Heylighen, Francis and Joslyn, Cliff: (2003) "Cybernetics and Second Order Cybernetics", in: Encyclopedia of Physical Science and Technology, v. 4, ed. RA Meyers, pp. 155-170, Academic Press, New York, https://doi.org/10.1016/B0-12-227410-5/00161-7

Cybernetics and Second-Order Cybernetics

Francis Heylighen
Free University of Brussels

Cliff Joslyn
Los Alamos National Laboratory

I. Historical Development of Cybernetics
II. Relational Concepts
III. Circular Processes
IV. Goal-Directedness and Control
V. Cognition

GLOSSARY

Variety A measure of the number of possible states or actions.
Entropy A probabilistic measure of variety.
Self-organization The spontaneous reduction of entropy in a dynamic system.
Control Maintenance of a goal by active compensation of perturbations.
Model A representation of processes in the world that allows predictions.
Constructivism The philosophy that models are not passive reflections of reality, but active constructions by the subject.

CYBERNETICS is the science that studies the abstract principles of organization in complex systems. It is concerned not so much with what systems consist of, but how they function. Cybernetics focuses on how systems use information, models, and control actions to steer toward and maintain their goals, while counteracting various disturbances. Being inherently transdisciplinary, cybernetic reasoning can be applied to understand model and design systems of any kind: physical, technological, biological, ecological, psychological, social, or any combination of those. Second-order cybernetics in particular studies the role of the (human) observer in the construction of models of systems and other observers.

I. HISTORICAL DEVELOPMENT OF CYBERNETICS

A. Origins

Derived from the Greek *kybernetes*, or "steersman," the term "cybernetics" first appears in Antiquity with Plato and in the 19th century with Ampère, who both saw it as the science of effective government. The concept was revived and elaborated by the mathematician Norbert Wiener in his seminal 1948 book, whose title defined it as "Cybernetics, or the study of control and communication in the animal

and the machine." Inspired by wartime and prewar results in mechanical control systems such as servomechanisms and artillery targeting systems, and the contemporaneous development of a mathematical theory of communication (or information) by Claude Shannon, Wiener set out to develop a general theory of organizational and control relations in systems.

Information Theory, Control Theory, and Control Systems Engineering have since developed into independent disciplines. What distinguishes cybernetics is its emphasis on control and communication not only in engineered, artificial systems, but also in evolved, natural systems such as organisms and societies, which set their own goals, rather than being controlled by their creators.

Cybernetics as a specific field grew out of a series of interdisciplinary meetings held from 1944 to 1953 that brought together a number of noted postwar intellectuals, including Wiener, John von Neumann, Warren McCulloch, Claude Shannon, Heinz von Foerster, W. Ross Ashby, Gregory Bateson, and Margaret Mead. Hosted by the Josiah Macy Jr. Foundation, these became known as the Macy Conferences on Cybernetics. From its original focus on machines and animals, cybernetics quickly broadened to encompass minds (e.g., in the work of Bateson and Ashby) and social systems (e.g., Stafford Beer's management cybernetics), thus recovering Plato's original focus on the control relations in society.

Through the 1950s, cybernetic thinkers came to cohere with the school of General Systems Theory (GST), founded at about the same time by Ludwig von Bertalanffy, as an attempt to build a unified science by uncovering the common principles that govern open, evolving systems. GST studies systems at all levels of generality, whereas Cybernetics focuses more specifically on goal-directed, functional systems which have some form of control relation. While there remain arguments over the relative scope of these domains, each can be seen as part of an overall attempt to forge a transdisciplinary "Systems Science."

Perhaps the most fundamental contribution of cybernetics is its explanation of purposiveness, or goal-directed behavior, an essential characteristic of mind and life, in terms of control and information. Negative feedback control loops which try to achieve and maintain goal states were seen as basic models for the autonomy characteristic of organisms: their behavior, while purposeful, is not strictly determined by either environmental influences or internal dynamical processes. They are in some sense "independent actors" with a "free will." Thus cybernetics foresaw much current work in robotics and autonomous agents. Indeed, in the popular mind, "cyborgs" and "cybernetics" are just fancy terms for "robots" and "robotics." Given the technological advances of the postwar period, early cyberneticians were eager to explore the similarities between technological and biological systems. Armed with a theory of information, early digital circuits, and Boolean logic, it was unavoidable that they would hypothesize digital systems as models of brains, and information as the "mind" to the machine's "body."

More generally, cybernetics had a crucial influence on the birth of various modern sciences: control theory, computer science, information theory, automata theory, artificial intelligence and artificial neural networks, cognitive science, computer modeling and simulation science, dynamical systems, and artificial life. Many concepts central to these fields, such as complexity, self-organization, self-reproduction, autonomy, networks, connectionism, and adaptation, were first explored by cybernetics during the 1940s and 1950s. Examples include von Neumann's computer architectures, game theory, and cellular automata; Ashby's and von Foerster's analysis of self-organization; Braitenberg's autonomous robots; and McCulloch's artificial neural nets, perceptrons, and classifiers.

B. Second-Order Cybernetics

Cybernetics had from the beginning been interested in the similarities between autonomous, living systems and machines. In this postwar era, the fascination with the new control and computer technologies tended to focus attention on the engineering approach, where it is the system designer who determines what the system will do. However, after the control engineering and computer science disciplines had become fully independent, the remaining cyberneticists felt the need to clearly distinguish themselves from these more mechanistic approaches, by emphasizing autonomy, self-organization, cognition, and the role of the observer in modeling a system. In the early 1970s this movement became known as *second-order cybernetics*.

They began with the recognition that all our knowledge of systems is mediated by our simplified representations—or *models*—of them, which necessarily ignore those aspects of the system which are irrelevant to the purposes for which the model is constructed. Thus the properties of the systems themselves must be distinguished from those of their models, which depend on us as their creators. An engineer working with a mechanical system, on the other hand, almost always know its internal structure and behavior to a high degree of accuracy, and therefore tends to de-emphasize the system/model distinction, acting as if the model *is* the system.

Moreover, such an engineer, scientist, or "first-order" cyberneticist, will study a system as if it were a passive, objectively given "thing," that can be freely observed, manipulated, and taken apart. A second-order cyberneticist working with an organism or social system, on the other

hand, recognizes that system as an agent in its own right, interacting with another agent, the observer. As quantum mechanics has taught us, observer and observed cannot be separated, and the result of observations will depend on their interaction. The observer too is a cybernetic system, trying to construct a model of another cybernetic system. To understand this process, we need a "cybernetics of cybernetics," i.e., a "meta" or "second-order" cybernetics.

These cyberneticians' emphasis on such epistemological, psychological, and social issues was a welcome complement to the reductionist climate which followed on the great progress in science and engineering of the day. However, it may have led them to overemphasize the novelty of their "second-order" approach. First, it must be noted that most founding fathers of cybernetics, such as Ashby, McCulloch, and Bateson, explicitly or implicitly agreed with the importance of autonomy, self-organization, and the subjectivity of modeling. Therefore, they can hardly be portrayed as "first-order" reductionists. Second, the intellectual standard bearers of the second-order approach during the 1970s, such as von Foerster, Pask, and Maturana, were themselves directly involved in the development of "first-order" cybernetics in the 1950s and 1960s. In fact, if we look more closely at the history of the field, we see a continuous development toward a stronger focus on autonomy and the role of the observer, rather than a clean break between generations or approaches. Finally, the second-order perspective is now firmly ingrained in the foundations of cybernetics overall. For those reasons, the present article will discuss the basic concepts and principles of cybernetics as a whole, without explicitly distinguishing between "first-order" and "second-order" ideas, and introduce cybernetic concepts through a series of models of classes of systems.

It must further be noted that the sometimes ideological fervor driving the second-order movement may have led a bridge too far. The emphasis on the irreducible complexity of the various system-observer interactions and on the subjectivity of modeling has led many to abandon formal approaches and mathematical modeling altogether, limiting themselves to philosophical or literary discourses. It is ironic that one of the most elegant computer simulations of the second-order idea that models affect the very system they are supposed to model was not created by a cyberneticist, but by the economist Brian Arthur. Moreover, some people feel that the second-order fascination with self-reference and observers observing observers observing themselves has fostered a potentially dangerous detachment from concrete phenomena.

C. Cybernetics Today

In spite of its important historical role, cybernetics has not really become established as an autonomous discipline. Its practitioners are relatively few, and not very well organized. There are few research departments devoted to the domain, and even fewer academic programs. There are many reasons for this, including the intrinsic complexity and abstractness of the subject domain, the lack of up-to-date textbooks, the ebb and flow of scientific fashions, and the apparent overreaching of the second-order movement. But the fact that the Systems Sciences (including General Systems Theory) are in a similar position indicates that the most important cause is the difficulty of maintaining the coherence of a broad, interdisciplinary field in the wake of the rapid growth of its more specialized and application-oriented "spin-off" disciplines, such as computer science, artificial intelligence, neural networks, and control engineering, which tended to sap away enthusiasm, funding, and practitioners from the more theoretical mother field.

Many of the core ideas of cybernetics have been assimilated by other disciplines, where they continue to influence scientific developments. Other important cybernetic principles seem to have been forgotten, though, only to be periodically rediscovered or reinvented in different domains. Some examples are the rebirth of neural networks, first invented by cyberneticists in the 1940s, in the late 1960s and again in the late 1980s; the rediscovery of the importance of autonomous interaction by robotics and AI in the 1990s; and the significance of positive feedback effects in complex systems, rediscovered by economists in the 1990s. Perhaps the most significant recent development is the growth of the *complex adaptive systems* movement, which, in the work of authors such as John Holland, Stuart Kauffman, and Brian Arthur and the subfield of *artificial life*, has used the power of modern computers to simulate and thus experiment with and develop many of the ideas of cybernetics. It thus seems to have taken over the cybernetics banner in its mathematical modeling of complex systems across disciplinary boundaries, however, while largely ignoring the issues of goal-directedness and control.

More generally, as reflected by the ubiquitous prefix "cyber," the broad cybernetic philosophy that systems are defined by their abstract relations, functions, and information flows, rather than by their concrete material or components, is starting to pervade popular culture, albeit it in a still shallow manner, driven more by fashion than by deep understanding. This has been motivated primarily by the explosive growth of information-based technologies including automation, computers, the Internet, virtual reality, software agents, and robots. It seems likely that as the applications of these technologies become increasingly complex, far-reaching, and abstract, the need will again be felt for an encompassing conceptual framework, such as cybernetics, that can help users and designers alike to understand the meaning of these developments.

Cybernetics as a theoretical framework remains a subject of study for a few committed groups, such as the *Principia Cybernetica Project*, which tries to integrate cybernetics with evolutionary theory, and the *American Society for Cybernetics*, which further develops the second-order approach. The *sociocybernetics* movement actively pursues a cybernetic understanding of social systems. The cybernetics-related programs on autopoiesis, systems dynamics, and control theory also continue, with applications in management science and even psychological therapy. Scattered research centers, particularly in Central and Eastern Europe, are still devoted to specific technical applications, such as biological cybernetics, medical cybernetics, and engineering cybernetics, although they tend to keep closer contact with their field of application than with the broad theoretical development of cybernetics. General Information Theory has grown as the search for formal representations which are not based strictly on classical probability theory.

There has also been significant progress in building a semiotic theory of information, where issues of the semantics and meaning of signals are at last being seriously considered. Finally, a number of authors are seriously questioning the limits of mechanism and formalism for interdisciplinary modeling in particular, and science in general. The issues here thus become what the ultimate limits on knowledge might be, especially as expressed in mathematical and computer-based models. What's at stake is whether it is possible, in principle, to construct models, whether formal or not, which will help us understand the full complexity of the world around us.

II. RELATIONAL CONCEPTS

A. Distinctions and Relations

In essence, cybernetics is concerned with those properties of systems that are independent of their concrete material or components. This allows it to describe physically very different systems, such as electronic circuits, brains, and organizations, with the same concepts, and to look for isomorphisms between them. The only way to abstract a system's physical aspects or components while still preserving its essential structure and functions is to consider relations: How do the components differ from or connect to each other? How does the one transform into the other?

To approach these questions, cyberneticians use high level concepts such as *order, organization, complexity, hierarchy, structure, information*, and *control*, investigating how these are manifested in systems of different types. These concepts are *relational*, in that they allow us to analyze and formally model different abstract properties of systems and their dynamics, for example, allowing us to ask such questions as whether complexity tends to increase with time.

Fundamental to all of these relational concepts is that of *difference* or *distinction*. In general, cyberneticians are not interested in a phenomenon in itself, but only in the difference between its presence and absence, and how that relates to other differences corresponding to other phenomena. This philosophy extends back to Leibniz, and is expressed most succinctly by Bateson's famous definition of information as "a difference that makes a difference." Any observer necessarily starts by conceptually separating or distinguishing the object of study, the *system*, from the rest of the universe, the *environment*. A more detailed study will go further to distinguish between the presence and absence of various properties (also called dimensions or attributes) of the systems. For example, a system such as billiard ball can have properties, such as a particular color, weight, position, or momentum. The presence or absence of each such property can be represented as a binary, Boolean variable, with two values: "yes," the system has the property, or "no," it does not. G. Spencer Brown, in his book "Laws of Form," has developed a detailed calculus and algebra of distinctions, and shown that this algebra, although starting from much simpler axioms, is isomorphic to the more traditional Boolean algebra.

B. Variety and Constraint

This binary approach can be generalized to a property having multiple discrete or continuous values, for example, which color or what position or momentum. The conjunction of all the values of all the properties that a system at a particular moment has or lacks determines its *state*. For example, a billiard ball can have color *red*, position x and momentum p. The set of all possible states that a system can be in defines its *state space*. An essential component of cybernetic modeling is a quantitative measure for the size of the state space, or the number of distinct states. This measure is called *variety*. Variety represents the freedom the system has in choosing a particular state, and thus the uncertainty we have about which state the system occupies. Variety V is defined as the number of elements in the state space S, or, more commonly, as the logarithm to the basis two of that number: $V = \log_2(|S|)$.

The unit of variety in the logarithmic form is the *bit*. A variety of one bit, $V = 1$, means that the system has two possible states, that is, one difference. In the simplest case of n binary variables, $V = \log_2(2^n) = n$ is therefore equal to the minimal number of independent dimensions. But in general, the variables used to describe a system are neither binary nor independent. For example, if a particular type of berry can, depending on its degree of ripeness, be either small and green or large and red (recognizing only two states of size and color), then the variables "color" and

"size" are completely dependent on each other, and the total variety is 1 bit rather than the 2 you would get if you would count the variables separately.

More generally, if the actual variety of states that the system can exhibit is smaller than the variety of states we can potentially conceive, then the system is said to be constrained. *Constraint C* can be defined as the difference between maximal and actual variety: $C = V_{\max} - V$. Constraint is what reduces our uncertainty about the system's state, and thus allows us to make nontrivial predictions. For example, in the previously cited example if we detect that a berry is small, we can predict that it will also be green. Constraint also allows us to formally model relations, dependencies, or couplings between different systems, or aspects of systems. If you model different systems or different aspects, or dimensions of one system together, then the joint state space is the Cartesian product of the individual state spaces: $S = S_1 \times S_2 \times \cdots S_n$. Constraint on this product space can thus represent the mutual dependency between the states of the subspaces, like in the berry example, where the state in the color space determines the state in the size space, and vice versa.

C. Entropy and Information

Variety and constraint can be measured in a more general form by introducing probabilities. Assume that we do not know the precise state s of a system, but only the probability distribution $P(s)$ that the system would be in state s. Variety can then be expressed through a formula equivalent to *entropy*, as defined by Boltzmann for statistical mechanics:

$$H(P) = -\sum_{s \in S} P(s) \cdot \log P(s). \qquad (1)$$

H reaches its maximum value if all states are equiprobable, that is, if we have no indication whatsoever to assume that one state is more probable than another state. Thus it is natural that in this case entropy H reduces to variety V. Again, H expresses our uncertainty or ignorance about the system's state. It is clear that $H = 0$, if and only if the probability of a certain state is 1 (and of all other states 0). In that case we have maximal certainty or complete information about what state the system is in.

We defined constraint as that which reduces uncertainty, that is, the difference between maximal and actual uncertainty. This difference can also be interpreted in a different way, as *information*, and historically H was introduced by Shannon as a measure of the capacity for information transmission of a communication channel. Indeed, if we get some information about the state of the system (e.g., through observation), then this will reduce our uncertainty about the system's state, by excluding—or reducing the probability of—a number of states. The information I we receive from an observation is equal to the degree to which uncertainty is reduced: $I = H(\text{before}) - H(\text{after})$. If the observation completely determines the state of the system ($H(\text{after}) = 0$), then information I reduces to the initial entropy or uncertainty H.

Although Shannon came to disavow the use of the term "information" to describe this measure, because it is purely syntactic and ignores the *meaning* of the signal, his theory came to be known as Information Theory nonetheless. H has been vigorously pursued as a measure for a number of higher-order relational concepts, including complexity and organization. Entropies, correlates to entropies, and correlates to such important results as Shannon's 10th Theorem and the Second Law of Thermodynamics have been sought in biology, ecology, psychology, sociology, and economics.

We also note that there are other methods of weighting the state of a system which do not adhere to probability theory's additivity condition that the sum of the probabilities must be 1. These methods, involving concepts from fuzzy systems theory and possibility theory, lead to alternative information theories. Together with probability theory these are called Generalized Information Theory (GIT). While GIT methods are under development, the probabilistic approach to information theory still dominates applications.

D. Modeling Dynamics

Given these static descriptions of systems, we can now model their dynamics and interactions. Any process or change in a system's state can be represented as a transformation: $T: S \to S: s(t) \to s(t+1)$. The function T by definition is *one-to-one* or *many-to-one*, meaning that an initial state $s(t)$ is always mapped onto a single state $s(t+1)$. Change can be represented more generally as a relation $R \subset S \times S$, thus allowing the modelling of *one-to-many* or *many-to-many* transformations, where the same initial state can lead to different final states. Switching from states s to probability distributions $P(s)$ allows us to again represent such indeterministic processes by a function: $M: P(s,t) \to P(s,t+1)$. M is a stochastic process, or more precisely, a *Markov chain*, which can be represented by a matrix of transition probabilities: $P(s_j(t+1) \mid s_i(t)) = M_{ij} \in [0, 1]$.

Given these process representations, we can now study the dynamics of variety, which is a central theme of cybernetics. It is obvious that a one-to-one transformation will conserve all distinctions between states and therefore the variety, uncertainty or information. Similarly, a many-to-one mapping will erase distinctions, and thus reduce variety, while an indeterministic, one-to-many mapping will increase variety and thus uncertainty. With a general many-to-many mapping, as represented by a Markov

process, variety can increase or decrease, depending on the initial probability distribution and the structure of the transition matrix. For example, a distribution with variety 0 cannot decrease in variety and will in general increase, while a distribution with maximum variety will in general decrease. In the following sections we will discuss some special cases of this most general of transformations.

With some small extensions, this dynamical representation can be used to model the interactions between systems. A system A affects a system B if the state of B at time $t+1$ is dependent on the state of A at time t. This can be represented as a transformation $T: S_A \times S_B \rightarrow S_B: (s_A(t), s_B(t)) \rightarrow s_B(t+1)$. s_A here plays the role of the *input* of B. In general, B will not only be affected by an outside system A, but in turn affect another (or the same) system C. This can be represented by another a transformation T': $S_A \times S_B \rightarrow S_C: (s_A(t), s_B(t)) \rightarrow s_C(\mathrm{t}+1)$. s_C here plays the role of the *output* of B. For the outside observer, B is a process that transforms input into output. If the observer does not know the states of B, and therefore the precise transformations T and T', then B acts as a *black box*. By experimenting with the sequence of inputs $s_A(t)$, $s_A(t+1)$, $s_A(t+2)$, ..., and observing the corresponding sequence of outputs $s_C(t+1)$, $s_C(t+2)$, $s_C(t+3)$, ..., the observer may try to reconstruct the dynamics of B. In many cases, the observer can determine a state space S_B so that both transformations become deterministic, without being able to directly observe the properties or components of B.

This approach is easily extended to become a full theory of automata and computing machines, and is the foundation of most of modern computer science. This again illustrates how cybernetic modeling can produce useful predictions by only looking at *relations* between variables, while ignoring the physical components of the system.

III. CIRCULAR PROCESSES

In classical, Newtonian science, causes are followed by effects, in a simple, linear sequence. Cybernetics, on the other hand, is interested in processes where an effect feeds back into its very cause. Such circularity has always been difficult to handle in science, leading to deep conceptual problems such as the logical paradoxes of self-reference. Cybernetics discovered that circularity, if modelled adequately, can help us to understand fundamental phenomena, such as self-organization, goal-directedness, identity, and life, in a way that had escaped Newtonian science. For example, von Neumann's analysis of reproduction as the circular process of self-construction anticipated the discovery of the genetic code. Moreover, circular processes are in fact ubiquitous in complex, networked systems such as organisms, ecologies, economies, and other social structures.

A. Self-Application

In simple mathematical terms, circularity can be represented by an equation representing how some phenomenon or variable y is mapped, by a transformation or process f, onto itself:

$$y = f(y). \tag{2}$$

Depending on what y and f stand for, we can distinguish different types of circularities. As a concrete illustration, y might stand for an image, and f for the process whereby a video camera is pointed at the image, the image is registered and transmitted to a TV screen or monitor. The circular relation $y = f(y)$ would then represent the situation where the camera points at the image shown on its own monitor. Paradoxically, the image y in this situation is both cause and effect of the process, it is both object and representation of the object. In practice, such a video loop will produce a variety of abstract visual patterns, often with complex symmetries.

In discrete form, Eq. (2) becomes $y_{t+1} = f(y_t)$. Such equations have been extensively studied as iterated maps, and are the basis of *chaotic dynamics* and *fractal geometry*. Another variation is the equation, well known from quantum mechanics and linear algebra:

$$ky = f(y). \tag{3}$$

The real or complex number k is an *eigenvalue* of f, and y is an *eigenstate*. Eq. (3) reduces to the basic Eq. (2) if $k = 1$ or if y is only defined up to a constant factor. If $k = \exp(2\pi i\, m/n)$, then $f^n(y)$ is again y. Thus, imaginary eigenvalues can be used to model periodic processes, where a system returns to the same state after passing through n intermediate states.

An example of such periodicity is the self-referential statement (equivalent to the liar's paradox): "this statement is false." If we start by assuming that the statement is true, then we must conclude that it is false. If we assume it is false, then we must conclude it is true. Thus, the truth value can be seen to oscillate between true and false, and can perhaps be best conceived as having the equivalent of an imaginary value. Using Spencer Brown's calculus of distinctions, Varela has proposed a similar solution to the problem of self-referential statements.

B. Self-Organization

The most direct application of circularity is where $y \in S$ stands for a system's state in a state space S, and f for a dynamic transformation or process. Equation (2) then

states that y is a fixpoint of the function f, or an *equilibrium* or *absorbing* state of the dynamic system: if the system reaches state y, it will stop changing. This can be generalized to the situation where y stands for a subset of the state space, $y \subset S$. Then, every state of this subset is sent to another state of this subspace: $\forall x \in y$: $f(x) \in y$. Assuming y has no smaller subset with the same property, this means that y is an *attractor* of the dynamics. The field of dynamical systems studies attractors in general, which can have any type of shape or dimension, including zero-dimensional (the equilibrium state discussed above), one-dimensional (a limit cycle, where the system repeatedly goes through the same sequence of states), and fractal (a so-called "strange" attractor).

An attractor y is in general surrounded by a *basin* $B(y)$: a set of states outside y whose evolution necessarily ends up inside: $\forall s \in B(y)$, $s \notin y$, $\exists n$ such that $f^n(s) \in y$. In a deterministic system, every state either belongs to an attractor or to a basin. In a stochastic system there is a third category of states that can end up in either of several attractors. Once a system has entered an attractor, it can no longer reach states outside the attractor. This means that our uncertainty (or statistical entropy) H about the system's state has decreased: we now know for sure that it is not in any state that is not part of the attractor. This spontaneous reduction of entropy or, equivalently, increase in order or constraint, can be viewed as a most general model of *self-organization*.

Every dynamical system that has attractors will eventually end up in one of these attractors, losing its freedom to visit any other state. This is what Ashby called the *principle of self-organization*. He also noted that if the system consists of different subsystems, then the constraint created by self-organization implies that the subsystems have become mutually dependent, or mutually adapted. A simple example is magnetization, where an assembly of magnetic spins that initially point in random directions (maximum entropy), end up all being aligned in the same direction (minimum entropy, or mutual adaptation). Von Foerster added that self-organization can be enhanced by random perturbations ("noise") of the system's state, which speed up the descent of the system through the basin, and makes it leave shallow attractors so that it can reach deeper ones. This is the *order from noise* principle.

C. Closure

The "attractor" case can be extended to the case where y stands for a complete state space. Equation (2) then represents the situation where every state of y is mapped onto another state of y by f. More generally, f might stand for a group of transformations, rather than a single transformation. If f represents the possible dynamics of the system with state space y, under different values of external parameters, then we can say that the system is organizationally closed: it is invariant under any possible dynamical transformation. This requirement of *closure* is implicit in traditional mathematical models of systems. Cybernetics, on the other hand, studies closure explicitly, with the view that systems may be open and closed simultaneously for different kinds of properties f_1 and f_2. Such closures give systems an unambiguous *identity*, explicitly distinguishing what is inside from what is outside the system.

One way to achieve closure is self-organization, leaving the system in an attractor subspace. Another way is to expand the state space y into a larger set y^* so that y^* recursively encompasses all images through f of elements of y: $\forall x \in y: x \in y^*$; $\forall x' \in y^*$: $f(x') \in y^*$. This is the traditional definition of a set y^* through *recursion*, which is frequently used in computer programming to generate the elements of a set y^* by iteratively applying transformations to all elements of a starting set y.

A more complex example of closure is *autopoiesis* ("self-production"), the process by which a system recursively produces its own network of physical components, thus continuously regenerating its essential organization in the face of wear and tear. Note that such "organizational" closure is not the same as thermodynamic closure: the autopoietic system is open to the exchange of matter and energy with its environment, but it is autonomously responsible for the way these resources are organized. Maturana and Varela have postulated autopoiesis to be the defining characteristic of living systems. Another fundamental feature of life, *self-reproduction*, can be seen as a special case of autopoiesis, where the self-produced components are not used to rebuild the system, but to assemble a copy of it. Both reproduction and autopoiesis are likely to have evolved from an autocatalytic cycle, an organizationally closed cycle of chemical processes such that the production of every molecule participating in the cycle is catalysed by another molecule in the cycle.

D. Feedback Cycles

In addition to looking directly at a state y, we may focus on the *deviation* $\Delta y = (y - y_0)$ of y from some given (e.g., equilibrium) state y_0, and at the "feedback" relations through which this deviation depends on itself. In the simplest case, we could represent this as $\Delta y(t + \Delta t) = k\, \Delta y(t)$. According to the sign of the dependency k, two special cases can be distinguished.

If a positive deviation at time t (increase with respect to y_0) leads to a negative deviation (decrease with respect to y_0) at the following time step, the feedback is *negative*

($k < 0$). For example, more rabbits eat more grass, and therefore less grass will be left to feed further rabbits. Thus, an increase in the number of rabbits above the equilibrium value will lead, via a decrease in the supply of grass, at the next time step to a decrease in the number of rabbits. Complementarily, a decrease in rabbits leads to an increase in grass, and thus again to an increase in rabbits. In such cases, any deviation from y_0 will be suppressed, and the system will spontaneously return to equilibrium. The equilibrium state y_0 is *stable*, resistant to perturbations. Negative feedback is ubiquitous as a control mechanism in machines of all sorts, in organisms (for example, in homeostasis and the insulin cycle), in ecosystems, and in the supply/demand balance in economics.

The opposite situation, where an increase in the deviation produces further increases, is called *positive* feedback. For example, more people infected with the cold virus will lead to more viruses being spread in the air by sneezing, which will in turn lead to more infections. An equilibrium state surrounded by positive feedback is necessarily unstable. For example, the state where no one is infected is an unstable equilibrium, since it suffices that one person become infected for the epidemic to spread. Positive feedbacks produce a runaway, explosive growth, which will only come to a halt when the necessary resources have been completely exhausted. For example, the virus epidemic will only stop spreading after all people that could be infected have been infected. Other examples of such positive feedbacks are arms races, snowball effects, increasing returns in economics, and the chain reactions leading to nuclear explosions. While negative feedback is the essential condition for stability, positive feedbacks are responsible for growth, self-organization, and the amplification of weak signals. In complex, hierarchical systems, higher-level negative feedbacks typically constrain the growth of lower-level positive feedbacks.

The positive and negative feedback concepts are easily generalized to networks of multiple causal relations. A causal link between two variables, $A \rightarrow B$ (e.g., *infected people* $\rightarrow$ *viruses*), is positive if an increase (decrease) in A produces an increase (decrease) in B. It is negative, if an increase produces a decrease, and vice versa. Each loop in a causal network can be given an overall sign by multiplying the signs ($+$ or $-$) of each of its links. This gives us a simple way to determine whether this loop will produce stabilization (negative feedback) or a runaway process (positive feedback). In addition to the sign of a causal connection, we also need to take into account the *delay* or *lag* between cause and effect, e.g., the rabbit population will only start to increase several weeks after the grass supply has increased. Such delays may lead to an oscillation, or limit cycle, around the equilibrium value.

Such networks of interlocking positive and negative feedback loops with lags are studied in the mathematical field of System Dynamics, a broad program modelling complex biological, social, economic and psychological systems. System Dynamics' most well-known application is probably the "Limits to Growth" program popularized by the Club of Rome, which continued the pioneering computer simulation work of Jay Forrester. System dynamics has since been popularized in the Stella software application and computer games such as SimCity.

IV. GOAL-DIRECTEDNESS AND CONTROL

A. Goal-Directedness

Probably the most important innovation of cybernetics is its explanation of goal-directedness or purpose. An autonomous system, such as an organism, or a person, can be characterized by the fact that it pursues its own goals, resisting obstructions from the environment that would make it deviate from its preferred state of affairs. Thus, goal-directedness implies regulation of—or control over—perturbations. A room in which the temperature is controlled by a thermostat is the classic simple example. The setting of the thermostat determines the preferred temperature or goal state. Perturbations may be caused by changes in the outside temperature, drafts, opening of windows or doors, etc. The task of the thermostat is to minimize the effects of such perturbations, and thus to keep the temperature as much as possible constant with respect to the target temperature.

On the most fundamental level, the goal of an autonomous or autopoietic system is survival, that is, maintenance of its essential organization. This goal has been built into all living systems by natural selection: those that were not focused on survival have simply been eliminated. In addition to this primary goal, the system will have various subsidiary goals, such as keeping warm or finding food, that indirectly contribute to its survival. Artificial systems, such as thermostats and automatic pilots, are not autonomous: their primary goals are constructed in them by their designers. They are *allopoietic*: their function is to produce something other ("allo") than themselves.

Goal-directedness can be understood most simply as suppression of deviations from an invariant goal state. In that respect, a goal is similar to a stable equilibrium, to which the system returns after any perturbation. Both goal-directedness and stability are characterized by *equifinality*: different initial states lead to the same final state, implying the destruction of variety. What distinguishes them is that a stable system automatically returns to its equilibrium state, without performing any work or effort. But a goal-directed

system must actively intervene to achieve and maintain its goal, which would not be an equilibrium otherwise.

Control may appear essentially conservative, resisting all departures from a preferred state. But the net effect can be very dynamic or progressive, depending on the complexity of the goal. For example, if the goal is defined as the distance relative to a moving target, or the rate of increase of some quantity, then suppressing deviation from the goal implies constant change. A simple example is a heat-seeking missile attempting to reach a fast moving enemy plane.

A system's "goal" can also be a subset of acceptable states, similar to an attractor. The dimensions defining these states are called the *essential variables*, and they must be kept within a limited range compatible with the survival of the system. For example, a person's body temperature must be kept within a range of approximately 35–40°C. Even more generally, the goal can be seen as a gradient, or "fitness" function, defined on state space, which defines the degree of "value" or "preference" of one state relative to another one. In the latter case, the problem of control becomes one of on-going optimization or maximization of fitness.

B. Mechanisms of Control

While the perturbations resisted in a control relation can originate either inside (e.g., functioning errors or quantum fluctuations) or outside of the system (e.g., attack by a predator or changes in the weather), functionally we can treat them as if they all come from the same, external source. To achieve its goal in spite of such perturbations, the system must have a way to block their effect on its essential variables. There are three fundamental methods to achieve such regulation: buffering, feedback and feedforward (see Fig. 1).

Buffering is the passive absorption or damping of perturbations. For example, the wall of the thermostatically controlled room is a buffer: the thicker or the better insulated it is, the less effect fluctuations in outside temperature will have on the inside temperature. Other examples are the shock absorbers in a car, and a reservoir, which provides a regular water supply in spite of variations in rain

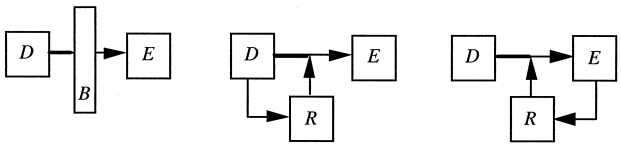

FIGURE 1 Basic mechanisms of regulation, from left to right: buffering, feedforward and feedback. In each case, the effect of disturbances *D* on the essential variables *E* is reduced, either by a passive buffer *B*, or by an active regulator *R*.

fall. The mechanism of buffering is similar to that of a stable equilibrium: dissipating perturbations without active intervention. The disadvantage is that it can only dampen the effects of uncoordinated fluctuations; it cannot systematically drive the system to a non-equilibrium state, or even keep it there. For example, however well-insulated, a wall alone cannot maintain the room at a temperature higher than the average outside temperature.

Feedback and feedforward both require *action* on the part of the system to suppress or compensate the effect of the fluctuation. For example, the thermostat will counteract a drop in temperature by switching on the heating. *Feedforward* control will suppress the disturbance *before* it has had the chance to affect the system's essential variables. This requires the capacity to *anticipate* the effect of perturbations on the system's goal. Otherwise the system would not know which external fluctuations to consider as perturbations, or how to effectively compensate their influence before it affects the system. This requires that the control system be able to gather early information about these fluctuations.

For example, feedforward control might be applied to the thermostatically controlled room by installing a temperature sensor outside of the room, which would warn the thermostat about a drop in the outside temperature, so that it could start heating before this would affect the inside temperature. In many cases, such advance warning is difficult to implement, or simply unreliable. For example, the thermostat might start heating the room, anticipating the effect of outside cooling, without being aware that at the same time someone in the room switched on the oven, producing more than enough heat to offset the drop in outside temperature. No sensor or anticipation can ever provide complete information about the future effects of an infinite variety of possible perturbations, and therefore feedforward control is bound to make mistakes. With a good control system, the resulting errors may be few, but the problem is that they will accumulate in the long run, eventually destroying the system.

The only way to avoid this accumulation is to use feedback, that is, compensate an error or deviation from the goal *after* it has happened. Thus feedback control is also called *error-controlled regulation*, since the error is used to determine the control action, as with the thermostat which samples the temperature inside the room, switching on the heating whenever that temperature reading drops lower than a certain reference point from the goal temperature. The disadvantage of feedback control is that it first must allow a deviation or error to appear before it can take action, since otherwise it would not know which action to take. Therefore, feedback control is by definition imperfect, whereas feedforward could in principle, but not in practice, be made error-free.

The reason feedback control can still be very effective is continuity: deviations from the goal usually do not appear at once, they tend to increase slowly, giving the controller the chance to intervene at an early stage when the deviation is still small. For example, a sensitive thermostat may start heating as soon as the temperature has dropped one tenth of a degree below the goal temperature. As soon as the temperature has again reached the goal, the thermostat switches off the heating, thus keeping the temperature within a very limited range. This very precise adaptation explains why thermostats in general do not need outside sensors, and can work purely in feedback mode. Feedforward is still necessary in those cases where perturbations are either discontinuous, or develop so quickly that any feedback reaction would come too late. For example, if you see someone pointing a gun in your direction, you would better move out of the line of fire immediately, instead of waiting until you feel the bullet making contact with your skin.

C. The Law of Requisite Variety

Control or regulation is most fundamentally formulated as a reduction of variety: perturbations with high variety affect the system's internal state, which should be kept as close as possible to the goal state, and therefore exhibit a low variety. So in a sense control prevents the transmission of variety from environment to system. This is the opposite of information transmission, where the purpose is to maximally conserve variety.

In active (feedforward and/or feedback) regulation, each disturbance from D will have to be compensated by an appropriate counteraction from the regulator R (Fig. 1). If R would react in the same way to two different disturbances, then the result would be two different values for the essential variables, and thus imperfect regulation. This means that if we wish to completely block the effect of D, the regulator must be able to produce at least as many counteractions as there are disturbances in D. Therefore, the variety of R must be at least as great as the variety of D. If we moreover take into account the constant reduction of variety K due to buffering, the principle can be stated more precisely as:

$$V(E) \geq V(D) - V(R) - K \qquad (4)$$

Ashby has called this principle the *law of requisite variety*: in active regulation *only variety can destroy variety*. It leads to the somewhat counterintuitive observation that the regulator must have a sufficiently *large* variety of actions in order to ensure a sufficiently *small* variety of outcomes in the essential variables E. This principle has important implications for practical situations: since the variety of perturbations a system can potentially be confronted with is unlimited, we should always try maximize its internal variety (or diversity), so as to be optimally prepared for any foreseeable or unforeseeable contigency.

D. Components of a Control System

Now that we have examined control in the most general way, we can look at the more concrete components and processes that constitute a control system, such as a simple thermostat or a complex organism or organization (Fig. 2). As is usual in cybernetics, these components are recognized as *functional*, and may or may not correspond to *structural* units.

The overall scheme is a feedback cycle with two inputs: the goal, which stands for the preferred values of the system's essential variables, and the disturbances, which stand for all the processes in the environment that the system does not have under control but that can affect these variables. The system starts by observing or sensing the variables that it wishes to control because they affect its preferred state. This step of *perception* creates an internal representation of the outside situation. The information in this representation then must be processed in order to determine: (1) in what way it may affect the goal and (2) what is the best reaction to safeguard that goal.

Based on this interpretation, the system then decides on an appropriate action. This action affects some part of the environment, which in turn affects other parts of the environment through the normal causal processes or dynamics of that environment. These dynamics are influenced by the set of unknown variables which we have called the disturbances. This dynamical interaction affects among others the variables that the system keeps under observation. The change in these variables is again perceived by the system, and this again triggers interpretation, decision and action, thus closing the control loop.

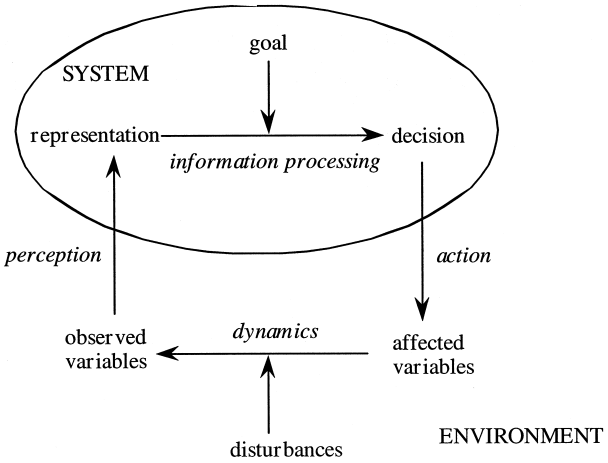

FIGURE 2 Basic components of a control system.

This general scheme of control may include buffering, feedforward and feedback regulation. Buffering is implicit in the dynamics, which determines to what degree the disturbances affect the observed variables. The observed variables must include the essential variables that the system wants to keep under control (feedback or error-controlled regulation) in order to avoid error accumulation. However, they will in general also include various nonessential variables, to function as early warning signals for anticipated disturbances. This implements feedforward regulation.

The components of this scheme can be as simple or as complex as needed. In the thermostat, perception is simply a sensing of the one-dimensional variable room temperature; the goal is the set-point temperature that the thermostat tries to achieve; information processing is the trivial process of deciding whether the perceived temperature is lower than the desired temperature or not; and action consists of either heating, if the temperature is lower, or doing nothing. The affected variable is the amount of heat in the room. The disturbance is the amount of heat exchanged with the outside. The dynamics is the process by which inside heating and heath exchange with the outside determine inside temperature.

For a more complex example, we may consider a board of directors whose goal is to maximize the long-term revenue of their company. Their actions consist of various initiatives, such as publicity campaigns, hiring managers, starting up production lines, saving on administrative costs, etc. This affects the general functioning of the company. But this functioning is also affected by factors that the board does not control such as the economic climate, the activities of competitors, the demands of the clients, etc. Together these disturbances and the initiatives of the board determine the success of the company, which is indicated by variables such as amount of orders, working costs, production backlog, company reputation, etc. The board, as a control system, will interpret each of these variables with reference to their goal of maximizing profits, and decide about actions to correct any deviation from the preferred course.

Note that the control loop is completely symmetric: if the scheme in Fig. 2 is rotated over 180°, environment becomes system while disturbance becomes goal, and vice versa. Therefore, the scheme could also be interpreted as two interacting systems, each of which tries to impose its goal on the other one. If the two goals are incompatible, this is a model of conflict or competition; otherwise, the interaction may settle into a mutually satisfactory equilibrium, providing a model of compromise or cooperation.

But in control we generally mean to imply that one system is more powerful than the other one, capable of suppressing any attempt by the other system to impose *its* preferences. To achieve this, an *asymmetry* must be built into the control loop: the actions of the system (controller) must have more effect on the state of the environment (controlled) than the other way around. This can also be viewed as an *amplification* of the signal travelling through the control system: weak perceptual signals, carrying information but almost no energy, lead to powerful actions, carrying plenty of energy. This asymmetry can be achieved by weakening the influence of the environment, e.g., by buffering its actions, and by strengthening the actions of the system, e.g., by providing it with a powerful energy source. Both cases are illustrated by the thermostat: the walls provide the necessary insulation from outside perturbations, and the fuel supply provides the capacity to generate enough heat. No thermostatic control would be possible in a room without walls or without energy supply. The same requirements applied to the first living cells, which needed a protective membrane to buffer disturbances, and a food supply for energy.

E. Control Hierarchies

In complex control systems, such as organisms or organizations, goals are typically arranged in a hierarchy, where the higher-level goals control the settings for the subsidiary goals. For example, your primary goal of survival entails the lower-order goal of maintaining sufficient hydration, which may activate the goal of drinking a glass of water. This will in turn activate the goal of bringing the glass to your lips. At the lowest level, this entails the goal of keeping your hand steady without spilling water.

Such hierarchical control can be represented in terms of the control scheme of Fig. 2 by adding another layer, as in Fig. 3. The original goal 1 has now itself become the result of an action, taken to achieve the higher level goal 2. For example, the thermostat's goal of keeping the temperature at its set-point can be subordinated to the higher order goal of keeping the temperature pleasant to the people present without wasting fuel. This can be implemented by adding an infrared sensor that perceives whether there are people present in the room, and if so, then setting the thermostat at a higher temperature T_1, otherwise setting it to the lower temperature T_2. Such control layers can be added arbitrarily by making the goal at level n dependent on the action at level $n+1$.

A control loop will reduce the variety of perturbations, but it will in general not be able to eliminate all variation. Adding a control loop on top of the original loop may eliminate the residual variety, but if that is not sufficient, another hierarchical level may be needed. The required number of levels therefore depends on the regulatory ability of the individual control loops: the weaker that ability, the more hierarchy is needed. This is Aulin's *law of*

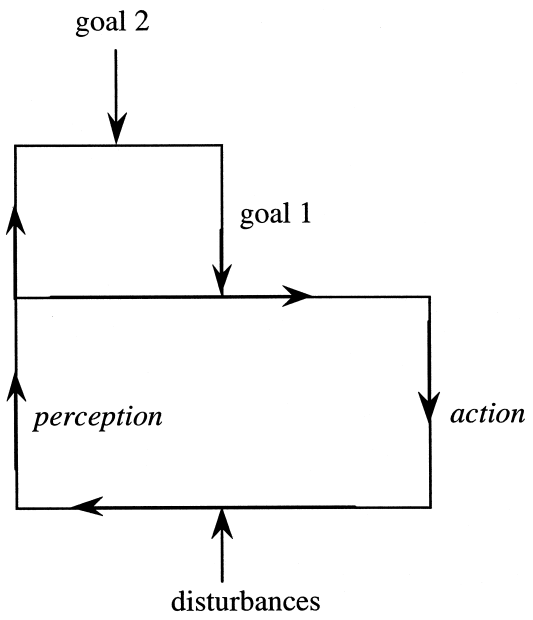

FIGURE 3 A hierarchical control system with two control levels.

requisite hierarchy. On the other hand, increasing the number of levels has a negative effect on the overall regulatory ability, since the more levels the perception and action signals have to pass through, the more they are likely to suffer from noise, corruption, or delays. Therefore, if possible, it is best to maximize the regulatory ability of a single layer, and thus minimize the number of requisite layers. This principle has important applications for social organizations, which have a tendency to multiply the number of bureaucratic levels. The present trend toward the flattening of hierarchies can be explained by the increasing regulatory abilities of individuals and organizations, due to better education, management and technological support.

Still, when the variety becomes really too great for one regulator, a higher control level must appear to allow further progress. Valentin Turchin has called this process a *metasystem transition*, and proposed it as a basic unit, or "quantum," of the evolution of cybernetic systems. It is responsible for the increasing functional complexity which characterizes such fundamental developments as the origins of life, multicellular organisms, the nervous system, learning, and human culture.

V. COGNITION

A. Requisite Knowledge

Control is not only dependent on a requisite variety of actions in the regulator: the regulator must also *know* which action to select in response to a given perturbation. In the simplest case, such knowledge can be represented as a one-to-one mapping from the set D of perceived disturbances to the set R of regulatory actions: $f: D \rightarrow R$, which maps each disturbance to the appropriate action that will suppress it. For example, the thermostat will map the perception "temperature too low" to the action "heat," and the perception "temperature high enough" to the action "do not heat." Such knowledge can also be expressed as a set of production rules of the form "if *condition* (perceived disturbance), then *action*." This "knowledge" is embodied in different systems in different ways, for example, through the specific ways designers have connected the components in artificial systems, or in organisms through evolved structures such as genes or learned connections between neurons as in the brain.

In the absence of such knowledge, the system would have to try out actions blindly, until one would by chance eliminate the perturbation. The larger the variety of disturbances (and therefore of requisite actions), the smaller the likelihood that a randomly selected action would achieve the goal, and thus ensure the survival of the system. Therefore, increasing the variety of actions must be accompanied by increasing the constraint or selectivity in choosing the appropriate action, that is, increasing knowledge. This requirement may be called the *law of requisite knowledge*. Since all living organisms are also control systems, life therefore implies knowledge, as in Maturana's often quoted statement that "to live is to cognize."

In practice, for complex control systems control actions will be neither blind nor completely determined, but more like "educated guesses" that have a reasonable probability of being correct, but without a guarantee of success. Feedback may help the system to correct the errors it thus makes before it is destroyed. Thus, goal-seeking activity becomes equivalent to heuristic problem-solving.

Such incomplete or "heuristic" knowledge can be quantified as the conditional uncertainty of an action from R, given a disturbance in D: $H(R \mid D)$. (This uncertainty is calculated as in Eq. (1), but using conditional probabilities $P(r \mid d)$). $H(R \mid D) = 0$ represents the case of no uncertainty or complete knowledge, where the action is completely determined by the disturbance. $H(R \mid D) = H(R)$ represents complete ignorance. Aulin has shown that the law of requisite variety (4) can be extended to include knowledge or ignorance by simply adding this conditional uncertainty term (which remained implicit in Ashby's non-probabilistic formulation of the law):

$$H(E) \geq H(D) + H(R \mid D) - H(R) - K \qquad (5)$$

This says that the variety in the essential variables E can be reduced by (1) increasing buffering K, (2) increasing variety of action $H(R)$, or (3) decreasing the uncertainty

$H(R \mid D)$ about which action to choose for a given disturbance, that is, increasing knowledge.

B. The Modeling Relation

In the above view of knowledge, the goal is implicit in the condition–action relation, since a different goal would require a different action under the same condition. When we think about "scientific" or "objective" knowledge, though, we conceive of rules that are independent of any particular goal. In higher order control systems that vary their lower order goals, knowledge performs the more general function of making *predictions*: "what will happen if this condition appears and/or that action is performed?" Depending on the answer to that question, the control system can then choose the best action to achieve its present goal.

We can formalize this understanding of knowledge by returning to our concept of a *model*. We now introduce *endo-models*, or models *within* systems, as opposed to our previous usage of *exo-models*, or models *of* systems. Figure 4 shows a model (an endo-model), which can be viewed as a magnification of the feedforward part of the general control system of Fig. 2, ignoring the goal, disturbances and actions.

A model starts with a system to be modeled, which we here call the "world," with state space $W = \{w_i\}$ and dynamics $F_a: W \rightarrow W$. The dynamics represents the temporal evolution of the world, like in Fig. 2, possibly under the influence of an action a by the system. The model itself consists of internal model states, or representations $R = \{r_j\}$ and a modeling function, or set of prediction rules, $M_a: R \rightarrow R$. The two are coupled by a perception function $P: W \rightarrow R$, which maps states of the world onto their representations in the model. The prediction M_a succeeds if it manages to anticipate what will happen to the representation R under the influence of action a. This means that the predicted state of the model $r_2 = M_a(r_1) = M_a(P(w_1)$ must be equal to the state of the model created by perceiving the actual state of the world

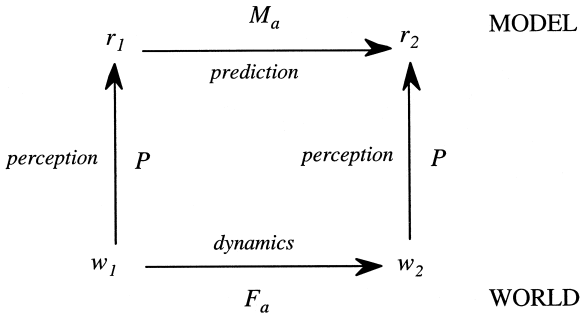

FIGURE 4 The modeling relation.

w_2 after the process $F_a: r_2 = P(w_2) = P(F_a(w_1))$. Therefore, $P(F_a) = M_a(P)$.

We say that the mappings P, M_a and F_a must *commute* for M to be a good model which can predict the behavior of the world W. The overall system can be viewed as a *homomorphic* mapping from states of the world to states of the model, such that their dynamical evolution is preserved. In a sense, even the more primitive "condition-action" rules discussed above can be interpreted as a kind of homomorphic mapping from the events ("disturbances") in the world to the actions taken by the control system. This observation was developed formally by Conant and Ashby in their classic paper "Every good regulator of a system must be a model of that system." Our understanding of "model" here, however, is more refined, assuming that the control system can consider various predicted states $M_a(r_1)$, without actually executing any action a. This recovers the sense of a model as a representation, as used in Section I.B and in science in general, in which observations are used merely to confirm or falsify predictions, while as much as possible avoiding interventions in the phenomenon that is modelled.

An important note must be made about the *epistemology*, or philosophy of knowledge, implied by this understanding. At first sight, defining a model as a homomorphic mapping of the world would seem to imply that there is an objective correspondence between objects in the world and their symbolic representations in the model. This leads back to "naive realism" which sees true knowledge as a perfect reflection of outside reality, independent of the observer. The homomorphism here, however, does not conserve any objective structure of the world, only the type and order of phenomena as perceived by the system. A cybernetic system only perceives what points to potential disturbances of its own goals. It is in that sense intrinsically subjective. It does not care about, nor has it access to, what "objectively" exists in the outside world. The only influence this outside world has on the system's model is in pointing out which models make inaccurate predictions. Since an inaccurate prediction entails poor control, this is a signal for the system to build a better model.

C. Learning and Model-Building

Cybernetic epistemology is in essence *constructivist*: knowledge cannot be passively absorbed from the environment, it must be actively constructed by the system itself. The environment does not instruct, or "in-form," the system, it merely weeds out models that are inadequate, by killing or punishing the system that uses them. At the most basic level, model-building takes place by variation-and-selection or trial-and-error.

Let us illustrate this by considering a primitive aquatic organism whose control structure is a slightly more sophisticated version of the thermostat. To survive, this organism must remain in the right temperature zone, by moving up to warmer water layers or down to colder ones when needed. Its perception is a single temperature variable with 3 states $X = \{too\ hot, too\ cold, just\ right\}$. Its variety of action consists of the 3 states $Y = \{go\ up, go\ down, do\ nothing\}$. The organism's control knowledge consists of a set of perception-action pairs, or a function $f: X \rightarrow Y$. There are $3^3 = 27$ possible such functions, but the only optimal one consists of the rules *too hot* → *go down*, *too cold* → *go up*, and *just right* → *do nothing*. The last rule could possibly be replaced by either *just right* → *go up* or *just right* → *go down*. This would result in a little more expenditure of energy, but in combination with the previous rules would still keep the organism in a negative feedback loop around the ideal temperature. All 24 other possible combinations of rules would disrupt this stabilizing feedback, resulting in a runaway behavior that will eventually kill the organism.

Imagine that different possible rules are coded in the organism's genes, and that these genes evolve through random mutations each time the organism produces offspring. Every mutation that generates one of the 24 combinations with positive feedback will be eliminated by natural selection. The three negative feedback combinations will initially all remain, but because of competition, the most energy efficient combination will eventually take over. Thus internal variation of the control rules, together with natural selection by the environment eventually results in a workable model.

Note that the environment did not *instruct* the organism how to build the model: the organism had to find out for itself. This may still appear simple in our model with 27 possible architectures, but it suffices to observe that for more complex organisms there are typically millions of possible perceptions and thousands of possible actions to conclude that the space of possible models or control architectures is absolutely astronomical. The information received from the environment, specifying that a particular action or prediction is either successful or not, is far too limited to select the right model out of all these potential models. Therefore, the burden of developing an adequate model is largely on the system itself, which will need to rely on various internal heuristics, combinations of preexisting components, and subjective selection criteria to efficiently construct models that are likely to work.

Natural selection of organisms is obviously a quite wasteful method to develop knowledge, although it is responsible for most knowledge that living systems have evolved in their genes. Higher organisms have developed a more efficient way to construct models: *learning*. In learning, different rules compete with each other within the same organism's control structure. Depending on their success in predicting or controlling disturbances, rules are differentially rewarded or reinforced. The ones that receive most reinforcement eventually come to dominate the less successful ones. This can be seen as an application of control at the metalevel, or a metasystem transition, where now the goal is to minimize the perceived difference between prediction and observation, and the actions consist in varying the components of the model.

Different formalisms have been proposed to model this learning process, beginning with Ashby's *homeostat*, which for a given disturbance searched not a space of possible actions, but a space of possible sets of disturbance → action rules. More recent methods include neural networks and genetic algorithms. In genetic algorithms, rules vary randomly and discontinuously, through operators such as mutation and recombination. In neural networks, rules are represented by continuously varying connections between nodes corresponding to sensors, effectors and intermediate cognitive structures. Although such models of learning and adaptation originated in cybernetics, they have now grown into independent specialisms, using labels such as "machine learning" and "knowledge discovery."

D. Constructivist Epistemology

The broad view espoused by cybernetics is that living systems are complex, adaptive control systems engaged in circular relations with their environments. As cyberneticians consider such deep problems as the nature of life, mind, and society, it is natural that they be driven to questions of philosophy, and in particular epistemology.

As we noted, since the system has no access to how the world "really" is, models are subjective constructions, not objective reflections of outside reality. As far as they can know, for knowing systems these models effectively *are* their environments. As von Foerster and Maturana note, in the nervous system there is no *a priori* distinction between a perception and a hallucination: both are merely patterns of neural activation. An extreme interpretation of this view might lead to solipsism, or the inability to distinguish self-generated ideas (dreams, imagination) from perceptions induced by the external environment. This danger of complete relativism, in which any model is considered to be as good as any other, can be avoided by the requirements for *coherence* and *invariance*.

First, although no observation can prove the truth of a model, different observations and models can mutually confirm or support each other, thus increasing their joint reliability. Thus, the more coherent a piece of knowledge is with all other available information, the more reliable it is. Second, percepts appear more "real" as they vary less

between observations. For example, an *object* can be defined as that aspect of a perception that remains invariant when the point of view of the observer is changed. In the formulation of von Foerster, an object is an eigenstate of a cognitive transformation. There is moreover invariance over observers: if different observers agree about a percept or concept, then this phenomenon may be considered "real" by *consensus*. This process of reaching consensus over shared concepts has been called "the social construction of reality." Gordon Pask's Conversation Theory provides a sophisticated formal model of such a "conversational" interaction that ends in an agreement over shared meanings.

Another implication of constructivism is that since all models are constructed by some observer, this observer must be included in the model for it to be complete. This applies in particular to those cases where the process of model-building affects the phenomenon being modelled. The simplest case is where the process of observation itself perturbs the phenomenon, as in quantum measurement, or the "observer effect" in social science. Another case is where the predictions from the model can perturb the phenomenon. Examples are self-fulfilling prophecies, or models of social systems whose application in steering the system changes that very system and thus invalidates the model. As a practical illustration of this principle, the complexity theorist Brian Arthur has simulated the seemingly chaotic behavior of stock exchange-like systems by programming agents that are continuously trying to model the future behavior of the system to which they belong, and use these predictions as the basis for their own actions. The conclusion is that the different predictive strategies cancel each other out, so that the long-term behavior of the system becomes intrinsically unpredictable.

The most logical way to minimize these indeterminacies appears to be the construction of a *metamodel*, which represents various possible models and their relations with observers and the phenomena they represent. For example, as suggested by Stuart Umpleby, one of the dimensions of a metamodel might be the degree to which an observation affects the phenomenon being observed, with classical, observer-independent observations at one extreme, and quantum observation closer to the other extreme.

However, since a metamodel is a still a model, built by an observer, it must represent itself. This is a basic form of self-reference. Generalizing from fundamental epistemological restrictions such as the theorem of Gödel and the Heisenberg indeterminacy principle, Lars Löfgren has formulated a *principle of linguistic complementarity*, which implies that all such self-reference must be partial: languages or models cannot include a complete representation of the process by which their representations are connected to the phenomena they are supposed to describe. Although this means that no model or metamodel can ever be complete, a metamodel still proposes a much richer and more flexible method to arrive at predictions or to solve problems than any specific object model. Cybernetics as a whole could be defined as an attempt to build a universal metamodel, that would help us to build concrete object models for any specific system or situation.

SEE ALSO THE FOLLOWING ARTICLES

ARTIFICIAL INTELLIGENCE • ARTIFICIAL NEURAL NETWORKS • CELLULAR AUTOMATA • COGNITIVE SCIENCES • DISCRETE SYSTEMS MODELING • HUMAN-COMPUTER INTERACTION • HUMANOID ROBOTS • MATHEMATICAL MODELING • ROBOTICS, COMPUTER SIMULATIONS FOR • SELF-ORGANIZING SYSTEMS • SYSTEM THEORY

BIBLIOGRAPHY

Ashby, W. R. (1956–1999). "Introduction to Cybernetics," Methuen, London (electronically republished at http://pcp.vub.ac.be/books/IntroCyb.pdf).

François, C. (ed.) (1997). "International Encyclopedia of Systems and Cybernetics," Saur, Munich.

Geyer, F. (1995). "The challenge of sociocybernetics," *Kybernetes* **24**(4), 6–32.

Heims, S. J. (1991)."Constructing a Social Science for Postwar America: The Cybernetics Group," MIT Press, Cambridge, MA.

Heylighen, F., Joslyn, C., and Turchin, V. (eds.) (1991–2001). "Principia Cybernetica Web" (http://pcp.vub.ac.be).

Holland, J. (1995). "Hidden Order: How Adaptation Builds Complexity," Addison-Wesley, Reading, MA.

Klir, G. (1991). "Facets of Systems Science," Plenum, New York.

Maturana, H., and Varela, F. (1998). "The Tree of Knowledge," (revised edition), Shambhala Press, Boston.

Richardson, G. P. (1991). "Feedback Thought in Social Science and Systems Theory," University of Pennsylvania Press, Philadelphia.

Meystel, A. (1996). "Intelligent systems: A semiotic perspective," *Int. J. Intell. Control Systems* **1,** 31–57

Umpleby, S., and Dent, E. (1999). "The origins and purposes of several traditions in systems theory and cybernetics," *Cybernetics and Systems* **30,** 79–103.

von Foerster, H. (1995). "Cybernetics of Cybernetics" (2nd ed.), FutureSystems Inc., Minneapolis, MN.