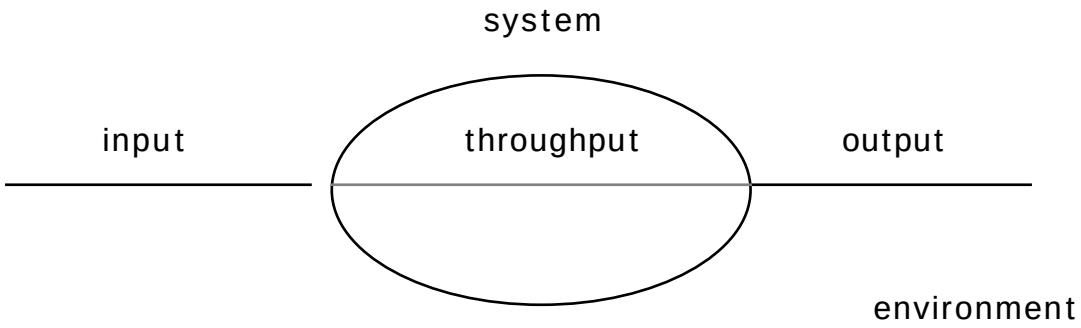
Autonomy and Cognition as the Maintenance and Processing of Distinctions

Francis Heylighen

Research Assistant NFWO
Transdisciplinary Research Group
Vrije Universiteit Brussel (TENA)
Pleinlaan 2, B-1050 Brussels, Belgium

1. Introduction : from classical to new cybernetics

Cybernetics can be viewed as the theory of how to *control* systems. A *system* is supposed to be a coherent whole which can be distinguished or separated from the rest of the universe by its boundary. Yet this does not mean that the system is completely separate or independent of this outside universe. In general there is always an interaction between the system and a part of the external world. This part is usually called 'the environment' of the system. Two components can be distinguished in what was called the 'interaction' : 1) the processes originating in the environment and influencing the system are called the 'input' of the system ; 2) the processes originating inside the system but influencing the environment are called the 'output' of the system (see figure).



The system itself can then be viewed as an process transforming input to output. In this sense the system embodies a causal relation, leading from cause (input) to effect (output). Remark that in this view cause and effect of the process are external to the system (exocausality) ; the system functions merely as a passive channel through which the information concerning the input flows. It can hence be seen as a medium or as an instrument for producing a specific output given a specific input.

For an outside observer, the problem of control or steering can now be formu- lated as follows : how to prepare the input so that the system will produce a specific output ? If the observer is able to make the system produce every output it can produce, then we may say that the observer controls the system.

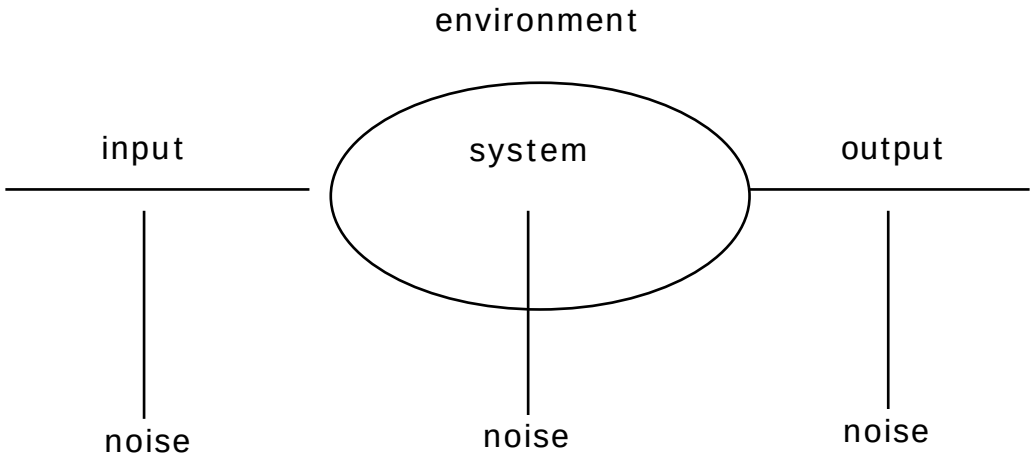
We should remark here that the present description is simplified in the sense that we neglect the aspect of time : we have not distinguished between input (or output) at this moment and input during a period of time. In general it is impossible to produce a specific output at this instant by providing a specific input at the same instant. The reaction of the system to the present input will depend upon the past history of all previous inputs. This history can be summarized by the 'state' of the system.

Before we can produce a specific output we should first prepare the system in the required state by letting it undergo the adequate sequence of inputs. This is called the problem of governability : a system is governable if all of its states can be reached by an adequate input sequence. A complementary problem is that of observability : if we wish to control a system we should be able to observe which state it is in. A system is called observable if it is possible to determine its initial state by looking ats its reaction (output) to a given input (see Mesarovic and Takahara, 1975).

In the present approach we will neglect this aspect of time and speak about input and output without specifying whether these denote sequences or isolated events. (this is equivalent to viewing input as the conjunction of the present input event and the present state, as far as this state is governable or observable; if it were not observable or governable, it could not be 'distinguished' by the observer (see section 2)).

As long as the process which transforms input to output is perfectly causal or deterministic, the problem of control is trivial. If we know the causal mechanism, then we can make the system produce any output we wish by providing it with the corresponding input. This is the domain of the purely mechanical systems, which is described by classical physics. In fact, we do not need any specifically *cybernetic* theory to understand or to steer such systems.

Cybernetics, however, was created to understand the control of systems whose interaction with the environment is not completely predictable. Hence their behavior is incompletely controllable. In classical cybernetics this is conceived as a perturbation of the causal process by unknown external influences (noise). This noise can have an effect at different stages of the process : input, throughput and output (see figure).



To counteract the effect of such noise *classical cybernetics* proposes a set of control mechanisms, such as feedback, feedforward and buffering, which redirect the process towards its intended output by anticipating and compensating the perturbations. This theory has many applications in the design of industrial systems, machines, management strategies, and so on.

However, there remain some basic categories of systems for which this approach is insufficient. These are systems for which the uncontrollability cannot be ascribed to some external source of noise, but for which it is an intrinsic factor. This means that their behavior is (at least partially) caused internally, i.e. independently of the environment (endocausality). They possess some form of *autonomy*. Examples of such systems are self-organizing systems in thermodynamics (dissipative structures), biological organisms, ecologies, human actors, organizations, ... Their lack of external controllability is compensated by some form of internal control of their interaction with the environment (self-steering). This internal control requires knowledge about how to interact with the environment without losing the system's autonomy (see e.g. Maturana anad Varela, 1980). Hence these systems can also be characterized by their capacity for *cognition.*

To understand the behavior of such systems, we need a *new cybernetics* . The aim of the present paper is to give a general formulation of the problem of how to design such a new cybernetics. A new conceptual framework will be sketched which is based on the concept of 'distinction'. This will allow us to distinguish clearly between the 'causal' systems studied by classical cybernetics, and the 'autonomous' systems studied by the new cybernetics. Furthermore, it will provide us with a possible strategy to tackle the general problem : how do autonomous systems function ?

2. Controllability and Distinction Conservation.

We shall now try to analyse different types of controllability by introducing a new concept : *distinction* . A distinction could be defined as the operation (or its result) of separating a class of phenomena from its complement, i.e. from all those phenomena which do not fit into the class. For an outside observer to describe the behavior (input, output) of a system he must be able to

distinguish different types of behavior, so that they can be grouped in classes to which some label can be attached. Different distinctions can be combined by conjunction so that an algebra of classes or distinctions is generated (Spencer-Brown, 1969). This static, logical framework can be extended to describe space-time and processes by introducing morphisms of distinction algebras (Heylighen, 1987a,b). Let us now see how the concept of distinction can be applied to the problem of control.

To control or to steer a certain process means that one is able somehow to choose between alternative outcomes of the process, and that he can ascertain that the chosen outcome will be effectively realized. In the input -output model this signifies that we suppose the observer can freely determine the value of the input variables and that, by selecting the appropriate input, he can make the system produce the desired output, i.e he can determine the value of the output variables.

Call :

$$O = \{ o_1, o_2, o_3, \ldots \}$$

the set of possible output values, and :

$$I = \{ i_1, i_2, i_3, \ldots \}$$

the set of possible input values.

Now a system can be said to be perfectly controllable if there is a mapping f from I onto O :

$$f : I \rightarrow O : i \rightarrow f(i) = o$$

such that each element of O is the image of one element of I , and each element of I is sent upon one element of O (f is a bijective function). In that case each element o of O can be reached by applying f to the appropriate element i of I. f can be said to symbolize a causal relation between the cause i and the effect o. Indeed, the definition of f entails the classical principle of causality :

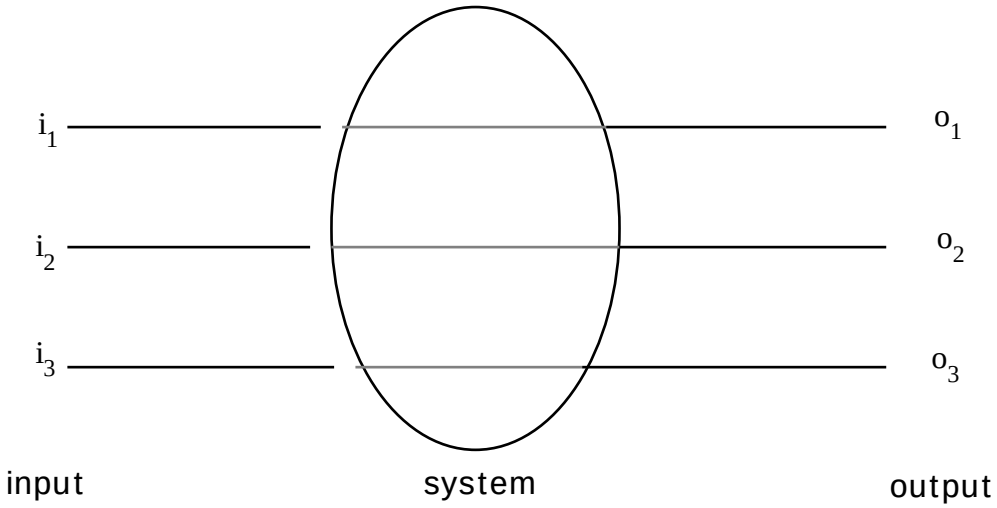
equal causes have equal effects, which may also be formulated as :
distinct causes have distinct effects.

Symbolically :

$$i_1 = i_2 \quad \text{if and only if} \quad f(i_1) = f(i_2),$$

which is equivalent to :

$$i_1 \neq i_2 \quad \text{if and only if} \quad f(i_1) \neq f(i_2).$$



We can now reformulate this property by introducing distinction conservation : *the distinction between i_1 and i_2 is conserved by f.* In other words, if the observer can distinguish between two outputs (inputs) of the system, then he can distinguish between the corresponding inputs (outputs).

It must be remarked that this definition of causality or perfect controllability is dependent upon the observer and the distinctions he makes. In general the symbol i_2 will denote a *class* of physically different phenomena, but which are assimilated by the observer to represent the same type of input. The observer does not distinguish between the members of this class, he only distinguishes between the members of different classes. It is this distinction of classes which is conserved by the causal system symbolized by the function f. Let us look at an example.

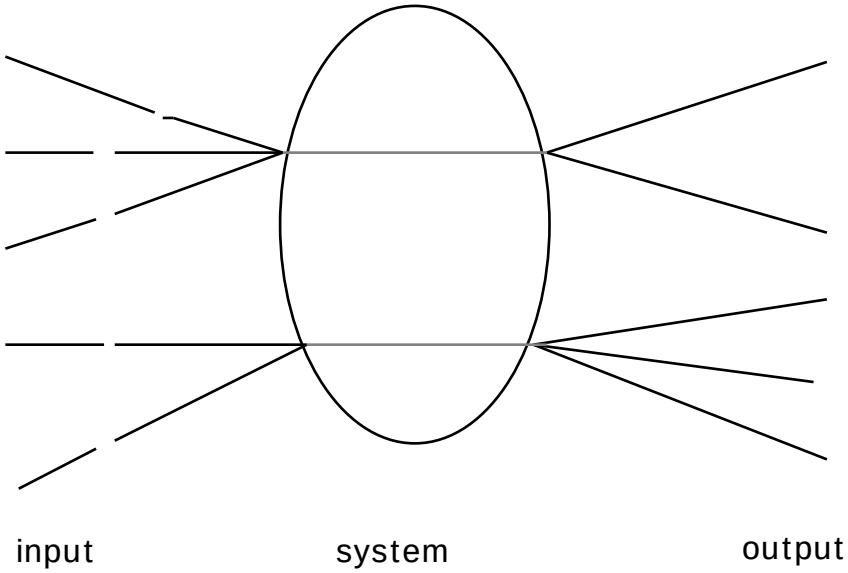
Consider a billiard-ball on a table. The ball can be viewed as a rigid body whose dynamics is determined by the laws of classical mechanics. This means that if we know the total force exerted on the ball, then we can predict its trajectory in a deterministic way. Moreover, its evolution is causal : distinct inputs (forces) correspond to distinct outputs (trajectories). Hence the ball is a perfectly controllable system : every conceivable trajectory can be produced by subjecting the ball to the appropriate force. This is what a skilled billiard-player tries to accomplish : to control the trajectory of the ball by exerting the appropriate force with his billiard-cue.

Yet he cannot control the internal tension on the different molecules constituting the ball. Indeed, the representation of the ball as a rigid body subjected to a total force is only an approximation : the ball is not completely rigid, its molecules can move somewhat relative to one another. This differential movement is not determined by the total force but by the different local forces exerted on the different parts of the ball. Yet in the model used by the billiard-player these differential forces and movements are not distinguished and hence are not controlled. The only thing he is interested in is the class of all differential forces which determine the same total force and hence the same trajectory of the ball considered as a whole.

Now that we have defined perfect controllability as seen by an observer making certain distinctions, we must look at situations where the controllability and hence the distinction conservation is incomplete. Consider for example a far away galaxy : whatever the observer does,

it will have no effect on the behavior of the system. This means that for all possible distinct actions (input) the observer can perform, the reaction (output) of the system will remain the same. There is absolutely no distinction conservation. With respect to the distinctions made by the observer such a system behaves as if it were closed : its behavior cannot be influenced in any way. There is no relation whatsoever between input and output.

The last case to be considered is that of a system which is *partially* distinction conserving. In general this means that there are distinct outputs which can be controlled by the observer, and others which cannot. There is a relation between the input and output sets I and O, but this relation is not a bijection (and in general not even a function). This means that distinct inputs can lead to the same output, and that the same input can lead to distinct outputs. Using some terminology borrowed from thermodynamics and theories of self-organizing systems (Haken, 1978; Prigogine, 1979), we may call the first phenomenon *equifinality* : different initial situations lead to equivalent final situations. In the second case we may speak about *bifurcation* (or branching) : the path leading from input to output bifurcates (see figure).



An example of equifinality can be found in the biological phenomenon of homeostasis : even for different outside conditions an organism succeeds to maintain an equilibrium value for certain parameters which determine its survival. E.g. a warm-blooded animal will maintain a constant body temperature even though the temperature of the environment changes drastically : distinct external temperatures (input) lead to the same internal temperature (output). An example of bifurcation can be found in self-organizing systems : the appearance of a dissipative structure in thermodynamics depends upon the boundary conditions of the system, yet it is not determined by them ; in general distinct structures can appear under the same boundary conditions (Prigogine, 1979). Another example can be found in the behavior of intelligent systems which is sometimes called 'free will' : in general it is impossible to predict how an individual person will react to a given situation ; the decision he will make is not completely determined by the sequence of all the inputs he ever received ; the same person could react to the same situation in distinct ways.

3. Autonomy as boundary maintenance.

A general property of incompletely distinction-conserving systems is that their boundary, which separates input from throughput and throughput from output, cannot be arbitrarily chosen, as is the case for causal systems. As long as all distinctions are conserved, the process which transforms input tot output can be conceived as the continuous transmission of information through the system. The number of bits, which is equal to the number of distinctions if all inputs are equiprobable, remains the same. The form or the medium of the signal may have changed but its information content is conserved. Since the process is continuous from this point of view, the discrimination of input, throughput and output as separate subprocesses is rather arbitrary. The actual choice of the boundary will then depend upon the observer and the objective he thas in mind while controlling the system.

In an incompletely distinction conserving system on the other hand, input distinctions (and hence information) are lost through equifinality. It is logical to choose the (input) boundary at that point of the process where the loss occurs, i.e. at that point where the process is not longer completely controllable. The boundary can then be interpreted as representing an obstacle to the inflow of external control, as something which (partially) insulates the system against outside influences. In an analogous way we can choose the output boundary behind the bifurcation points, i.e. at that point where the process is controllable again. Such a boundary can be spatial (topological), but it need not be so. Let us look at some examples.

Consider a house which is kept at a constant temperature by a heating installation with a thermostat. This means that it is impossible to control the internal temperature by changing the outside temperature : distinct outside temperatures lead to the same inside temperature. It is natural to choose the walls of the house as the boundary of this system. Indeed, it is through the insulation of the walls that the heating installation succeeds to compensate the fluctuations of the external temperature; the transition from externally controlled temperatures to internally controlled temperatures takes place at the walls. Another example of such a spatial boundary is a shell, which protects the shell fish living in it against external attempts to control it (e.g. by a predator which would like to eat it).

However, we may also conceive of boundaries which have no support in physical space. Consider a secret organization A (e.g. a masonic lodge) whose members are also members of another larger organization B (e.g. society at large). Since the same person belongs both to A and to B, it is clearly impossible to find a physical boundary separating A from B. However, if we look at this problem from the viewpoint of information or distinction transmission between systems, it is clear that there is a boundary. The information inside the secret organization is passed between the members but it does not leak outside. A member of A when acting outside the context of A, will not say or do anything which might signal what has happened inside the organization. On the other hand, a person who does not belong to A cannot control or influence what happens inside, even though he may communicate with one of its members. We could say that the barrier or boundary separating A from B resides inside the head of the members of A : they somehow know when to switch from information processes in the context of A to information processes outside this context.

The example of a secret society is perhaps somewhat extreme, but the same reasoning applies to a less degree to any organization where there is more information exchange between the members of the organization than across its boundaries. The general principle we should remember is that partially distinction conserving systems do exchange information with their environment (otherwise they would correspond to closed systems), but that this exchange is restricted (or 'filtered') by their boundaries. In this sense such a system can be said to be 'autonomous' : it does not respond directly to any signal from its environment. Its behavior is partially caused by the information coming from its environment, partially by the information inherent in its internal organization (endocausality). Such systems may also be called 'self-organizing', since the processes occurring within the boundary are not controlled or 'organized' by an external agent, but (at least partially) by the system itself.

However, I would like to reserve the word 'autonomy' for a more specific type of systems. Although a self-organizing system is incompletely controlled by its environment, this control may be sufficient to destroy the system, and hence to destroy its boundary. For example, most dissipative structures depend strongly on their boundary conditions for survival. E.g. in the Bénard phenomenon (Prigogine, 1979), the structure of hexagonal cells will disappear if the heating of the water layer is interrupted. I would call a system autonomous if it would possess some internal mechanism to counteract such destructive fluctuations of its boundary conditions, i.e. if it would somehow be able to control and hence to maintain its boundary. To maintain its boundary means to maintain the separation or the distinction between itself and its environment, hence to maintain its *identity* . (Such a system might also be said to *produce* its boundary and hence its identity or self, and might therefore be called *autopoietic* (i.e. self-producing), cfr. Maturana and Varela, 1980; Varela, 1979). Remark that in this case the boundary is not determined by the objective of the outside observer, but by the internal objective of the autonomous system : self-maintenance.

For a system to compensate the external perturbations and hence to secure its identity, it must be able to *adapt* to a changing situation, i.e. to change its relation with its environment so as to stabilize its boundary. I use the word 'adaptation' here as well in the sense of a system which changes its internal structure in order to survive in an invariant environment, as in the sense of an invariant system which changes its environment in order to cope with the strains it exerts. For example, I could adapt to an environment where there are a lot of wild animals by becoming stronger so that I could defend myself against any attack. But I might also change the environment by building protective walls or by poisoning the animals. Both are cases of adaptation, i.e. of changing the *relation* between myself and the environment in order to enhance my chances for survival. Adaptation is the steering by a system of its boundary conditions so as to secure the maintenance of this boundary. These conditions depend as well on the situation inside as on the situation outside the boundary. It is the relation between both which counts.

4. Cognition as problem-formulating and problem-solving ability.

In order to adapt a system must be able somehow to evaluate its present situation (i.e. its boundary conditions) in order to determine how likely it is that this situation will lead to the destruction of its boundary in some not to far away future. It can then try to change this situation to another situation which is less likely to lead to desintegration (normally it is impossible to find a situation which will never lead to the destruction of the system; we all know that we are going to die some day; in practice we shall try to postpone this day for as long as possible by leading a safe, healthy, satisfactory life).

These two situations as perceived by the system define a *problem* (for a general theory of problems and problem-solving see e.g. Newell and Simon, 1972) :

how to transform the actual situation A to a new situation B where the maintenance of the boundary is better secured ?

To solve this problem the system must carry out a search process starting from the initial state A and directed towards the goal or desired state B (see figure). (In the terminology of of Maturana and Varela (1980), A may be conceived as the state of the system which is 'perturbed' by the environment, B as the state where the perturbation has been 'compensated' by an adequate reaction of the system).

The efficiency of this problem-solving process will depend on the *knowledge* the system has about its relationship with the environment. If it would have no knowledge it could only try at random certain actions and see whether they would result in the desired state B. Except in the most simple cases, it is clear that such a blind exploration would have only the slightest chance ever to reach its goal. Therefore to be really autonomous a system must know how to adapt. It must be able to select the actions which are likely to lead to a more secure situation, and to eliminate the actions which are likely to lead to a less secure situation. If it is unable to carry out this selection internally, by itself, then it will be subjected to the selection by the environment (i.e. 'natural' selection), which will eliminate all systems which behave in an inappropriate way. (That is why knowledge can be conceived as a 'vicarious selector' (see Campbell,1974), which selects adequate behavior as if it were a 'vicar' or represen- tative of the environment.)

Hence an autonomous system should be able to distinguish adequate actions (likely to lead to the goal) from inadequate actions (unlikely to lead to the goal). This ability may be called heuristic or *problem-solving knowledge*. However, it should also be able to analyse, interpret and evaluate the different actual and potential situations in order to formulate an adequate goal (i.e. a goal which satisfies the basic need for self- maintenance and which can be reached within a reasonable time interval). This ability may be called *problem-formulating knowledge*. It presupposes that the system be able to distinguish those features of the environment which are

relevant for its general survival strategy, i.e. those features which directly determine whether the system will survive or not, together with the features which have a causal relationship with the former and which can be manipulated in some way by the system. For example, in order to survive I should be able to distinguish dangerous animals from harmless ones (problem-formulating knowledge), but I should also be able to distinguish effective means for eliminating the danger (e.g. poisoning the pool where the dangerous animals come to drink) from ineffective means (e.g. doing nothing at all) (problem-solving knowledge).

We may conclude that an autonomous system should be able to formulate and to solve adaptation problems, and that this activity is basically one of making the adequate distinctions. What is called cognition is just the process of making and manipulating those distinctions. The organized whole of distinctions and their (partially) causal relations may be called a representation of the adaptation problem, hence an *adaptive representation* (Heylighen, 1987a,b). Such a representation is an abstract, information-processing structure, which 'represents' in a certain sense the possible adaptive (i.e. securing the self-maintenance) interactions between the autonomous system and its environment.

This concept is a generalization of the concepts of 'knowledge representation' (Bobrow and Collins, 1975; Charniak and McDermott, 1986) and 'problem representation' (Newell and Simon, 1972; Korf, 1980) used in Artificial Intelligence and cognitive science. Whereas these more traditional representation concepts stress the correspondence (or mapping) between the elements of the representation and the elements (objects, properties) of the outside world, the adaptive representation concept emphasizes the correspondence between representation elements and possible adaptations, i.e. changes of the relation between the system (self) and the outside world. This means, first, that dynamic features are more important than static features, second, that the external phenomena are only represented insofar as they are related (directly or indirectly) to the system itself. In particular it implies that each adaptive representation must have an aspect of what might be called *self-representation*. It must be noticed, however, that self-representation can only be partial (cfr. the paper of Löfgren, this volume, and Heylighen, 1987a).

You may have noticed that I introduced a distinction as something which is used by the observer to describe the behavior of a system he wants to control from the outside, whereas I now use the concept of 'distinction' to describe how an autonomous system functions from the inside. The reason is that we as observers are autonomous systems ourselves, and hence are bound to make distinctions in order to solve our adaptation problems. The problem of how to control a particular system is just an instance of the general adaptation problem : how could we use this system as a means for effectuating the changes which would make our situation more likely to lead to self-maintenance ?

One of the basic characteristics of the new cybernetics is that there is no longer an absolute separation between the system which is to be controlled or observed and the system which is doing the controlling or observing (cfr. the paper by Glanville, in this volume). If we are designing a theory of how autonomous systems function as seen by an observer, then we are at the same time designing a theory of how the observer functions as an autonomous system, which may be observed by another autonomous system. One of the main advantages of the distinction concept is that it can be used as well to describe the 'object', which is distinguished, as the 'subject', who is distinguishing. The operation of 'distinguishing' is merely a partially distinction conserving process, leading from the object to the subject.

An implication of this view is that there is no such thing as absolute or objective knowledge. All knowledge consists of a finite collection of distinctions. On the other hand the

number of potentially distinguishable phenomena in the universe is infinite. The only criterion I have for selecting one particular finite set of distinctions (and hence rejecting the infinite set of all remaining distinctions) is that they appear adequate to solve my particular adaptation problems. Hence the cognitive distinctions I use are determined by the maintenance of my own boundary, i.e. the maintenance of my self-environment distinction, and not so much by the objective features of the world. In this sense knowledge is dependent upon autonomy.

The reason why knowledge appears more or less independent of such purely subjective criterion is that there are a number of intermediate levels between what we call 'rational knowledge' and the subjective desire for individual self-survival. What was called an adaptive representation is a very complex, multi-level organization of distinctions, connecting the primitive self-environment distinction to the more general, less self-centered distinctions associated with the rational manipulation of concepts. If we wish to understand how cognition, and hence autonomy, functions in general - from the most primitive organisms to the most advanced societies - we should unravel this organization and fill in the gaps which presently exist in our conception of these phenomena. The approach I want to advocate here is based on a *dynamics of distinctions* (Heylighen, 1987a,b), which would describe how new distinctions are created during the general (phylogenetic or ontogenetic) evolution of a system, so as to enhance its capacity for adaptation. To somewhat simplify this problem I propose first to distinguish different levels of complexity in the strategies used by a system for adaptation.

5. A hierarchy of distinction levels.

The question I wish to address now is : what kind of (cognitive) distinctions should a system make in order to secure the maintenance of its self-environment distinction ? To make this problem more concrete we will look at examples of systems which become gradually more complex, and analyse how the adaptation problems may be solved on each complexity level. We shall begin with an extremely simple autonomous system : a house kept at a constant temperature with the help of a thermostat.

The self-environment distinction the system attempts to maintain is that between an inside temperature which is higher than some fixed value (e.g. 21° C) and a fluctuating outside temperature. The only instrument the system has at its disposal is a heating apparatus with two states : off and on. What must the system be able to do in order to be autonomous ? First it should be able to perceive and to interpret those features of its situation which are relevant with respect to its boundary maintenance (problem-formulating knowledge). The only feature which should be distinguished in this respect is between a temperature lower than 21° C and a temperature higher than or equal to 21° C. Second the system should be able to distinguish between adequate and inadequate actions to perform in order to restore the desired situation (problem-solving knowledge).

Since there are only two possible actions (turning the heating on, respectively off), the decision as to which action is adequate for the given situation is very simple : if the detected temperature is lower than 21° C the heating should be turned on, otherwise it should be turned off. This is the only 'knowledge' the system needs about the way it can adapt to environmental

changes. It can be represented as an elementary causal (i.e. distinction conserving relation) between the distinction made by 'perception' and the distinction leading to 'action' :

(temp. $<$ 21° C) $\rightarrow$ (heating on)
(temp. $\geq$ 21° C) $\rightarrow$ (heating off)

In this case there is just one distinction (i.e. two alternatives) perceived by the system and there is just one type of action or change which can be effectuated by the system in order to make the transition from the alternative evaluated as negative for self-maintenance to the alternative evaluated positively. However, for more complex systems the self-maintenance will require a larger set of actions which respond to a more differentiated set of external situations. This means that the system will now have to choose between more than two possible actions to be performed in a given situation. Therefore it should make a more elaborated classification of distinct situations, i.e it should make more perceptual distinctions.

An elementary example of such a system is a living cell. The processes inside the cell are controlled by the DNA, which functions as a coupled set of genes. Each gene can be active (i.e. producing a particular type of protein) or inactive : it has two distinct states. The genes constitute a chemical network, in the sense that the activity of one gene can activate or desactivate another gene by producing certain enzymes. The activity of the network as a whole will depend on the chemical 'situation' of the cell, which depends on the environment. The 'knowledge' inherent in the DNA is that which distinguishes a particular class of situations by activating a specific set of genes responding to that class. If the DNA is well-adapted to its environment, this particular activation pattern should be sufficient to compensate the perturbation associated with that class.

However, it could be that the situation the system is confronted with does not fit in the 'preprogrammed' or 'wired-in' classes of situations for which there is a known, adequate response. In that case the system can only try a certain behavior and hope that it will prove adequate. If it does not, the resulting new perturbed situation may still be recognized so that the overall behavior can be corrected by feedback. This is an elementary form of problem-solving : different paths of actions are tried until a more or less satisfying situation is reached.

On a higher level of autonomy or intelligence the system will *learn* from such mistakes : the knowledge that a particular action is adequate or inadequate in a particular situation will somehow be stored in the network of interconnected distinctions. This means that it will be possible for the system to create new distinctions and new associations between distinctions. An example of such systems are animals with a central nervous system. This can again be conceived as a network of interconnected distinctions : in a simplified model each cell (neuron) of the nervous system can be in two states, active or inactive, and this activity is passed from one neuron to the other ones through their connecting synapses. The learning occurs through a change of the pattern of interconnections.

On a yet higher level, the number of distinct states of the representation through which the system must search for a goal state may become so large that the learned set of associations becomes insufficient to guide the problem-solving process efficiently. In that case the system should be able to select a small subset of distinctions and associations, which are more invariant or reliable than the other ones and which would provide an easily manipulatable subrepresentation because they could be handled more or less independently of each other. These stable distinctions may be called 'concepts' and their associations (production or deduction) 'rules'. The combination of such concepts according to the rules will allow the system to explore classes of situations

which were neither wired-in nor learned, but which correspond to logically conceivable 'possibilities'. Such systems may be called 'rational' or 'logical-conceptual'. An example of a rational system is a normal human being.

We could even go further and imagine a level of autonomy where also the rational mechanisms of cognition would be transcended. Such a system might for example be characterized by its capacity for methodically adapting its own concepts and rules (i.e. changing its logic) according to the situation it is confronted with. This would require yet another level of distinctions which would represent and hence allow to control the level of concepts and rules below.

6. Towards a dynamics of distinctions.

This global construction may be conceived as a hierarchy (see Mesarovic, Macko and Takahara, 1970) of representation or distinction levels (in Campbell's (1974) terminology : a nested hierarchy of vicarious selectors ; in Nauta's (1972) terminology : a hierarchy of semiotic or informational levels).

The fundamental level is determined by one distinction : that between survival and destruction of the system, which corresponds to the distinction between system and environment (i.e. the 'boundary'). Above this level comes a first level of 'vicarious' distinctions, which classify situations and choose between actions according to the likeliness that they will lead to survival or to destruction. In biological systems this level is realized by the DNA, and its adequacy is guaranteed by a phylogenetical variation-and-selection process. On a following level we may find a set of learned distinctions which effectuate a more fine-grained classification and selection of situations and actions. This distinction level results from an ontogenetical history of trial-and-error experiences. A yet higher, rational level of distinctions may be added through the exposition of an individual to the concepts and rules provided by language and culture. These were arrived at through a process of discovery, based on trial-and-error, by the culture as a whole.

On each level the creation of new distinctions appears to be determined by a process of blind-variation-and-selective-retention (i.e. trial-and-error), whereby new distinctions are tried in a blind way, after which those which prove to be adequate according to the criteria determined by the already existing distinctions are selected and retained. The distinctions which prove to be inadequate (i.e. the classes they select do not fit to the already existing classes, which themselves ultimately depend on the survival-destruction or self-environment distinction) are simply eliminated. The general evolutionary tendency leads to an ever larger pyramid of distinctions with the survival-destruction distinction at the top.

This process can also be conceptualized as a generalization of the problem-solving method of *means-ends analysis* (Newell and Simon, 1979). This is a technique which consists in factorizing the change which is to be brought about (from the initial state A to the final state B), into a set of smaller changes for which adequate actions are known. To do this the 'difference' (i.e. distinction) between A and B is represented as a conjunction of 'smaller' differences. To solve the problem it now suffices to find actions (or action sequences) which would reduce each of these 'smaller' differences. For autonomous systems the basic difference to be factorized is that between

survival and destruction. The smaller differences correspond to the cognitive distinctions the system must make in order to solve its survival problem in an efficient way.

What should be done now is to construct a formal description of this complex set of processes and structures which determines the functioning and creation of distinctions, allowing an autonomous system to adapt to a changing environment. This description would be an adaptive representation of adaptive representations, or in other words : an *adaptive metarepresentation.* It would allow to analyse adaptive representations, to classify them, to evaluate them according to their adequacy, to transform them so as to make them more adequate. Such a theory would provide the real basis for a general theory of autonomy and cognition, and hence for a new cybernetics.

A first sketch of a formal-conceptual framework in which an adaptive metarepresentation could be formulated is proposed in (Heylighen, 1987a,b). It is based on an analysis of existing classical and non-classical scientific representations and on a (partial) reconstruction of their fundamental substructures (objects, predicates, proposition logic, state space, operators, time, dynamical constraints) by means of a dynamic algebra of distinctions. Although the framework is general enough to allow the description of non-distinction conserving processes, it still lacks a clear conception of why certain distinctions are created while other distinctions are eliminated. The present analysis of autonomy and adaptation may well be a first step towards the resolution of this problem. The next step should be a more thorough, formal analysis of how higher order distinctions are formed by blind variation and selective retention determined by distinctions which are more basic in the hierarchy.

7. Conclusion

I have argued that a new cybernetics is to be developed which would be concerned with the modelling of systems characterized by autonomy and hence by cognition. Such systems were contrasted with the mechanical systems which inspired classical cybernetics by introducing the concept of distinction conservation. In a classical, causal, mechanistic system all distinctions made in the input can be found back in the output. From this point of view, the system is just a channel waiting passively for input signals and transforming them to the corresponding output signals. The form of the signal has changed but its information content has remained. This is not longer true when there is an interference with noise, causing a loss of information and hence of distinctions. The methods of classical cybernetics are designed to minimize the effect of such noise, and hence to bring the system back to its causal mode of operation.

An autonomous system, however, is characterized by the fact that it is *intrinsically* non-distinction-conserving. The change of distinctions (creation and destruction) during the process is not due to some external source of noise, but to the functioning of the system itself. The system *must* create and annihilate distinctions in order to maintain its autonomy. Indeed, autonomy can be defined as the ability to maintain a self-environment boundary. This stable boundary is produced by counteracting or compensating external influences which may be destructive, i.e. by evading the control by the environment.

To succeed in this task the system must solve adaptation problems. This requires the application of cognitive distinctions in order to reduce (i.e. represent, factorize and solve) the problem. However, the distinctions made by the system are different from the distinctions made by the observer. They are determined by the goal of self-survival of the system, whereas the distinctions made by the observer are determined by the objective the observer has in mind while trying to control the system. From the viewpoint of the observer this means that the system does not conserve distinctions. If the observer nevertheless attempts to control the system by forcing it to conserve the distinctions he makes, he not only denies the autonomy of the system, he even may destroy it (cfr. the 'allopoietic' use of 'autopoietic' systems, Maturana and Varela, 1980).

If we wish to model the functioning of such a system, we must construct a representation of the way it makes distinctions and hence represents its adaptation problems. In other words, we need an adaptive metarepresentation. To build such a metarepresentation, it was proposed first to arrange different types of distinctions in a hierarchy, with the survival-destruction distinction at the bottom and the rational- conceptual distinctions at the top, second to design a dynamics of distinctions, based on a variation-selection process represented within a dynamic distinction algebra. This should lead to a framework which is broad enough for the foundations of a new cybernetics.

References :

Bobrow D.G. and Collins A. (eds.) (1975) : *Representation and Understanding. Studies in Cognitive Science*, (Academic Press, New York).

Campbell D.T. (1974) : *Evolutionary Epistemology* , in : The Philosophy of Karl Popper, Schilpp P.A. (ed.), (Open Court Publishing, La Salle, Illinois), p. 413.

Charniak E. and McDermott D. (1985) : *Introduction to Artificial Intelligence*, (Addison-Wesley, Reading, Mass.).

Haken H. (1978) : *Synergetics. Non-Equilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*, (Springer, Berlin).

Heylighen F. (1986) : *Towards a General Framework for Modelling Representation Changes*, in : Proceedings of the 11th International Congress on Cybernetics, (Association Internationale de Cybernétique, Namur, Belgium).

Heylighen F. (1987a) : *Representation and Change. An Integrative Meta- representational Framework for the Foundations of Physical and Cognitive Science*, (Ph. D. thesis, Vrije Universiteit Brussel).

Heylighen F. (1987b) : *Formal Foundations for an Adaptive Metarepresentation* , in : Proceedings of the 7th International Congress of Cybernetics and Systems, (Thales publications, St. Annes-on-Sea, Lancashire).

Korf R.E. (1980) : *Toward a Model of Representation Changes*, Artificial Intelligence 14, p. 41.

Maturana H. and Varela F. (1980) : *Autopoiesis and Cognition : the realization of the living*, (Reidel, Dordrecht).

Mesarovic M.D., Macko D. and Takahara Y. (1970) : *Theory of Hierarchical , Multilevel Systems*, (Academic Press, New York).

Mesarovic M.D. and Takahara Y. (1975) : *General Systems Theory : Mathematical Foundations*, (Academic Press, New York).

Nauta D. (1972) : *The Meaning of Information*, (Mouton, Den Haag).

Newell A. and Simon H.A. (1972) : *Human Problem Solving* , (Prentice Hall, Englewood Cliffs).

Prigogine I. (1979) : *From Being to Becoming : Time and Complexity in the Natural Sciences*, (Freeman, San Francisco).

Spencer Brown G. (1969) : *Laws of Form*, (Allen & Unwin, London).

Varela F.J. (1979) : *Principles of Biological Autonomy*, (North Holland, New York).