same as asking whether the predicted behavior of the model is like the observed behavior of real objects.

The first question is not whether the model describes the phenomenon, but just what the model *does* describe. To have a model, one must be able to figure out what the model would do, operating strictly by its own rules, without any reference to what the real world does. What it would *necessarily* do.

When you know what a model must do according to its proposed properties, you then have something to compare with observation. If you can't say what the model will do—if anyone can't say what it will do, even not knowing what it is a model of—then you don't have a model. It's that simple. Models must behave.

Once you can say what a model must necessarily do, you can compare that with what actually happens. Newton, if he ever actually considered the third-power law, would have found that there can be no orbits in that model of reality that match observable orbits. So he would immediately have junked that model and gone on to another—say one in which the power is 2.99999, and so on until he hit exactly 2. Which shows, of course, that he did no such thing; the invention of models isn't that systematic, which makes the stunning successes of the method of modeling all the harder to understand. We make models better than we ought to be able to make them.

In the life sciences one will often come across descriptions and diagrams that are purported to be models of a behaving system. Many of these are simply mnemonics representing observed behaviors—this leads to that. We can discount those because they are just diagrams of the rules, and don't explain why the observed rules *have* to hold true.

In other cases we find models that really do seem to propose underlying structures or organizations inside a behaving system. Some of them get pretty elaborate. But with only a couple of exceptions I know about, these models fail the simplest requirement. Nobody, not even their inventors, could say what the proposed arrangement would actually do out of its own stated properties. The authors of such models never explain their diagrams element by element, stating the rule by which one part reacts to its inputs by producing outputs, and how (and by what means) the lines carrying those outputs to other parts of the model affect those other parts. Given only

such a diagram, one couldn't possibly guess what kind of organism it represented or what kind of behavior it was supposed to generate. Such models imitate some of the forms of real models while leaving out all the substance that adds up to a real model.

Diagrams make models easy to grasp, but aren't always necessary. The essence of a model is contained in the properties of its parts and the connections between the parts that establish organization. Such things can be described symbolically as well as being drawn. There are models in the life sciences that use no diagrams (except optionally, to illustrate relationships). But even though such models come close to being real models, they are too simple. They are referred to in psychology as "intervening variable" theories. They are uniformly based on the idea that the responses of organisms follow from stimuli (or "situations") that impinge upon organisms. We control theorists think that this is not the organization of behavior. But even granting that aspect of the model, the real problem (as other kinds of psychologists will readily explain) is that the simple introduction of an intervening variable offers no proof that some other intervening variable, or chain of other variables, isn't there instead. If the output from the intervening process is the behavior we observe, and the input to it is the stimulus we observe, then the intervening process is totally superfluous—it's just a way of representing the rule relating stimulus input to behavioral output. So a behaviorist would argue.

That, of course, would not be the control theorist's basis for objecting to intervening variables. There are two objections from our point of view. One is that the input is never in fact independent of the output; the closed loop is being ignored. But the other is more important here: the models usually proposed are not well-enough defined to produce any behavior by themselves. They can be matched to observation only after the fact. They are therefore useless, which is worse than being wrong. There's nothing necessary about them.

Control theory is, or leads to, a working model of specific individual behaviors, and by extension to properties that are shared by all human beings, or within all species. But control theory itself is subsidiary to the concepts that contain it, pri-

marily the principle of modeling as it is applied in the physical sciences. First one must adopt the method of modeling with all the demands for perfection that follow. Then one can begin asking *which* model works best. Control theory could not have arisen as a model of behavior unless there were behind it the modeling approach itself.

Control theory can be compared to other models that have been tried within the same general approach; among these I would include neural models (both digital and analogue) and computer models that try to imitate human mental functions. Such models can be compared simply by seeing which of them predicts the best. Some of the other models may well become part of the control-system model—for example the "perceptron" model, which is now applied, I would say erroneously, to modeling behavior, but which may prove an excellent model of certain functions within a control system, such as perceptual functions. There are many part-models that pertain to motor functions, biochemical functions, and genetic functions that could eventually become part of a more developed version of the control-system model.

There is, however, no point of contact between control theory and the non-modeling approaches that predominate in the social and psychological sciences. That is why I say they are not sciences. I do not say that control theory is a better theory than any proposed under the conventional approach. I say that it belongs to a different universe of discourse, the method of modeling rather than the method of trying statistically to deduce meaningless dependencies. I adopt the physical sciences as my criterion for judging success. And I explain that success by attributing it to the use of the method of modeling.

I do *not* attribute the success of the physical sciences to their choice of a simple or highly predictable aspect of nature. That excuse often given by behavioral scientists will not hold water. The models of physics are immensely more complex than the models (and concepts) used in the life sciences. Organisms may in fact be more complex than the phenomena of the nonliving world. But in that case, shouldn't the principles of explanation also be more complex? That is obviously not the case.

The next ten years

In considering the goals of the Control Systems Group, I suggest that we must shape them according to the method of modeling, and aim as high as that method permits us to aim. I think we should consider our efforts as aimed at nothing less than the development of a science of life. In our group are represented at least a dozen scientific specialties and several professions concerned with human behavior. We cannot cover all the specialties and professions that some day must be brought together, nor can we even do justice to every aspect of our own specialties. But we can try to do something useful toward reaching our goal.

We will not lack help. Most scientists, once they understand that an effort like this might have a chance of succeeding, will support it. Our main task is not to do all the work that needs to be done, but to prepare the ground by submitting our own ideas and works to the same criteria we propose that others should accept. We must demonstrate the result of doing this.

The method of modeling suggests that perfection exists. The physicist would put that a little differently: nature operates according to laws, and is only and exactly what it is. Therefore when we make models, or propose properties, we are approximating the real laws of nature. If we then look at the difference between what our propositions imply and what actually happens, we should be able to see how to make the approximations better—to make them come closer to a representation of What Is.

Maybe there isn't just one necessary model of reality. But we can go far by pretending that there is. Instead of just hoping that our models will predict correctly a little more often than they fail, we can raise our standards and treat every failure of a model as indicating that it needs revision. That's how the physical sciences got where they are. A few years ago I proposed to this group that we stop accepting any correlation less than about 0.95. I've changed my mind. Now I think the goal should be 1.000.... In other words, the point of modeling isn't to accumulate facts, but to improve models. The facts will take care of themselves, but the models will *always* need work. As long as we can measure any difference between what the

model does and what the real organism does, the model needs work. That ought to take care nicely of anyone who wants to form a "school."

Perhaps the most immediately useful project we could undertake would be to do ourselves what we are always demanding that other unfortunates do: criticize our own assumptions. No member of this group arrived but from some other way of thinking. No one is free of unexamined assumptions. Even I, who came from totally alien fields, often feel hopelessly contaminated. We all have axes to grind; these axes are the models we have accepted without having any reason to suppose they are right. To get anywhere, to understand what we are really asking of others, we must bring these hidden models to light and subject them to the same demands we place on those we propose formally. I am going to suggest that we try this as a group to see just what kinds of difficulties are involved, or more hopefully, whether we can come closer in this way to realizing our own ambitions.

I hope that within ten years we will have learned enough, done enough, and taught enough to be justified in founding a Center for the Study of Living Control Systems. I hope that this Center will continue and improve upon our tradition of non-exclusion, of welcoming anyone who wants to learn and judging no one on the basis of what they believed before they understood control theory.

I hope the Center will be a center, not a fence around a perimeter. And I hope that some day this Center will dissolve, for the simple reason that the disciplines in the life sciences have come together to form a single science based on successful methods, accepting whatever is left by that time of control theory. Then my work will be done.

[ca. 1989]

The Epistemology of the Control System Model of Human Behavior

In the paper in which this model of human behavior was introduced (Powers, Clark, & McFarland, 1960), the following statement appears:

> ... 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. (pp. 78-79)

In 1973, I published a book called *Behavior: The Control of Perception* (Powers, 1973). Reading "control system" for "feedback system," "perceptions" for "feedback signals," and "perceptual function" for "feedback function," the later usages,

Unpublished manuscript.

it can be seen that this title reflects the basic epistemology adopted from the start. Almost from the start. This continuity reflects not mere consistency, but necessity, for if behavior is seen as a process of controlling input, and if models of behavior are constructed in the attempt to simulate this kind of process, the result is to demonstrate this epistemological relationship between what a behaving system can experience and the world around it. The epistemology comes out of the model.

The human observer of such a system can see that inside the system only certain aspects of the environment are represented, precisely those aspects that come to be under control by the system, or are potentially controllable by it. The behavior of the system can be understood completely in terms of controlling the inner representation instead of the objective reality. To adopt the point of view of such a system is to imagine experiencing a world made up only of those perceptual signals it derives from its contact with a world that is otherwise unknown to it. In that world, behavior can be only control of perception.

As long as we allow ourselves the luxury of standing on a cloud from which we can see processes both inside and outside the behaving system, we do not really have any epistemological problem: it is simply a matter of engineering design and analysis to decide what kind of system can produce the sorts of behaviors we see. Epistemology isn't the observer's problem; it's the system's problem. For the most part the control system model I espouse is presented and analyzed in just that innocent way. This will continue to be the most practical point of view toward developing a model to represent human behavior and even experience more and more exactly.

The implied objectivity of this approach, however, contains a number of traps; these traps are no longer very dangerous since they are full almost to the brim of psychologists, neurologists, biologists, and the like, but it is still embarrassing: one's whole attitude toward science depends on whether one falls into or skirts these traps, and one's experiments, data, and interpretations are strongly affected.

The point where epistemology becomes crucial is where the observer becomes an experimenter, and begins to exert control

over his own perceptions, some of which are experienced as the very system under study. That system and its environment instantly become subjective impressions on the part of the observer. Of course this is what they were all along, but only action makes this fact obvious. As long as we only observe and analyze, taking in data passively and churning it through our internal logical processes, it seems that we are simply reporting the facts and deducing their implications. A description seems as good as a model, an inference as good as a test.

When we act, however, we discover the real nature of the models we have been half-consciously assuming. These models are at the bottom of our expectations. We do not know they are models until we apply some new action to perceived reality, and imagine applying it to the model at the same time—in other words, generate an expectation. What usually happens is that the model does one thing, and reality does something else. Suddenly the model is a model, and we can see the difference between what we are imagining and what we are perceiving.

This divergence tells us, or should tell us, that there is a difference between the world outside us and the models inside us. Generally the models have a great deal more detail than the experienced world. In the experienced world the scientist sees a few meter-readings and some knobs he can twist. In the model, he sees electrons, resistance, inductance, capacitance, charge, fields, entropy, information, neurones, biochemicals, and that whole vast array of conceptions that we call knowledge of the universe. The models and the perceptions share a few common variables, the ones we can perceive directly like meter readings. But to see what is "really" going on, the scientist turns away from direct perception and looks at this imagined model, for it is his only way of perceiving, or seeming to perceive, the real world that lies behind the unscientific world of direct experience.

We need to establish some epistemological position in order to deal with human behavior and experience, where the physicist and chemist can get away without one. We are modelling not just some external phenomenon, but ourselves, the modellers. If we build models of ourselves, then those models contain the functions needed to allow us to do that, or they

are not models of human beings. There must be some limits on how successfully and completely we can do this, but there is no point in worrying about them until we encounter them. If we can make a model of a muscle and a nerve, maybe we can make a model of a reflex. If we can make a model of a reflex, maybe we can make a model of a system that coordinates two reflexes. If two, then maybe 20. And so on, peeling off what we can model and making more explicit what we have yet to tackle. The trick is to find an epistemology that remains consistent with this process, without fooling us into thinking that we are back in the objective world and the omniscient viewpoint. How we might do this is the long-postponed subject of this paper.

Realities

In the Drug Decades, with which we have not finished, there grew up an epistemological position that became very popular. It can be stated simply: my reality is as good as your reality. With a little chemical aid, epistemology became laughably easy. What is reality? It is simply whatever pops into your head. The principal effect of the various chemical substances, as far as the philosopher is concerned, seems to have been instant destruction of the capacity to be skeptical about the propositions that constantly occur to any creative person.

I do not think that anyone's reality is as good as anyone else's reality. There are competent and incompetent realities, workable and unworkable realities, realities we can experience and realities we can only imagine. Notice that I do not say there is only one reality. If there is any positive result to be found from the the Drug Decades, perhaps it is the widespread realization that people construct many different worlds for themselves out of the raw material with which we all begin. But this diversity of worlds implies more than freedom; it also implies a serious problem, a scarey problem when you think about it.

Constructing a workable, competent, and above all communicable reality is a task that seems more impossible the more closely one examines what we start with. All the findings of physical and biological science conspire to tell us that we

are each alone, isolated inside a nervous system. To live in the world, and more impossibly, to live with other similar (?) people in a world, the brain must solve an unsolvable problem. It, with the rest of the central nervous system, is presented with a world consisting of millions of tiny signals, each telling the state of a single sensory receptor quite independently of the state of any other sensory receptor. This world contains no objects, no relationships, no sequences, no transitions, no logic, no persons, no science, no nothing —nothing but intensity, the report of how much of something is present. From that, the brain must create shoes and ships and sealing-wax, not to mention capitalism and communism. Impossible. Science tells us that science is impossible.

It is obviously not impossible. We have to accept that we have done it. Somehow there are links that allow our realities to converge, both toward each other and toward at least some congruence with physical non-living nature. Every day millions of new brains are brought into contact with this network of realities, and immediately begin doing what is necessary to poke, test, modify, hypothesize, revise, and converge toward the common, or roughly common, consensual domain (the felicitous term of Humberto Maturana). No epistemology can make any sense if it ignores these facts of experience. If science leads us to ignore these facts, then science makes no sense.

Clearly, in order for us to have some confidence about this convergence toward a reasonably common consensual domain, we have to consider more than passive analysis of experience. We can in fact construct any number of arbitrary perceptual realities, if perception is all we choose to consider. The unique contribution of control theory to solving this problem—while not destroying itself in the process—is the introduction of action as a phenomenon that is a priori in a different class from perception. Action affects the world, whatever it is. Perception represents that world to us, in potentially uncountable ways. But the combination of action and perception quickly rules out great areas of possibility, for when we act, what happens in perception is not what might happen, or should happen, but what does happen. We cannot predetermine that this action shall have that perceptual consequence. We can choose any action we like, and construct

perceptions according to any scheme we please, but when we act, it is something else that decides what the perceived result will be.

This is why we have to learn how to control what happens to us, instead of just knowing how. Whatever picture of reality we have constructed, its behavior is drawn from the behavior of all those millions of tiny intensity signals that science has told us about. And the behavior of those tiny signals results from the actions of something outside the nervous system. It isn't necessary to accept any particular scientific or intuitive picture of what lies out there. But whatever it is, our actions affect it, and it affects our perceptions. In that passage of effects out of us and back into us, reality gets into the act.

It gets into the act in two ways. First, there is simply the matter of which action affects which perception. Second, there are disturbances: our perceptions tend to change when we are not acting on them—in fact, that's the principal reason we have to learn to control them: they won't spontaneously come to the states we want, and they won't stay there unless we keep acting. Without these two factors we would truly have no choice but solipsism, however we struggle and squirm to escape it.

One epistemology

In 1953, Robert Kenley Clark and I decided to take on a project that seemed worthy of our efforts: revolutionizing everything (I was then 27 years old, which now doesn't seem very old). We both worked under the inspiration of Wiener and Ashby, from afar, and I added a working familiarity with control systems, which I had more or less understood for all of eight years from experience fixing them in the Navy. Bob Clark was a genuine physicist with a PhD; I had a bachelor's degree in physics. I don't know where Bob is today, and I wish I did. We began from the basis of knowing that the behavior of control systems (the obviously right model) depended crucially on their input processes, and much less so on their output processes. We therefore did much the same thing that the cyberneticists were doing in the Macy conferences, unbeknownst to us. We decided that to classify behav-

ioral organizations, we had to classify perceptions, which meant a study of language.

This took us immediately into what is known today as semantic network analysis, or perhaps even semiotics. We began looking for the ways in which the meanings of some perception-words depended on, were defined in terms of, other perception-words, an obvious start toward a hierarchy of control systems. This project went tremendously well; we built up quite a structure almost immediately, both of us being pretty glib. It made perfect sense, top generalization to bottom detail. But then one of us noticed an alternative classification, somewhere in the middle, a much better one. Of course this required altering everything that lay above it, which proved to be remarkably easy. Then, if I remember correctly, Clark and I each noticed two different possible restructurings at the same time, and developed the structure anew in two different ways, which we then managed to merge.

To cut the story short, two years into this project I looked at Clark and said, "Bob, let's not build another castle out of words." He instantly agreed. We had run into the same problem that librarians face with the Dewey Decimal System. Sure, you can divide all human knowledge into ten main categories, and subdivide the categories, and subdivide the subdivisions, and so on. Ten numbers or so do the trick. But then someone comes up with a problem. Is a book called *Specialty Restaurants in Europe and the Middle East* about cooking, or travel? The answer is yes, depending on whether you're looking for recipes or planning a trip. Semantic networks simply aren't unique. You can turn them upside down and inside out and they still make sense. The organization of words in language is simply not a useful model for the organization of perception. Two years down the drain, which seemed like a lot at the time.

Then, defeated by words, we started to think. Control systems control their own inputs (we had not wasted all our time). This means that all they know of reality is what their sensors tell them. It means that the organization of the sensors and any computing processes that combine their outputs (all analogue, then) determines what reality is going to look like to them, and what will apparently be there to control. Clearly, if we were control systems too, all this applied to us. In this way

we discovered epistemology, without enough sophistication for either of us to mention that term. I didn't even know what it meant. Perhaps the most important realization was that organisms don't control the names of perceptions: they control the perceptions themselves. And that proves to be a totally, totally different matter.

I think it was Clark who came up with the solution. When you have a really convergent pair of consensual domains going, it is sometimes hard to tell where an idea came from, and it doesn't seem to matter. The solution was to take a close look at subjective experience, and try to place all the contents of that experience into some structure that was usable. Here is what we arrived at.

First, there is Directly Experienced Reality. That is really all there is. It's what you can't doubt experiencing, because all you have is the experience. Nothing can be more real than that.

Then there is Deduced Reality. This is a subdivision of DER. It consists of all the statements you can make about experiences, all the opinions you have, all the doubts and beliefs, all the analyses, all the interpretations, all statements to the effect, "This is a that." It's simply that part of Directly Experienced Reality that we call our thoughts and opinions about reality. You can't be mistaken about being aware of thinking that the world is flat, although that statement, being part of Deduced Reality, can be judged in terms of truth or falsity within the whole network of knowledge. Science is part of Deduced Reality. So is epistemology, and control theory, and cybernetics, and the idea of consensual domains. Deduced Reality contains opinions like "That theory is right" or "wrong." There is no other game in town: if you want to play the science or philosophy game, this is where it happens. It's a subdivision of, not an alternative to, Directly Experienced Reality.

There's another subdivision of Directly Experienced Reality: we called it Constructed Reality. By this we meant not what von Glasersfeld means, and what I mean now, but Imagined Reality. This is the world that consists of perceptions that are generated completely within the brain: signals produced by some means or other that do not end up as actions, but which substitute for the perceptions that normally depend on exter-

nal events, whatever they are. The idea that there are external events is part of Deduced Reality. Constructed/Imagined Reality is where we plan, dream, imagine, and make models of the reality that we have no way to experience directly. Part of Deduced Reality says that Constructed Reality is produced by internal feedback loops in the brain. If you want to argue about that, you will be arguing in terms of Deduced Reality.

What I am describing now is the proposal that Clark and I came up with as a useful addition to Deduced Reality. It solved all the problems we thought were important at the time. It still looks good to me.

Look what we could do with this picture. Taking the standpoint of Deduced Reality, we could do physics, electronics, mathematics, engineering, and all those "objective" things freely. When we got into trouble, we could expand the horizon to include all of Directly Experienced Reality: clearly, if Deduced Reality contradicted Directly Experienced Reality (which, remember, includes all that is given in experience), we knew there was something wrong in Deduced Reality. We looked at paradoxes, the stuff Douglas Hofstadter likes to play with, and dismissed them as trivial. If you have a problem with a game, change the rules. There aren't any paradoxes. Long before we ever heard of Heinz von Foerster, we discussed self-referential sentences, and decided that there is no such thing: sentences can't refer to anything, much less themselves. Some function of the mind uses sentences as pointers, and if it gets confused about what it is doing, tough. It doesn't have to get confused, because it doesn't have to play that game. If you want to make verbal paradoxes, go ahead. Just don't expect to capture my attention.

Pretty cocky. But that is why we construct epistemologies and decide they're good enough for our purposes. There isn't, of course, any "right" epistemology. You can't get "meta" enough to exempt epistemology from being a part of Deduced Reality. We construct the notions that seem good enough for now, and when experience shows that they aren't good enough any more, we reconstruct them. Certainty, even about epistemology, just isn't in the rules of this game.

References

Powers, W. T., 1973, *Behavior: The Control of Perception*, Aldine Publishing Co., Chicago.

Powers, W. T., R. K. Clark, & R. L. McFarland, 1960, "A general feedback theory of human behavior: Part I," *Perceptual and Motor Skills 11*(1), August, 71-88.

[1989]

The Fallacy of Behaviorism

In the "experimental analysis of behavior" an attempt is made to adhere to a strictly empirical point of view. Some behaviorists even refer to this as the "scientific" point of view. No conjectures are made about the internal organization of a behaving system; all conclusions and all models are constructed only from observable phenomena. It is apparently thought that by working within such strict guidelines, one can achieve in the life sciences the same kind of rigor and objectivity that is found in the physical sciences.

That myth was invented by J. B. Watson, the founder of behaviorism (although he probably got its basis from biologists). While Watson is now considered old-fashioned even by behaviorists, the basic principles he expounded are still defended. Watson proposed that by varying the conditions surrounding any organism and observing subsequent changes in its behavior, he would be able to amass enough data to describe how a given organism will respond under any conditions. It was not necessary, he claimed, to invoke such mysterious concepts as consciousness, goals, motivations, emotions, or any other such thing that is known only subjectively. It was not even necessary to know how an organism works in order to observe the laws that govern its behavior.

I call that set of assertions a myth because in fact behaviorists (including, in his time, Watson) do make assumptions about how organisms work. If these assumptions had been made explicitly and competently and had been tested in the

Unpublished manuscript.

normal scientific way, they would soon have been found wanting; the underlying model would have had to be modified to fit the facts better. Behaviorism fails not because it tries to deal only with observations, but because it proposes an unworkable and untested model of organisms to explain those observations.

There are few behavioral scientists left who call themselves behaviorists, but many who believe in the same model that behaviorists have put forth. Scientific Method itself (as named and practiced by behavioral scientists) depends on this model. That is why I focus this discussion on behaviorism, even though the current fad is to pronounce it dead.

The reason that the failure of the model to work properly has not been discovered is that its correctness is assumed. When predictions based on the model fail, as they very often do, the failure is explained by saying not that the model is incorrect, but that organisms are inherently variable. The model is simply assumed to be correct, without demonstration. In this paper I will show what that model is and why it does not work.

Operant conditioning

B. F. Skinner is deservedly famous for creating a new kind of experiment, one in which the actions of an organism provide "reinforcers" rather than the reinforcers being administered arbitrarily by an experimenter. Skinner built an apparatus, since highly refined, that automatically converted actions such as pressing a bar or pecking on a key into delivery of reinforcements. This apparatus also converted the details of behavior into one common kind of "cumulative record," which produces a tracing of total responses against time. The slope of these traces represents the basic measure of behavior that Skinner and his adherents still use: the rate of responding. Ticks on the traces, representing deliveries of reinforcers, allow recording reinforcements in the same units: rate of reinforcement. We will look first at the properties of the apparatus, and then at the properties of the test-organism as described by behaviorists.

The apparatus

The automation of the administration of reinforcements meant that the experimenter could establish a precisely known manner of dependence of the reinforcement rate on the behavior rate. The simplest such "schedule" was the ratio schedule: for every n presses, one bit of reinforcement was delivered to the test subject. Knowing only the behavior rate, one can compute the exact reinforcement rate that goes with it.

The second primary kind of schedule was the interval schedule. The apparatus could be set so that during a time interval of length I (starting at the previous reinforcement event) no responses had any effect; the first response after this interval had elapsed produced a reinforcement. If the test subject responded at any rate slower than 1/I responses per time unit, just one reinforcer would be delivered per press, so the schedule would be equivalent to a ratio schedule with n set to 1. For rates of responding more rapid than 1/I responses per time interval, the effective ratio of reinforcements to responses decreased as the reciprocal of the rate of responding.

These are not schedules in the same sense that a train runs by a schedule; here the term means only the way in which reinforcement depends on behavior. Without knowing what behavior is occurring, there is no way to predict what reinforcement pattern will occur, so the term schedule does not mean a fixed pattern of administration of reinforcements.

Both the ratio and interval schedules can be represented as functions, piecewise continuous at least, that describe how rate of reinforcement depends on rate of responding. When random variations are added to the schedules, the average relationships can still be represented, or computer simulations can reproduce the relationships exactly. Thus we can easily incorporate these schedules of reinforcement into a computational model.

The test-organism

The effect of reinforcers on behavior is variously described by behaviorists as "increasing the probability of a response," or "maintaining behavior," or "increasing behavior." In short,

there is a claim that behavior depends on reinforcement in a generally positive way. This is where the strict principle of using only observational knowledge becomes an impediment.

Behaviorists have avoided proposing any specific model of the inner functioning of organisms. Yet they still claim that reinforcement has some effect on behavior, a generally positive effect. That, like it or not, is a model. It is a vague model, a scarcely-defined model, but it is a model: it is different, for example, from a model that claims that behavior decreases as reinforcement increases.

A behaviorist would object to calling the proposed relationship a "model of the organism." It is certainly not intended to be taken as a model: it is a generalization, supposedly, from observation. The observation is that behavior and reinforcement generally increase together, at least in operant-conditioning experiments done in the standard way.

A naive onlooker, however, might notice that we could predict the same relationship without knowing anything about the organism, and without even having a test-organism present. If any agency presses the bar or pushes down on the key at a certain rate, reinforcers will be delivered at a corresponding rate, totally predictable if we know how the apparatus is programmed. The rate of reinforcement is totally, physically, determined by the properties of the apparatus and the manner in which the bar or key is pressed. That is true whether or not an organism is receiving the reinforcements and emitting the responses. We could program a computer to press on the bar in a preprogrammed way, and there would be a corresponding totally predictable pattern of reinforcement. The reinforcement rate would generally increase with behavior rate—although a behaviorist might still insist on speaking as if the reinforcement were the independent variable and say that increasing the reinforcement rate produces a rise in the behavior rate.

We know that such a computer would be unaffected by the reinforcements. What, then, is the difference between the patterns of behavior when the organism and the computer are "responding"? It is significant that this test has never (to my knowledge) been done. As far as I know, nobody has ever compared the cumulative records obtained when (a) an organ-

ism is operating the bar or key, and (b) an artificial device programmed to behave arbitrarily is operating the bar or key.

I venture that many of the "findings" from operant-conditioning experiments would remain unchanged if the bar or key were being pressed arbitrarily, even at random. The "matching law," for example, would continue to be supported as before, and would tend to fail as before—or so I claim. Until a behaviorist troubles to do this experiment, we have no reason to believe that the results of operant-conditioning experiments reveal anything of importance beyond the properties of the apparatus.

That is not to say that organisms don't show characteristic behavior in the Skinner Box. They do. But they cannot be showing it for the reasons that behaviorists have proposed. There may be reason to believe that organisms in an operant-conditioning experiment are revealing facts of importance about themselves. Until a workable model of the organism is proposed and found adequate, those facts will remain undiscovered.

Simultaneous equations

As we already have one relationship between the behavior rate and the reinforcement rate (established by the apparatus), we really need a second relationship between these same two variables in order to be able to solve the two relationships as a simultaneous system and thus calculate the expected rates of responding and reinforcement. But doing that would amount to making a proposal about how the test-organism itself works —a proposal about the function inside the organism that makes behavior rate depend on reinforcement rate (and any other stimuli that may be involved).

Without an explicit proposal as to how the behavior rate depends on the reinforcement rate—that is, without a model of the organism as explicit as our model of the apparatus—the whole organism-apparatus system is underdetermined. There are two variables (reinforcement rate and behavior rate) and only one equation (the description of the schedule of reinforcements). Because the system is underdetermined, there is no way to predict the level either of reinforcement or of behavior.

All we can say is that whatever the pattern of behavior, the reinforcement will follow a corresponding pattern according to the properties of the apparatus.

I cannot believe that no behaviorist has ever faced this elementary problem of analysis. Every beginner in engineering or the physical sciences learns about it before learning anything else about analysis. One has to have one equation for each variable in a system. A single constraint doesn't allow one to say anything about the behavior of a system involving two variables. How is it, then, that behaviorists make such statements anyway? I think it is because of the model they tacitly use.

B. F. Skinner has said repeatedly that he "insists" that the environment is what causes behavior. Most behaviorists seem to interpret behavior as if this is necessarily the case. But if something in the environment causes behavior, then that something must be an independent variable.

If a covariance is observed between a measure of behavior and a measure of an environmental variable, the interpretation is always that the behavior depends on the environmental variable. While behaviorists (admitted or covert) say aloud that covariances do not prove causation, they still normally speak of "the effect of A on B," where A is never the behavior.

Often an independent variable is truly independent because it is totally under the experimenter's control. But in an operant-conditioning experiment, that is not true. The rate of reinforcement is a dependent variable in relation to behavior, and we can prove it by pressing on the bar and showing that we can exactly predict the appearance of reinforcers. But it's also believed that behavior depends on reinforcement; any reasonable person would suspect that, too. Which is the truth?

Neither is the truth. This is a system of two variables, related by two functions at the same time. One function describes the apparatus, the other the organism. If by environment one means reinforcers, then it's simply wrong to treat reinforcers as independent variables. They clearly depend on behavior.

Both the reinforcement rate and the behavior rate are *dependent* variables. What they depend on is not obvious. We have to guess. That is one purpose in constructing a model: to guess at the true independent variables that may be acting.

As I said, I cannot believe that no behaviorist has ever faced this problem, but I can make an educated guess as to what might have prevented publication of the outcome of facing it. Models are not like intervening variables—one is not totally free to make any assumption one pleases about the properties of the model. Every proposal for a model implies a behavioral result when that model is coupled to the environment. If the computed or simulated result doesn't fit observation, the model is wrong. What would be the result of assuming that behavior is a monotonic rising function of reinforcement?

If we say that behavior is some constant times reinforcement, which would fit all verbal descriptions of the effect of reinforcement on behavior, only one outcome is possible: both behavior and reinforcement rates are zero. If we include dynamics in this model, the solutions expand to include violent oscillations or unlimited growth toward infinity. There is an indeterminate case but the probability of observing it is zero. This simple model of the organism would probably be tried first by anyone looking for a workable model, but it would fail. So would any model that involves a positive linear or increasing nonlinear function; probabilistic or deterministic, static, sequential, or dynamic. The only way to fit the facts with a model in which behavior increases with reinforcement is to choose just the right nonlinearity, such that as reinforcement increases, behavior increases more and more slowly (for example, a logarithmic relationship). Then by carefully selecting the constants in the model one could fit it to a specific example of behavior. It would not, however, cover much of a range of behaviors, because the constants would have to be adjusted anew for every new circumstance—even for every new ratio or interval.

That, of course, would be unsatisfying and unconvincing; ad hoc models are of little use. If I had tried out such models as explanations of operant conditioning, I would not have seriously proposed them to anyone, or (to be honest) even mentioned them. It's too obvious why they can't work.

I have proposed a control-system model to explain operant conditioning (which then is no longer conditioning at all). It explains observations over a wide range of conditions, but not for all conditions—particularly those involving extremes of

deprivation, meaning the standard conditions for such experiments. My aim here, however, is not to sell control theory but to show why we *must* make models, whether they are good or bad ones. We can't observe how organisms work inside by watching them from outside. But we can make systematic guesses about what the inner organization might be, and systematically improve those guesses in the same way that a physicist systematically tests and improves models of inanimate nature.

And that is my final point, which I have made before and will no doubt try to make again. The successful sciences all depend on making models of natural processes that can't be directly observed. Watson quite obviously didn't know that when he said he was taking physics as his model for how to explore nature. Not only Watson but most life scientists who have little detailed knowledge of physics regularly assume that physicists must know what they are talking about, that they have the really objective facts of nature in hand. Scientific Method as practiced in the life sciences is supposedly an imitation of the way the physical sciences make discoveries—yet very few physical scientists make any use of that method. The whole fabric of physics and chemistry is built on models—models many times revised and refined, models tested against the most stringent demands on precision, but still models, figments of the human imagination. Behaviorism, by refusing to use that kind of imagination, has doomed itself to being forever limited to what the eye can see, and to juggling a single equation endlessly in the attempt to find answers that one can get only by treating two equations simultaneously—one of them a guess.

[1989]

Human Nature from the Standpoint of Control Theory

The most common conception of human nature is based on empiricism: watching how people behave, remembering and classifying everything, and sequencing the classes. To understand human nature then means being able to say "Under circumstances A, people in category B will react to event C by producing behavior D." This entails finding ways of determining

- that circumstances A do in fact exist.

- that a person is in fact in category B.

- that event C has in fact occurred.

- that behavior D does then in fact occur.

Each part of such a statement of human nature carries an uncertainty. Also, one or more of the parts may actually entail multiple uncertainties. For instance, people are characterized not just as belonging to one category, but as belonging to many at once: [female][heads of households][with no more than a 7th-grade education][with more than three children] [receiving ADC payments][between 1/1/85 and 12/31/88][in the state of New York]. Seven independent conditions must be established as existing in order to say that a person is in category B. If each condition has a probability of 0.97 of actually

Unpublished manuscript.

existing, there is then an 81 per cent probability that the person is in category B. If all four statements have that probability of truth, the entire statement has a 57 per cent chance of being wrong each time it is applied.

Statistical analysis of statements in the above form attempts to answer the question, "given the antecedents, did the consequent follow in a systematic way or only by chance?" I have never seen anyone ask the two obvious questions: "What was the probability that the antecedent actually was present?" and "What was the probability that the apparent behavior actually took place?" If the experimenter administered the instructions from a written sheet, if the subject filled in answers to all the questions, if the experimental manipulation was then executed, if the behavior was then recorded, it is taken for granted that there was no chance for error or misrepresentation in any of these steps. But not every person understands every word of the instructions the same way (or at all); not every subject tells the truth or remembers correctly when filling out a form; not every execution of a manipulation is identical to its predecessors; not every measure of behavior is recorded correctly. Of course if the implied variations are random then presumably the overall measure of variance is correct. But (a) the ultimate statement of fact assumes that all the variance is in the connection from antecedent to consequent rather than in determination of either the antecedent or the consequent, and (b) *systematic* variations in each element could not be detected. Experimenters may make systematic errors in reading the instructions (hesitating over "penis"); subjects may systematically lie about age, sexual preference, income, marital status, education, place and length of residence, political affiliation, number of abortions, debts, and so on. An experimenter recording data may consistently interchange "right" and "left," show hysteresis in categorizing behaviors, perceptually substitute stereotypes for individual behaviors that are really different. What is the probability of *systematic* error in recording both antecedents and consequents? Not zero.

With so many sources of uncertainty, one might expect that statements of human nature in the above form would prove to be wrong considerably more often than they are right. That is

correct: they do. The best you can hope for is that by insisting on the truth of such statements every time they look applicable, you will be correct in a predictable fraction of the cases. This means you will be wrong in a predictable fraction of the cases.

Back to the subject. The empirical method of characterizing human nature rests on predicting what people will do by observing what they have done before under certain circumstances. In the long run we are probably better off for being able to make such predictions. But we aren't much better off than if we just guessed.

The real problem comes when we try to use empirically-determined facts in combinations. If it's true that female heads of households (etc.) buy more bread than meat (probability 0.8), and if it's true that children raised on bready diets are sick more often than those who eat less bread and more meat (probability 0.8), then is it true that mothers in this class will have to contend more often with sick children? Yes (probability 0.64). And if that is true with that probability, and it is also true that sick children do better than non-sick ones at reading, having more time for it (probability 0.8), is it then true that the above mothers are likely to raise children who are sick but read well? Yes (probability 0.512). Here we have combined only three facts to make a prediction, and we would do just as well by flipping a coin. Add one more fact and we would do better by flipping a coin.

That is what is most wrong with the empirical approach to understanding human nature. It limits our understanding to isolated little groups of facts that can't be connected to form a Big Picture. The moment we try to deduce anything important about human behavior, we find that we get the wrong answer more often than the right one (although there is no law of nature saying that anyone has to admit this fact out loud). So there is no way to develop the empirical way of understanding human nature into a science.

That's the first thesis: that an understanding of human nature based on empiricism is inherently limited in scope and inherently unreliable. It can serve some purposes and obviously has sufficed to keep the human race alive for many millennia because it's better than nothing. But it doesn't work as well

as science works.

You'll notice that I am reserving "science" to mean a method more advanced than empiricism.

The next thesis is that science is based on constructing and refining models of reality, not on empiricism.

Physics is an experimental science, but it is not an empirical science in the same sense that we have been discussing. The point of doing experiments in physics is not to find out what causes lead to what effects (although such information is not thrown away) but to test the predictions of models.

The term "model" in this context does not mean a miniature of the real thing, but an imaginary picture of things and processes at a level of reality underlying the level that we can experience. The model-based approach does not take apparent cause-effect relationships at face value. It asks *why* such a cause-effect relationship is seen. Noticing a relationship is just a beginning.

The answer to this "why" is not a description of a motive or a purpose, not in physics, but a description of an underlying structure that *logically must* behave so as to create the observed relationship. The underlying structure consists of entities that we can't directly observe.

For example, we can see that as the sun rises, the landscape becomes bright and visible. Speaking empirically, we can say that the presence of the sun in the sky causes us to see the world around us if we are looking. But that empirical observation is explained in physics by saying that the sun emits *light*, which is partially absorbed and partially reflected by objects around us. The reflected light travels in all directions, some of it entering the lenses of our eyes. The rays are focused on our retinas, giving rise to neural impulses that constitute seeing.

The concept of *light* is part of a physical model of reality. We can't see light. Light travels through empty space from the sun to our back yards; we can see the back yards and the sun, but we can't see the light moving from the one to the other. We can't see it after it bounces and heads toward our eyes. We see the sun and the back yard and nothing else. The light is imaginary.

The model of light says that light carries visible energy and travels in straight lines at about 186,000 miles per second. If

that is true, then we would expect that holding something opaque between the sun and a patch of grass in the back yard would prevent direct rays of light from reaching the grass. We would predict that the region where light is blocked would look darker than the regions around it, because our eyes would receive less reflected light from it.

We would predict that the outline of the dark area would match the outline of the opaque object as seen from the direction of the sun. We would predict that if we moved the object sideways, the dark area would move in the same way, with no lag that we could discern.

These predictions are not based on the observation that most objects have cast shadows in the past (on clear days), even though that is true. They are based on the *properties of light* as we imagine them to be. We imagine that light moves in straight lines and that it can't pass through most solid objects. We imagine that we see because of light reflected from objects into our eyes. From those assumptions it follows that there must be shadows.

That deduction is completely independent of the actual existence of shadows. It's a *necessary consequence* of the assumed properties (and existence) of light. We can predict that there will be shadows even without knowing whether any actual shadows have ever been observed.

We can also predict the effects of new experiments. If we obtained a large mirror and masked it so it would reflect light from the sun onto the ground in a patch matching the shadow of the object, we could make the shadow disappear—and create a shadow of the mirror. We would predict that surrounding the shadow with a barrier that blocked light reflected from the surroundings would make the shadow darker. There is no end to the predictions we could make and there is no need for any predicted effect ever to have occurred before.

This model of light would seem to most people just to be a fact of nature. Everybody knows that light comes from the sun and reflects off objects. But everybody has not always known this. Before the 17th Century, a different model was used. Objects could in fact "cast shadows"—that expression is a vestige of the old model, as is "looking." The shadows of clouds at sunset, extending upward against a bright sky, were

called "anticrepuscular rays"—rays of darkness. The "facts of nature" were different then.

There is nothing new about making models. When Antony said (according to Shakespeare) that Caesar was not "ambitious," he was acknowledging a model of personality, in which an invisible thing called ambition could reside inside a person, accounting for certain kinds of behavior. When ancient Egyptians called the Sun by the name of Ra, and said that it traveled in a chariot across the sky, they were making a model of something they could not see to account for what they could see. It is not the making of models *per se* that makes physics into a science. It is the care with which they are made, and the ambition of making them perfect.

The point of creating a model in which light carries energy in straight lines is not to lead to the discovery that objects will cast shadows. We already knew that empirical fact. The point of making the model was to test the model, and the point of testing the model was to create an explanation of the phenomenon.

The model of light is accepted in physics not because it predicts some phenomena correctly under most conditions, but because it predicts all pertinent phenomena correctly under all conditions. No exceptions, no time-outs. This did not come about by luck, but through applying a method of testing and refining models.

The key to this method is to set high standards. It is not enough for a model to be plausible or "make sense." It must generate a prediction of behavior out of its own properties and that prediction must be quantitative as well as qualitative. The model must *behave,* in that it must generate a prediction of observable variables changing through time. Altering something that in principle must affect some part of the model will cause the model to change its action in some way, and that action will alter the world in some respect that we can sense or detect with instruments. And the consequence that the model predicts must match the consequence we observe *to the limits of our ability to measure.*

To repeat, this kind of result does not come about through lucky guessing. It comes about through making mistakes. No physical model ever worked properly the first time it was

proposed. Most beginning models predict one consequence while nature insists on producing a different one. But in that difference there is information telling the modeler how to change the model to move its prediction nearer to success. The difference between the phlogiston model of combustion and the oxygen model was subtle, for both predicted most phenomena correctly. Lavoisier finally realized that a subtle change in the model—the direction in which the critical effect flowed—would retain all the correct predictions and take care of the remaining wrong ones. So in the new model, oxygen combined with the burning substance instead of the substance losing phlogiston.

That's the kind of fiddling that eventually produces models that always work. There's no luck involved, except in coming up with each new possible change that will be, most likely, discarded. Always the modeler is asking what the effect would be if one or another aspect of the model were changed—how the model would then have to behave. Sometimes the necessary revision is obvious and is easily made. Other times, as in the case of phlogiston, we might have to wait 150 years for the simple and obvious revision to occur to someone.

Physics and other closely related sciences are successful because they use a rigorous form of the method of modeling, a form that demands perfection. This demand is what made progress after Galileo simply explode.

The idea of demanding perfection is an audacious one, as seen from one point of view. But from another it is simply a way of proposing that the natural world works as it does *because it has stable and repeatable properties*. A model that was made of entities that actually, although unobservably, exist, and that are related to each other precisely in the manner of natural laws that actually govern the world, would be a perfect model. Its calculated behavior would exactly match what we observe. Any failure of a model to predict exactly what occurs is therefore evidence that the model differs in some way from the perfect model—that is, from reality. It is at the same time a hint as to how the model must be changed to bring it closer to the right form, the *one* right form.

Whether there is only one correct model of the world is not very important. What is important is that we construct models

as if that were the case. Instead of excusing imperfect predictions, the modeler must welcome them as leads toward still another improvement in the model. If we find that a model predicts as perfectly as we know how to measure phenomena, we are moved to devise more accurate means of measurement so that we can detect smaller mistakes. When optical theorists had a model capable of predicting that there will be shadows, they then looked more closely at the edges of the shadows and discovered dark and light bands that the model didn't predict. That led to a modification of the model: light did not consist of rays, but of waves. Now the model could generate not only shadows, but interference fringes at the edges of shadows.

Aristotle's model of motion explained that a projectile was kept in flight by air rushing in to fill the vacuum behind it. When the initial impetus was exhausted, the projectile simply dropped vertically to the ground. Neither Aristotle nor any of the established authorities between his time and the dawn of modern science ever asked whether this model predicted correctly, or even whether it predicted at all. The only requirement placed on a scientific explanation was apparently that it seem plausible. This mode of explanation has been accepted since the start of history; in many fields of study that call themselves sciences it is still accepted. Once the explanation has been found that seems to fit the facts, it too becomes a fact. Freud invented the Id, Ego, and Superego, and built them into a model run by internal pressure of some sort. Given these terms and some plausible lines of reasoning involving them, psychoanalysis came into being.

But is there any such thing as an Id, an Ego, or a Superego? Does nature in fact "abhor" a vacuum? Is there anything corresponding to the idea of impetus? Questions like those, that would occur to any modern physicist, are missing from the prescientific way of understanding nature. The reason is that the idea of testing models is missing. No model can be tested if it is stated in terms from which no predictions can be made for comparison with observation.

And no predictions can be made from a model if it is specified so loosely that there is no way to deduce its behavior without knowing the real behavior it is supposed to match. Here we have an Id, an Ego, and a Superego. Without any

reference to the behavior of real people, could anyone deduce the kind of behavior these entities would exhibit under any conditions? That would be impossible: those ideas, like most others in the behavioral and social sciences, can be used only after the fact to offer a plausible description of what led to the observation. An ancient farmer could explain that the rain came because of weathermaking powers in his amulet. But if asked during a time of normal climate what the effect of the amulet is, he could only say, "Why, to keep things the way they are." The idea of testing this power of the amulet—of seeing if it would make even more rain than needed—wouldn't come up.

The way around this inability of models to predict is, of course, empiricism. If one bases knowledge of the natural world on a recording of past relationships and uses them to predict future ones, then of course nobody expects that predictions will work every time, or even most of the time. When that is the basic mode of research, the requirements on models are correspondingly relaxed. They don't *have* to work in every circumstance or all of the time. As a result, models are accepted even in the face of blatant counter-examples. Perhaps under some circumstances, an observer could see a projectile describing a smooth arc from start to end of its flight. But that single observation could not invalidate Aristotle's model in which projectiles drop straight down at the end of their flight.

Under the physical method of making models, the observation of a projectile following a smooth arc would, if accepted as competent, destroy Artistotle's model immediately. Freud's hydraulic model would never get off the drawing board because it can't, of its own properties, predict any behavior. The demands placed on physical models are such that the only possible outcome is a model that always works—or no model at all.

Control theory is based on the physical method of modeling.

When most people speak of human nature, they are thinking in terms of how people behave in relationship to the world around them. They may claim that people are gregarious, or selfish, or altruistic, or territorial, or aggressive, or freedom-loving, or susceptible to being led. Whatever characteristics are said to make up human nature, they are justified by citing

observations of people going about their varied businesses. Are people basically good or basically evil? Just think of Nazi concentration camps, or the Gulag Archipelago, or the suppression of student demands for democracy in China, or the torture of political prisoners in Central and South America (or practically anywhere else). Or, if you want the opposite conclusion, think of Amnesty International, or the students demonstrating for freedom, or the Resistance smuggling Jews out of Europe, or the Christian Martyrs, or the Marshall Plan.

The usual approach to understanding human nature seems to be to search for just those categories that will capture the essence of human beings, just the characteristics that will let us classify human beings once and for all. Both scientists and laymen seem to have the idea that somewhere, among all those hundreds of systems of classification that have been tried and remain to be tried, there must be one that will reveal universals of human behavior: characteristics, tendencies, or traits shared by all members of the human species. This hope is regularly frustrated, but the idea of finding universals of human nature is apparently less resistable than is the temptation to conclude that it isn't succeeding.

One obvious conclusion that might be drawn is that there is no such thing as human nature at the level of behavior. It is the nature of human behavior to be miscellaneous. One could even say that it is the nature of life to be miscellaneous. Some people seem to believe that the only level at which human nature can be defined is the biological level, the level at which we share most details of construction with many lower species. By reducing human nature to the level of biochemistry we can definitely find characteristics common to all, or very nearly all, human beings—certainly all normal ones. But to do this is to say that human behavior is unimportant, of interest only in a sort of subjective way. This does not satisfy many people.

A parable of the machine

Sometimes an analogy can make the cause of a difficulty easier to see. An appropriate modern analogy might be the behavior of computers. If we were to approach the question of

computer nature as many people approach the question of human nature, we would ask what computers tend to do when given certain inputs. To explore this question scientifically, we could think up a wide variety of inputs and outputs that computers can handle—lists of characters, tables of numbers, numbers representing the outputs of sensing devices, numbers that are converted into actions by printers, display screens, on-off switches, and so on. In the search for universals we would do controlled experiments: varying just one aspect of the inputs at a time, we would look for typical effects on output. We could subject single computers to a variety of inputs, or many computers to specific kinds of inputs. By doing statistical analyses we might find that there are tendencies for the computer to generate or avoid generating certain outputs when presented with, or forbidden, certain inputs. In short, we would manipulate the effects of the environment on computers and look for patterns in the responses.

Obviously this kind of investigation could go on indefinitely, as we seem to detect certain patterns and try to devise variations in the experiments to bring them out more clearly. We could propose endless hypotheses about effects of inputs, interactions among inputs, influences of surrounding circumstances on the effects and the interactions. As patterns were found and verified as being statistically significant, we could easily get the feeling that we were getting somewhere, that we were building up a body of basic scientific knowledge about computers. If this body of knowledge seemed a bit disorganized, if we found it difficult to understand what it all means, we could at least be satisfied in knowing that we were providing the raw material for future scientists to use in their progress toward understanding. Perhaps it is somewhat puzzling to know that computers presented with both even numbers and odd numbers tended to type out more "A"s than computers presented with odd numbers only. But some day future generations will be able to explain that effect, thanks to the fact that we noticed it.

Now why is it that this way of understanding computer nature seems somewhat ludicrous? The more you know about computers the more ludicrous it becomes, but most people know enough to understand what is the matter here. Anyone

who runs games or balances books or does word-processing or creates graphic pictures or sends and receives eletronic mail knows that what a computer *does* is determined by what floppy disk you stick into it. A computer is not an adding-machine, although it can behave like one. It isn't a word-processor, although it can do word-processing functions. It isn't any of the things it is used for. The behavior of the computer depends not on the nature of the computer but on the nature of the program that is running in it.

To put that idea in terms a little closer to the problem of understanding human nature: what makes a computer useful for so many different applications is that its organization is determined by an abstraction called a program, and programs can have any organization whatsoever. When we study the behavior of a computer as it responds to various inputs, we are really seeing only the program that happens to be running in it. When you press the space bar on the keyboard, you may be used to seeing a space appear on the screen, but there's nothing inherent in computer nature to say that the result couldn't be the letter Q, or the string of characters "This is not a space," or even clearing of the screen and termination of the program. The effect of any input on the computer's actions is determined completely by the program, and not at all by the nature of the computer or of the input.

This conclusion may seem to put us at the mercy of the biologists. Programs are simply strings of ones and zeros, which in turn are really just voltages or lacks of voltage (or magnetic fields aligned one way or another way), so these simple physical variables are really what cause the physical computer to act as it does. To see why this is not the correct way to understand computer nature, either, we can carry the analogy a little farther.

Suppose we give the computer and its floppy disks to a team of physicists and chemists charged with telling us why this machine behaves the way it does.

One team looks at the floppy disks, where something called a "program" supposedly resides. They remove the disk of plastic from its paper folder and begin to analyze the patterns of magnetization on its surface. Scanning the surface they find that the magnetic patterns are confined to separate circular

tracks. By scanning these tracks they discover some basic regularities in the patterns of magnetization, which they then analyze for the distribution of magnetic energy. They find that there is a power spectrum easily discernible, with many changes in magnetization occuring at intervals of a few thousandths of an inch, and that various parts of the disk show definite differences in these power spectra. It is even possible to find areas scattered over the disk that can be classified as belonging to the same species on the basis of similarities in spectra. The next step, once this exploration is done, is to scrape off the magnetic material and bring the big guns to bear on it: the cryogenic magnetic detector, the nuclear magnetic resonance analyzer, the mass spectrograph, and the chemical analysis using radioactive tracers.

While that was going on, the other team was taking the computer apart. First all the connections were traced and the courses of current flow were recorded. Then the integrated circuits were removed and opened. The faintly-patterned chips of essentially pure silicon inside were then given the same treatment as the floppy disks, finally being ground up into powder and subjected to the most sophisticated possible analysis.

The report of these teams would not, of course, say anything about word-processing or running games. It would not mention that the computer could carry out logical operations or add, subtract, multiply, and divide numbers. This team would come across neither logical nor arithmetic variables; it would never find the letter "A"; it would never see anything of the actual functions of this machine. Yet it might well announce that it now understood essentially all the basic properties of the computer.

So we now have two versions of computer nature, one reported by those who investigated the behavior of the computer, and the other reported by those who investigated its physical organization and composition. Both, of course, claim that they have been exploring the true nature of the computer. In fact neither group, as any computer expert would be happy to explain, has even come close, even though neither group has made any mistakes.

The computer expert will point out that we can run the same

program on many different machines, even machines that accomplish the basic operations with different circuitry, and that accept inputs in different forms—magnetic tapes or even punched holes in paper tapes, instead of floppy disks. Clearly, the physical form of a program makes no difference. But beyond that, even the *program* makes no essential difference in the operation of the computer. All that counts in the final analysis is that certain basic operations be executable, somehow, by the computer, and that the computer be able to operate using stored information. That is computer nature. It is the nature of a computer to be programmable.

Metatheories of human nature

From this analogy we can't conclude that human nature is defined as programmability (or that brains are computers). To do that would be to miss the point of the story. The real point is that there is a level of analysis that is neither behavioristic nor biological at which we can, in fact, hope to find definitions of human nature that are truly universal.

At this level we do not ask either what the system does, nor how it physically accomplishes what it does. We ask how the system is organized. This is not the same as asking how its *behavior* is organized, or how its *components* are organized. It is to ask what functions the system must be able to perform in order to do *all of what we see it doing*. In this kind of analysis we do not care how, physically, any function is carried out, nor do we care whether one example of it or another is used.

At this level of inquiry, we are not interested in statistics or in physical analysis. We are asking how this system has to be organized in order to behave at all, in order to show any example of behavior whatsoever that we happen to notice. It follows that whatever we propose, there can be no exceptions to it. Even a single counterexample would show that the proposition is wrong, at least in some respect. It may seem that by putting such severe restrictions on accepting proposals, we rule out success from the start. But that isn't true. In fact most sciences, including sciences of behavior, accept certain statements about human nature that have, or at least are believed to have, just this degree of universality.

These statements have been called *metatheories*. They are theories of a science that are assumed to be correct, as the basis for justifying all more detailed theories. Most scientists do not treat them as theories at all—they call them fundamental principles of science. In the behavioral sciences, the metatheory, the fundamental principle, that underlies all more detailed theories about human nature, could be called simply "causality."

This theory shows up in many disguises, disguises that can conceal the basic concept so thoroughly that the underlying agreements are ignored even while rival theories that depend on the same principle are debated. The basic idea is so simple and straightforward, so well supported by observation, that questioning it just wouldn't occur to most scientists. The idea is that the behavior of organisms, including that of human beings, is governed by what happens to them.

This metatheory can be justified as follows. Organisms are made of "irritable tissue," which means that small influences on them can have large effects on them. The prime example is the sensory system of a higher organism. Minute amounts of energy impinging on a sensory nerve make it generate trains of neural impulses that pass inward into the nervous system and brain. At many stages in the nervous system, the activity of one nerve triggers off activities in many more of them, so that the initial small effect spreads and becomes larger. Ultimately these consequences of stimulation reach the muscles, more irritable tissue, which respond to small electrochemical discharges by converting stored energy into large physical forces. These physical forces produce the movements we observe as behavior.

Few behavioral scientists are actually interested in behavioral phenomena at this level of detail. Many of them are far more interested in what happens to those trains of impulses as they pass through the networks of the brain. Others still are interested only in the overall effect, the final observable relationship between external events that affect the organism and the consequences of muscle activity. But when we look at the way any scientific results are obtained and reported, we can see that without exception all workers at all levels of abstraction accept the metatheory that behavior is caused by what

happens to an organism.

The clue lies in what is called "scientific method," the experimental method with which all hypotheses must be tested before they can be published. Whatever the phenomenon, it must be investigated by varying some condition and looking for a correlated change in behavior. Most scientists have learned to be cautious about asserting causality, but they assert it just the same. No matter what the circumlocution, the implication of a report is always that if it were not for the manipulation that the experimenter applied, the behavioral effect would not have occurred.

This is why control experiments are done. The experiment is done both with and without the manipulation, and the behavioral changes are recorded. The reason for the control experiment is precisely to show that it was the manipulation, not chance, that was responsible for the behavioral effect. Whether the manipulation is baldly said to have caused the effect, or is delicately said only to show a correlation with the effect, we are still talking causality. If it were not for A, we would not have observed B under these conditions—at least not as often.

So there is an assertion about human nature implicit in the very methodology of the behavioral sciences. It is the nature of a human being to transform incoming effects into outgoing behavior. This, until very recently, was the fundamental article of faith of all the sciences of life, not just those concerned with human beings. To "be scientific" meant to accept this basic fact of nature as a fact, no matter what level of abstraction one preferred.

To understand what control theory has to say about human nature (the subject of this article in case you had forgotten), one must see that it presents an alternative not just to existing theories of behavior, but to an almost-universally accepted metatheory of behavior. It is not possible for control theory to be right and for the older metatheory also to be right. They are mutually exclusive.

Human nature according to control theory

The control-theory version of human nature—or the nature of organisms, for that matter—can be put succinctly. Organ-

isms control. Whatever we see them doing, at whatever level of analysis we prefer, we see them controlling, not reacting.

The old metatheory says that there is a one-way path through the organism, from cause to effect. The final effect is behavior. The new one says that there is a closed loop of action that has neither a beginning nor an end. The old concept says that behavior can be expressed as a function of independent variables in the environment. The new concept says that behavior is varied by the organism in order to control its own inputs. The old concept says that environments shape organisms. The new one says that organisms shape both themselves and their environments.

The details of the control-theoretic argument can be found elsewhere—they are founded on good physics, biology, and mathematics and cannot be dismissed as trivial, or by trivial arguments. The point here is not to defend control theory, but to make plain its implications. Control theory offers a new conception of human nature and of the nature of organic life itself. As a metatheory, it applies to all examples of behavior but does not depend on any specific examples of it. Its applicability is independent of the material properties of organisms; it applies at all levels of analysis including the biochemical. It is an organizational principle, not a behavioral law.

According to control theory, it is the nature of human beings to control what happens to themselves as individuals and as a species. They are organized specifically as required to work in this way. That is a new conception of human nature.

[1990]

Standing at the Crossroads

I'd like to try today to give you the sense that psychology is standing at a crossroads—and not only psychology, but all the sciences of life. We are about to experience the advent of something for which many people have searched, an organizing scheme that pulls together all the disparate schools of thought, specializations, movements, and evanescent fads that make up various fragmented branches of the life sciences.

The organizing scheme is called "control theory." This theory explains a phenomenon, as theories are supposed to do. The phenomenon in question is called control. Everyone has heard this word, and most people have occasion to use it from time to time, but in science it has become part of the meta-language rather than designating a subject of study. A scientist does a control experiment, or demonstrates how manipulation of stimuli and rewards can control an animal's movements, or advocates a proper diet to control cholesterol level, or competes for control of a department. This word is used as part of a background of ordinary language, but it has not been part of the technical language of the life sciences.

The reason is quite simple: nobody in or out of science understood the process of control until about the beginning of World War 2. By understanding the process, I mean being able to define it, characterize it, measure its parameters, predict how it will proceed, and recognize it in a real system. This doesn't mean that control was impossible to accomplish before World War 2: after all, most people accomplish digestion with-

Distributed at the Control Systems Group Meeting, August 15-19, 1990, at Indiana, Pennsylvania.

out understanding any biochemistry. But control is as natural a process as digestion, and like digestion can be understood in a scientific way only by studying it and learning how it works.

World War 2 started only about 50 years ago. Perhaps you can see why this fact implies some problems with studying control as a natural process. If control is a natural process, it was occuring in 1840, 1740, 1640, and so on back to the primordial ooze. In 1940, the sciences of life were already something like 300 years old (and their prehistory was far older than that). If nobody understood control until 1940, it's clear that these sciences went through a major part of their development without taking it into account. The next question is obvious: how *did* they explain the phenomena that arise from processes of control?

Many of the puzzles and controversies that occupied early researchers could have been resolved if scientists had realized that they were dealing with control processes. Purpose could have been studied scientifically instead of merely theologically. We can see now that all these early researchers, not recognizing a control process when they saw one, were drastically misled by some side-effects of control. The principal side-effect that deceived them resulted from the way control systems act in the presence of disturbances of the variables they control. When a disturbance occurs, a control system acts automatically to oppose the incipient change in the controlled variable. But if this opposition is not recognized (it's not always obvious), the observer will inevitably be led to see the cause of the disturbance as a stimulus and the action opposing its effects as a response to the stimulus. Furthermore, this opposition results in stabilizing some aspect of the environment or organism-environment relationship. That stabilization conceals the role of the stabilized variable in behavior; the better the control, the lower will be the correlation between the controlled variable and the actions that stabilize it. The variable under control is the one that is actually being sensed, but the logic of control makes it seem that the disturbance is the sensory stimulus.

Donald T. Campbell has proposed a "fish-scale" metaphor of scientific progress. Each worker constructs just one small scale that overlaps those already laid down by others. Even-

tually the whole fish will be covered completely. But what if the fish is a red herring? Then all these patient workers will devote their lives to covering the wrong fish. The converse of the fish-scale metaphor is that a person who is concentrating on fitting one little scale to others already laid down is bound to have a very localized view of the problem. Seeking to extend the accomplishments of others, a single worker can make what seems to be progress—but it is unlikely that a single worker will discover that something is wrong with the whole design. The result can easily be the diligent application of fish-scales to a giraffe.

I submit that something like this has happened in the life sciences. A fundamental misconception of the nature of behavior, natural but nevertheless horrendous, has pointed the life sciences down the wrong trail. Nearly all life scientists, particularly those who try to achieve objectivity and uniform methodology, have interpreted behavior as if it were caused by events outside an organism acting on a passive mechanism. This hypothesis has become so ingrained that it is considered to be a basic philosophical principle of science. To explain behavior, one varies independent variables and records the ensuing actions; to analyze the data, one assumes a causal link from independent to dependent variable and calculates a correlation or computes a transfer function. This leads in turn to models of behaving systems in which inputs are transformed by hypothetical processes into motor outputs; those models lead to explorations of inner processes (as in neurology and biochemistry) predicated on the assumption that one is looking for links in an input-output chain. One assumption leads to the next until a whole structure has been built up, one that governs our thinking at every level of analysis from the genetic to the cognitive.

Control theory, by showing us an alternative way of understanding this entire structure, therefore threatens the integrity of practically every bit of knowledge about behavior that has ever been set down on paper.

This is, of course, a message of the type that leads to a high mortality among messengers. That is why you are listening to a person with no reputation to lose and no fame to protect, instead of a Nobel Prize winner. In an utterly predictable way,

scientists have for the last 50 years gone to great lengths to avoid learning control theory or else to assimilate it into the existing picture of behavior. Failing that, they have simply declared it irrelevant to their own fields, with the result that the authoritative literature of behavioral control theory is almost completely insulated from the mainstream. It appears in publications like proceedings of the Institute of Electrical Engineers division on Man, Machines, and Cybernetics, or in human factors and manual control publications, or in Xeroxed papers passed from hand to hand. There is a scattered literature on control theory in the life sciences, but nothing on this subject gets past the referees into a standard journal without first having its teeth pulled.

Despite all the defenses, the concepts of control theory are spreading. When our descendants look back on the latter half of the 20th Century, they will probably be amazed at the speed with which control theory became accepted: 50 years in the course of a science is nothing. We control theorists have nothing to complain about. Our greatest successes have come not through pounding at locked doors, but through continuing to explore the meaning of this new approach and learning how to apply it in many different disciplines. If we do our job correctly, acceptance will take care of itself. That job is not something one can toss off overnight, nor can it be done by just a handful of people. We are coming to a time of rigorous re-evaluation of all that is known or presumed to be known about the nature of organisms. The more people that are involved in this enormous project, the sooner it will be accomplished. That is why we are all so glad to welcome our guests at this session: after the party, you will be invited to help do the dishes.

There has been progress in understanding how organisms work, the wrong model notwithstanding. Biochemical reactions are not going to change because of control theory. Muscles and nerves will continue to operate as they are known to operate. Even at more abstract levels of analysis, many phenomena will continue to be accepted as valid observations; for example, phenomena of perception, of memory, of cognition. If competently observed, these phenomena will still be part of the legacy of earlier workers. When we pull the stopper on the

old theories, we must keep a strainer over the drain and let only the bathwater out.

Part of the task of reconstructing the sciences of life consists of separating valid observations of components from invalid conjectures about how they work together. Consider biochemistry as an example. Biochemistry is an odd mixture of solid research and wild leaps of undisciplined imagination. The research reveals chemical processes taking place in the microstructure of the body. The wild leaps propose that the chemical reactions somehow directly produce the behavioral effects with which they are associated. It's as though a specialist in solid-state physics were to propose that electrons flowing through wires and transistors are responsible for the music that comes out of a radio. While it's true that a shortage of electrons will make the music faint, and that without the electrons you wouldn't get any music, the physicist would be laughed out of town for suggesting that electrons *cause* music, or that you could fix a weak radio just by putting some more electrons into it. You can't understand the role of the electrons without grasping the principles of organization that make the radio different from a radio kit.

In the same way, if shortages or excesses of chemicals like enzymes and neurotransmitters are found to be associated with functional and behavioral disorders, all we then know is that these substances play some role in the operation of the whole system that creates organized behavior. If there's a shortage of some chemical substance, then some other system has reduced its production of that substance, and some other system still has decreased its effect on the driving system, and so on in chains and causal loops. Nothing in a system as complex as the human body happens in isolation. If biochemistry is to have anything to say about the organism at any higher level, biochemists are going to have to study whole systems, not isolated reactions. We need a functional theory to supplement the microscopic laws of chemistry.

There are workers in biochemistry who are investigating feedback control processes. One significant process involves an allosteric enzyme that is converted into an active form by the effect of one substance, and into an inactive form by the effect of another. When these two substances have the same concen-

tration, the transition from active to inactive is balanced; the slightest imbalance of the substances causes a highly amplified offset toward the active or the inactive form. In one example, the active form catalyzes a main reaction, and the product of that reaction in turn enhances the substance that converts the enzyme to the inactive form—a closed-loop relationship. The feedback is negative, because the active form of enzyme promotes effects that lead to a strong shift toward the inactive form. This little system very actively and accurately forces the concentration of the product of the main reaction to match the concentration of another substance, the one that biases the enzyme toward the active form. This allows one chemical system to control the effects that another one is having on the chemical environment.

A person without some training in recognizing control processes might easily miss the fact that one chemical concentration is accurately controlling the product of a different reaction not directly related to the controlling substance. The effect of this control system is to create a relationship among concentrations that is imposed by organization, not simply by chemical laws. This is the kind of observation that a reductionist is likely to overlook; reductionism generally means failing to see the forest for the trees. Even the workers who described this control system mislabeled what it is doing—they concluded that this system controls the outflow of the product, when in fact it controls the concentration and makes it dependent on a different and chemically-unrelated substance.

To shift through several gears, consider the lines of research that began with Rosenblatt's perceptron. This device was conceived as a behavioral system that could be trained to react to patterns contained in its input information. First this idea was shown, by something of a hatchet job, to be impractical, and then it was shown to be practical again if several levels of training could occur within it (I haven't seen any apologies to Frank Rosenblatt, who died without vindication). In all its incarnations, however, the perceptron has been thought of as a system that learns to "respond correctly" to a stimulus pattern.

From the standpoint of control theory, however, organisms do not respond to stimuli but control input variables. So does

that invalidate all that has been learned about perceptrons? Not at all. Control-theoretic models desperately need something like a perceptron to explain how abstract variables can be perceived. In a control model, however, the perceptron is only one component: it provides a signal that represents an aspect of some external state of affairs. It's easy to show that behavior can't be explained simply by converting such a signal into an output action. But behavior can be based on the difference between the perceptron's output signal and a reference signal that specifies the state of the perception that is to be brought about. The control-system model shows where the functions that are modeled as perceptrons belong in a model of the whole system.

Shifting gears again: some theorists are trying to model motor behavior in terms of "motor programs" and "coordinative structures." In these models, command signals are presumed to be computed such that when applied to elastic muscles they produce the movements of a real limb. These models contain some impressive mathematics, taking into account the linkages of the limb and the dynamics of movement of the limb masses. But control theory says that behavior is not produced by computing output; it is produced by comparing inputs with desired inputs, and using the difference to drive output. No complicated "motor program" computer is needed. Does this mean that the mathematical analysis by the motor program people is spurious and ought to be discarded?

Again, not at all. At some point in elaborating the control model, we must show how the driving signals actuate muscles to cause the movements we actually see. This entails solving all the physical equations for muscle and limb dynamics, just as the motor programmers have done. If they did their arithmetic right, it will still be right when we substitute the control-system model for the central-computer model. Both models have to produce the same driving signals. The only thing that will change is that control theory will show how the required driving signals arise naturally from perception and comparison against reference signals, instead of being computed blindly from scratch.

Finally, shifting to overdrive, what do we do about Artificial Intelligence? We take advantage of whatever it really has to

offer, modifying it only where we know it fails to explain enough. One place where it fails to explain enough is in the way it deals with action. Basically, it doesn't deal with action. It starts its analysis with perception of abstract variables in the form of symbols, constructs models that imitate human symbol-handling processes as well as possible, and finishes by generating more strings of symbols that describe actions to be taken. It says nothing useful about how a description of an action, in symbols, gets turned into just those muscle tensions that will in fact produce an action that fits the description. When devices are built that are run by symbol-processing computers, the critical transformations that make action out of symbols are simply put into the device by its builders. Many of those critical parts turn out to be servomechanisms—control systems.

The assimilation of control theory into the life sciences will require a lot of this kind of reanalysis. Some old ideas will have to go, some will stay. This job is best done by people who are already competent in existing fields. Of course these also have to be people who can see that there is room for improvement along lines other than the standard ones.

In the current membership of the Control Systems Group we have representatives of at least a dozen disciplines of the life sciences, and a few persons representing some unlikely occupations such as piano teaching and law. When these people meet, there is little difficulty in communicating because all of them have a basic understanding of control theory. But communication isn't the only factor that makes these meetings valuable. The most important lesson comes from seeing how control theory applies in someone else's field.

The biggest problem with introducing control theory to scientists in conventional disciplines is that each scientist tends to think only of the scientific problems that are defined in that one field. The problem in question may involve behavior, but behavior is generally taken on faith to work the way some other specialist says it works. In fact most scientists tend to dismiss details involving other fields, assuming (often quite wrongly) that somebody else understands them well enough. We therefore find some very detailed biochemistry or neurology or personality-testing, all done competently, being used to

explain behavioral phenomena that are very poorly analyzed and in many cases don't actually occur. The sociobiologist concludes that behavior patterns are inherited, not knowing that only the consequences of motor outputs, not the outputs themselves, repeat. What does a geneticist really know about the actions through which a bird catches a bug? You can inherit the control systems that are capable of catching bugs, but you can't inherit acts that happen to take you where a particular bug is going next. The combination of narrow expertise in one field and naive conceptions in every other field leads to facile explanations that are right only at one point.

Specialists must see the need for a model of behavior that applies in *all* disciplines, even those in which the specialist is not competent. Once the Artificial Intelligence researcher understands exactly why organized behavior cannot be produced by computing outputs, he or she will modify the AI model so it will work correctly with more detailed systems actually capable of organized behavior. Important effects of learning how control theory applies in other fields will occur at the boundaries between disciplines—exactly where we need to work if we are ever to have a unified science of life. At Control Systems Group meetings, specialists from many fields hear other specialists talking about the way control theory has made them rethink the problems in a different field. Because of the common understanding, this inevitably reveals one's own hasty assumptions, and encourages still more rethinking.

One last remark about the CSG. The CSG does not represent any one scientific discipline. It has no agenda of its own beyond encouraging the application of control theory within existing disciplines—no agenda, that is, except perhaps lowering the barriers between disciplines. The psychologists in the group are still psychologists, the sociologists are still sociologists, the therapists are still therapists, the engineers still engineers. This is not a political movement nor an alternative to established science. It is simply a vehicle for promoting interaction among people interested in using or learning more about control theory in any specialty whatsoever. When all the branches of the life sciences have assimilated and begun using control theory, the CSG, its work accomplished, will have no

further reason to exist.

In this presentation I have talked around control theory, alluding to some of its conclusions without attempting to justify or explain them. Learning control theory can't be done by listening to a half-hour's talk. I hope that some of you will find the promise of a unifying principle for the life sciences appealing enough to go further into this subject.

[1990]

The X-Phenomenon

Many people in cybernetics and other fields are talking about control and control theory. In most cases, the principle apparently used to explain control is that of compensation, following Ashby and other notable figures. Others seem to think that control is a matter of prediction, analytic or stochastic. Still others appear to believe that control can be achieved by a system that analyzes its environment and computes the output that will have the desired effect. Some have believed that control can be achieved by (somehow) matching a measure of information in a system's output to a measure of information in disturbances. A few have assumed that control is a process that takes place entirely inside an organism, modifying what is otherwise an input-output conversion process.

In the face of all these explanations of control, I propose to abdicate temporarily as an authority on whatever the phenomenon is to which they refer and to apply for a position as the discoverer of a new phenomenon. I will call it the X-phenomenon; we can defer proposals about a better term to substitute for X.

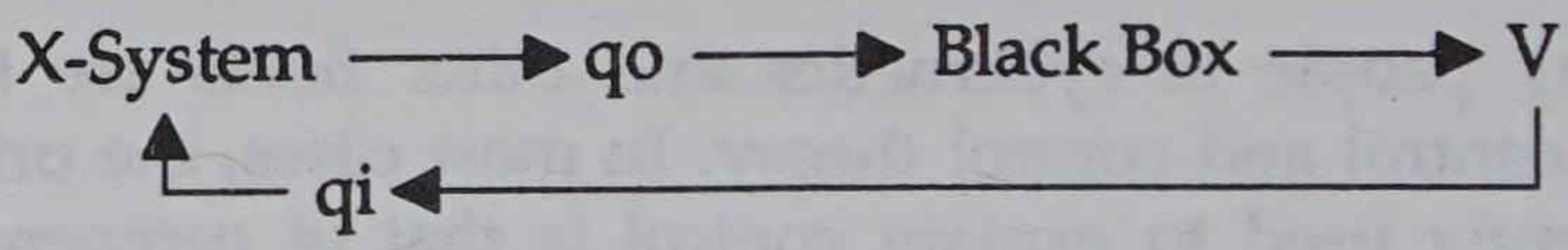
The X-phenomenon is created by the interaction between an environment and a type of system I will call an X-system. The nature of an X-system is recursively defined by the X-phenomenon, which can be observed both in the laboratory and in natural systems. Only living organisms naturally exhibit the X-phenomenon. I will describe a canonical experiment that can distinguish between an X-system and other organizations.

Distributed at the Control Systems Group Meeting, August 15-19, 1990, at Indiana, Pennsylvania.

Setting the conditions

An X-system has a sensory input and a motor output.

The motor output generates qo, the output quantity. In this experiment, qo affects a black box in the environment. Out of the black box comes an effect that determines the state of some remote environmental variable, which we will label V. The input of the X-system, called the input quantity qi, represents the state of V, the remote effect of action. The sensory input qi is simply proportional to V. The arrangement is:



The black box conceals coefficient M and variable A that transform qo into V according to:

$$V = Mqo + A.$$

Hidden inside the black box, A and M vary smoothly but unpredictably over limited ranges. M is greater than 0; A varies between positive and negative limits. To assure that A and M vary unpredictably, we could derive them from successive pairs of digits in the decimal representations of Pi and e. We could also use a random-number computer algorithm, because all that counts is that the X-system be unable to predict variations in A or M. A smoothing and scaling algorithm limits the speed and amplitude of variations to values representative of natural variations in macroscopic physical quantities and relationships.

To repeat: processes inside the black box are known to the X-system only through their effects in transforming qo into V.

The X-phenomenon

The X-phenomenon is this: if the system is an X-system, it will produce just those output variations that will keep V at some predetermined constant or varying value r(t), generated inside the X-system, within small error limits. The waveform r(t) determines the behavior of V, but not the behavior of the

system's output quantity qo.

Also without any design change, the X-system will vary its output so that a smoothly-varying extraneous disturbance acting directly on V will be canceled by changes in the output of the black box, again within small error limits, while V continues to be determined by r or r(t).

The error limits can be made very small in comparison with the range of practical values of r or r(t): one per cent or less. These limits are not statistical means, but absolute limits on the amount of error that is seen at any time.

A workable model of an X-system

There are many ways to model a successful X-system operating through various black boxes. In the present case, the simplest way is to treat r(t) as a waveform generated inside the X-system, and make the output quantity, qo, proportional to the time-integral of the difference r(t) - qi:

$$qo = k\ int[r(t) - qi].$$

The larger k becomes, the smaller will be the error limits within which V is maintained. This design will work with the black box as it is defined here. It will also work with many other black box definitions.

The behavior of an X-system

An X-system stabilizes the state of a remote environmental variable or causes that variable to change in an arbitrary but repeatable way. These effects can be terminated either by preventing the output of the X-system from affecting the black box or by preventing the variable V from affecting the input quantity of the X-system. We know, therefore, that the X-system is the agent responsible for the regular behavior of V and that it makes use of sensory information about the state of V.

We also know, however, that the regular effects can't be created by corresponding regularities of action by the X-system: the black box sees to that. Somehow the X-system is

producing just the variations in its output quantity that will nullify the changes in offset and scale factor that continually occur inside the black box, while maintaining the specified net effect on the remote environmental variable.

About 100 years ago, William James said that the hallmark of a living system is its ability to create constant ends by variable means. In a variable environment, the production of constant ends can be accomplished *only* by varying the means, the system's actions. Repeating the same action would generally create a different result. The phenomenon on which James remarked was the X-phenomenon.

An example of the X-phenomenon

A driver who keeps a car on a straight road for 20 miles is exerting forces on the steering wheel (qo). These forces act through the front wheels, which generate sideward forces on the car according to a variable coefficient of friction (M). Other forces caused by wind, tilts, and bumps (A) independently produce sideward accelerations. The net sideward acceleration is such that the position of the car on the road (V) remains the same, mile after mile. The driver senses the position of the car visually (qi). If the steering linkage fails or if the driver is blindfolded, successful steering ceases. The driver can specify the path of the car, but can't predict the steering efforts.

Unworkable explanations

The conditions of the canonical experiment rule out many classes of explanation that might appear to work under other circumstances.

1. *Centrally-computed commands* can't explain the behavior of V; there is no command that can specify the effector output that will produce a specified state of V.

2. *Sensory guidance* can't direct effector output so as to stabilize V near any preselected value: for any sensed state of V, the effector action required to hold it there or to change it to a given other state might be of any amount and in either direction. See 1.

3. *Statistical or quantitative predictions* of the effect of action

on V can't result in actions that will maintain V close within small limits to a given state. The correlation of the output quantity with V will be very low. Short-term predictions cannot produce long-term stability; long-term predictions cannot prevent large short-term variations in V.

4. *Guidance by the goal-state of V* is impossible, because the system senses only the actual state. The goal-state is set by r(t), inside the system and not in its environment.

5. *Generation of compensating actions based on sensing disturbances* will not work because no information about the interior of the black box (A and M) is available to the system. Only the combined effect of disturbances and actions is visible in the state of V.

6. *Matching the "variety" of action to the "variety" of disturbances* will not work for two reasons. First, many different patterns of action and disturbance would yield the same measure of "variety," whereas only one pattern of action, qo = [r(t) - A]/M, will stabilize V. Second, the variety of the disturbances is not measurable by the system because the disturbances A and M are not sensed (see 5).

Conclusions

With the advent of cybernetics, many workers thought that control theory could explain the fundamental phenomenon of behavior described here. The servomechanisms that first interested Wiener and his colleagues were in fact X-systems. Early cybemeticians accepted the proposition that control theory explained this closed-loop kind of behavior. Most of them have chosen, however, to interpret control theory in ways that rely on one of the principles ruled out above.

Control theory as developed by engineers (and especially as applied in the field of analogue computing) correctly predicts the behavior of X-systems. Control theory as described by many cybemeticians, psychologists, and others does not.

From this fact we can draw either of two conclusions: (1) The X-phenomenon is new and requires an explanation other than control theory, or (2) those who have tried to explain closed-loop behavior according to the principles just listed have not understood how control theory explains it.

If the second conclusion is accepted, we can simply substitute the word "control" for "X." In that case I will gladly relinquish my claim to discovery of a new phenomenon, and return the credit to those who deserve it: the engineers of the 1930s who first understood how the process of control actually works.

Published Works by William T. Powers on Living Control Systems, 1989-1991

1989a "Volition: A semi-scientific essay," in Wayne A. Hershberger, ed., *Volitional Action: Conation and Control*, North-Holland, Amsterdam, et al., 21-37.

1989b "Quantitative measurement of volition: A pilot study," in Wayne A. Hershberger, ed., *Volitional Action: Conation and Control*, North-Holland, Amsterdam, et al., 315-332.

1989c (with Richard S. Marken) "Levels of intention in behavior," in Wayne A. Hershberger, ed., *Volitional Action: Conation and Control*, North-Holland, Amsterdam, et al., 409-430.

1989d "Problems with the Smiths," *Continuing the Conversation: A Newsletter of Ideas in Cybernetics* (16), Spring, 3-4.

1989e "By any other name," *Continuing the Conversation: A Newsletter of Ideas in Cybernetics* (16), Spring, 11-12.

1989f "Things I'd like to say if they wouldn't think I'm a nut: Or, overgeneralizations that aren't *that* far over," *The Control Systems Group, Inc. Newsletter*, May, 5.

1989g "Foreword," in Edward E. Ford, *Freedom from Stress*, Brandt Publishing, Scottsdale, Arizona, ix-xi.

Prepared by Gregory Williams.

1989h (editor, with Richard J. Robertson) *Introduction to Modern Psychology: The Control Theory View*, Richard J. Robertson, Chicago, iii + 316 pp.

1989i *Living Control Systems: Selected Papers of William T. Powers*, The Control Systems Group, Gravel Switch, Kentucky, xx + 300 pp.

1989j "Powers contributions," *Control Systems Group Newsletter*, November 3, 3.

1989k (with Richard S. Marken) "Random-walk chemotaxis: Trial and error as a control process," *Behavioral Neuroscience 103*(6), December, 1348-1355.

1990a "Dewey, the libertarians, the determinists, and control theory," *The Control Systems Group, Inc. Newsletter*, February, 4-6.

1990b (editor, with Richard J. Robertson) *Introduction to Modern Psychology: The Control-Theory View*, The Control Systems Group, Gravel Switch, Kentucky, x + 238 pp.

1990c "Control theory: A model of organisms," *System Dynamics Review 6*(1), Winter, 1-20.

1990d "Control theory and statistical generalizations," *American Behavioral Scientist 34*(1), September/October, 24-31.

1990e "From Bill Powers," *The Control Systems Group, Inc. Newsletter*, October, 1.

1991a "Commentary on Bandura's 'human agency,'" *American Psychologist 46*(2), February, 151-153.

1991b "Skinner's Mistake," *The Control Systems Group, Inc. Newsletter*, February, 7-14.

1991c (with Tom Bourbon and Ed Ford) "The method of levels and internal conflict," *Closed Loop* (2), May, 1-13.

1991d (with Gary Cziko, Wayne Hershberger, and Rick Marken) "'Conditioning,'" *Closed Loop* (2), May, 13-24.

1991e (with Izhak Bar-Kana, Gary Cziko, Ed Ford, Joel Judd, Rick Marken, Mark Olson, Mary Powers, and Chuck Tucker) "Competition, morals, religion, and science," *Closed Loop* (3), July, 1-35.

1991f (with Izhak Bar-Kana, Tom Bourbon, Gary Cziko, Wayne Hershberger, and Rick Marken) "A (control) engineer among the psychologists," *Closed Loop* (4), October, 1-43.

1991g (with Ed Ford, Dag Forssell, David Goldstein, and Rick Marken) "'Reality therapy' and 'experiential therapy,'" *Closed Loop* (4), October, 43-51.

1991h "From Bill Powers," *The Control Systems Group, Inc. Newsletter*, November, 3.

PSYCHOLOGY/BIOLOGY

LIVING CONTROL SYSTEMS II

Selected Papers of William T. Powers

"Powers turned the millennia-old idea that living systems act to produce intended perceptions into a formal theory of behavior: perceptual control theory. Perceptual control theory identifies behavior as the necessarily variable means by which organisms control their perceptions of the world. Working first on a build-it-yourself computer, then on a first-generation IBM personal computer, Powers created elegant demonstrations in which the simple-idea-turned-formal-model generates remarkably accurate quantitative simulations and predictions of behavior and its consequences. He identified a first principle for behavioral, social, and life sciences and showed the way to a new foundation of theory and method."

—from the Foreword by W. Thomas Bourbon

All of the papers in this second collection of writings by William T. Powers are published here for the first time. Their underlying theme, developed over three decades, is control theory's potential for revolutionizing the life sciences, particularly psychology. Powers critiques the theories of mainstream behavioral scientists, showing how their defects are avoided by applying control theory, instead. He also demonstrates the necessity for constructing truly *generative* models if a genuine science of living control systems is to be developed.

Published by
The Control Systems Group
460 Black Lick Rd.
Gravel Switch, KY 40328

ISBN 0-9624154-2-1