Psyke & Logos
nr. 2, 1982

SELF-REFERENCE AS A PROBLEM IN THE CONTROL OF COMPLEX SYSTEMS

Erik Hollnagel & Morten Lind

1. Introduction

The purpose of this paper is to discuss the relation between self-reference and some aspects of the control of complex systems. This may at first sight appear to be rather esoteric, but we strongly believe that it is a relation where philosophy and psychology may conceivably make a valuable contribution to systems control and the design of man-machine systems (MMS). The present stage of technological development has produced systems so complex that they defy traditional approaches to control. Thus neither Control Theory nor Human Factors can provide any viable solutions to the problems arising from this complexity. New alternatives appear possible from a combination between Systems Theory, Cybernetics, or Artificial Intelligence and the more traditional disciplines of Cognitive Psychology, Human Factors Engineering or Control Theory. In these efforts we face problems that are relatively new to the technological disciplines, but which are familiar elsewhere, such as in Philosophy of Science, Epistemology and Phenomenology. One such problem is self-reference or self-reflectivity. Since the relevant sciences in the technical field, such as Cybernetics, lack a proper philosophical foundation, we see this as an apportunity for philosophy to demonstrate the practicality of its conceptual framework, and further to make a valuable contribution to solve a practical problem. Technology should not be anathema to philosophy, nor should philosophy be banned from technology (cf. Ihde, 1979).

2. Definition of the Control Problem

As a basis for the following discussion we will give a definition of the control problem in very general terms. We will describe the control problem independent of whether the decision-maker is a human operator, a machine (computer), or a MMS where man and machine share the decision-making responsibility. The latter situation gives rise to some interesting philosophical problems that will be discussed later.

In essence control theory deals with a decision-making situation that involves the interaction between three systems as outlined in figure 1. The sy-

stem 'A' represents the process as an activity which must be controlled. It has some inputs 'x' which can be manipulated to modify its behaviour and some outputs 'y' which provides information about the actual state or internal conditions of the system. Furthermore, system 'A' has another set of inputs 'z' which represents a source of disturbances that contribute to the uncertainty of the behaviour of 'A'. The disturbing input 'z' is the output of another system 'B' which includes all phenomena in the environment of 'A' (except for the system 'C', cf. below) which may have an influence on its behaviour. The third system 'C' in Figure 1 is a control system that interacts with 'A', through the input 'y' and output 'x', to produce an overall behaviour of the total systems which satisfies a specified goal. As an example, 'C' might manipulate 'x' so as to maintain 'y' constant or to keep it within specified limits in spite of the disturbances 'z'. This is an abstract desciption of the principles behind e.g. thermostats for the regulation of room temperature or homeostatic mechanisms in biological systems. The structure in Figure 1 can be used as a model for natural and artificial phenomena alike, but in every case it is absolutely essential that the goal of 'C' can be identified i.e. explicated. Otherwise it is not possible to distinguish between 'A' and 'C'. The whole point in applying the control model of Figure 1 then disappears since 'A' and 'C' might equally well have been represented by one system. Another critical point in control theory is the identification of 'x', 'y' and 'z' or in other words the identification of 'A', 'B' and 'C' as separate systems.

In the design of control systems the situation is somewhat different since the whole purpose of the design activity is to synthesize a system 'C' which will make the total system ('ABC') behave according the the designer's intentions. Here the structure in Figure 1 is used as guidance for the design and a major part of the work is modelling the controlled system 'A' and its possible disturbances (the system 'B'). These models provide the basis for the choice of 'C', in particular its decision-making strategies. It has actually been proved that a good controller should have a model of its environment (Francis & Wonham, 1976). However, even if a proof had not been given it is quite obvious that it is necessary for a controller, be it a human operator or a machine, to have acces to information about the properties of the systems that must be controlled, i.e. have a model of its environment. Without a model, the controller would not be able to make predictions and plan its actions and could accordingly not produce purposeful behaviour.

2.1. Control of Large Systems

Modern control theory provides a wide range of highly developed mathematical techniques for control synthesis but the coverage of these techniques is not sufficient to deal with the complex decision-making problems in the control of large industrial systems such as nuclear power plants or che-

mical production units. These systems are large in terms of the number of components involved in their operation. In a typical power plant the number of valves, pumps, tanks, etc. is in the order of several hundred, and similar to other large systems (social, for instance) they cannot be described with one single model. These systems are called multidimensional because the possible modes of interaction between components are enormeous, and require several distinctly different models to capture their nature.

Historically, industrial processes have always been controlled or supervised by human operators. But as automation technology has advanced they have become increasingly automated. The structure of these systems is depicted in Figure 2. The control system comprise the operator, the automated controls and a man-machine interface which supports the communication between the operator, the production process and the automation. Due to the complexity of the production process they cannot be completely automated. It is simply inpossible to predict all possible modes of malfunction which can occur in these systems. In the highly automated processes we have today the operator's role is to supervise the automated control systems i.e. to ensure that they function as intended and to take care of the remaining control tasks which cannot or have not yet been automated.

The deficiencies of this approach to automation become apparent in the case of infrequent but serious plant disturbances which have not been anticipated by the designer of the automated systems. In these situations it may be necessary for the operator to take over the function of some of the automated controls, a task for which he may not be well prepared. This is the background for characterizing the operator's situation as 99 percent boredom and 1 percent terror (Bibby et al., 1975). It is clearly an undesirable working situation for an operator. Neither is it desirable from a production and economic point of view, because the infrequent but serious disturbances usually involve risks for loss of equipment or for the plant environment. This can be in terms of a sudden release of energy (explosions) or poisenous materials (e.g. radioactivity or dioxin, as in Soveso). Accordingly, there is no reason to accept this situation and one of the remedies is to change the nature of the man-machine interface.

3. The Problem of MMS: Coping With Complexity

Proper design of the man-machine interface is essential to the operator if he shall be able to respond properly to plant malfunctions. This does not just involve the traditional ergonomic concerns of the physical properties of the operators work place such as the layout of indicators and instruments, the control panel, or the size of colour of knobs and dials etc. The cognitive aspects of the operator's work situation must also be considered. Problems related to how the interface supports the operator in identifying the plant state, hence the control problem that must be solved, are of the outmost

importance – especially in the infrequent and therefore less trained disturbance situations which require thinking and problem solving. A major deficiency of existing man-machine interface is that it does not support cognitive activities. The operator is expected to be able to diagnose process malfunctions on the basis of thousands of individual alarms. He is left completely on his own with the problem of correlating complex alarm patterns with the plant knowledge acquired during training or daily routine.

There is presently an effort to provide a systematic basis for the redesign of the man-machine interface by taking advantage of the information processing capacity of computers. The basic idea is to use a computer to provide information to the operator related to different ways the functional properties of the plant can be represented (Rasmussen & Lind, 1981). In this approach the plant is considered as a multidimensional system which must be described from different perspectives in order to capture its functional nature. This can be demonstrated by an example chosen form the everyday life. Suppose you should describe the function of a mechanical watch. You may choose to describe its operation in terms of the movement of the parts which make up the watch. But this description would only capture a fraction of the information you need to describe the nature of a watch. To complement this simpel, mechanistic view of a watch you most also describe how the movements of the parts are coordinated to constitute the functions of the watch such as the wheel train and the escapement. Furthermore, you must also describe the purpose of this mechanisms as a timekeeping device. These descriptions of the watch deal with different aspects of the same physical phenomena. In daily use only the description of the watch as a timekeeping device is necessary, but if something fails you must consider the other aspects of the system in order to diagnose the fault or compensate for the effects of wear, etc. You can imagine how difficult it would be to diagnose multiple faults if you did not considered the watch from different perspectives and, for instance, only regarded it as a complex of interacting parts.

If we return to the problems of diagnosis in large industrial processes conventional man-machine interfaces will only supply the operator with information about the individual parts. Information about higher level functions is not directly presented to the operator. He has to produce this information on the basis of complex inferences from working experience and general knowledge about the plant. This is often an impossible task especially during emergencies. The goal of current research in MMSs for the control of large systems is to support the operator with state information related to different perspectives as explained in our simple watch example. The different models applied in this constitute a so-called abstraction hierarchy. By means of that it is possible to organize plant information such that a computer may help the operator to cope with the complexity of operating and supervising large processing units.

4. Operator Models and Machine Images

One of the attempts of solving the MMS problem concerned with the control of complex systems is a synthesis called Cognitive Systems Engineering (Hollnagel & Woods, 1982). The basic principle is that the parts of a MMS should be considered as cognitive systems. A cognitive system is an adaptive system which bases its actions, and particularly planning and modification of actions, on knowledge about itself andt its environment, cf. the discussion of the control problem above. Thus neither man nor machine should be regarded as simply reacting to the information received, but rather as acting on basis of that.

The knowledge about the environment is commonly referred to as a model for the environment. To make a distinction between a person and a machine, we shall use the terms 'model' and 'image', and refer to the person's model of the environment and the machine's image of the environment. In particular the person will have a model of the machine he is working with, while the machine will have an image of the operator. In most cases the machine has been equipped with an image of the operator as a part of its design, but in principle nothing prevents the machine from developing this image itself. In the discussion of this image it is useful to make a distinction between several levels.

4.1. Images on the First Level

All machines are artificial systems because they are designed with a specific purpose in mind. Any tool, for instance, is an example of that. (As the examples chosen will indicate we shall generally use the term 'machine' in a very broad sense, which goes beyond simple mechanistic connotations). From the beginning machines were made by the individual user for himself, but gradually machines were made by one person for the use of another. Thus every machine was designed with another person in mind, although this was not really conspicuous before the industrial revolution. In case a machine, is made for a particular person, e.g. a house or a custom built bicycle, the image included in it is fairly easy to discern. But most machines are made for a group of users, and the image reflects the characteristics of the group rather than of any particular person.

A guitar, for instance, is made on the assumption that the operator is right-handed, that he has five fingers on each hand, that he possesses a certain muscular strength, etc. The image on this level thus mainly implies physical relations. The same goes for a hammer, a car, a shirt, a TV, a stove, etc. The machine on which this is written, and the journal in which this is printed are two more instances, and there is obviously no end to the examples that could be mentioned.

4.2. Images on the Second Level

The second level is characterized by adding assumptions about functional relations. A traffic light, for instance, assumes that the person is able to discriminate the colors and choose the proper activity based on that, i.e. to follow a rule. On this level we generally have machines that emit information, and in doing so apply a specific image of the user or group of users. We are here dealing with moderately complex technological systems, and it should be fairly obvious that some consideration of the user's characteristics have been made during the design. Human Factors is a typical example of a discipline that operates on this level.

4.3. Images on the Third Level

For most of the machines which we use in daily life it is sufficient to remain with the image on the second level. But the technological development has brought about systems that require yet another level of images. On this, the third level, the machine must make assumptions about the person as a cognitive system. Conversely that implies that it must also in some way realize that it is itself a cognitive system on a par with the person.

It is easy to find an example of this in the human domain. Every person communicates with others on the basis of a model of them as cognitive systems, i.e. as being of the same kind as himself. And this is the basis for the complexity and efficiency of human communication (despite isolated examples to the contrary).

If we return to the realm of machines, this third level of the image may seem rather remote. And there are probaly few systems today which exhibit this characteristic. The important point, however, is that it is necessary for the control of very complex processes, to make it possible for man to cope with the complexity, and to create the basis for a humanized work environment for the operator.

The failure to realize this means that the image on the second level is extended to situations where it is inadequate. This is in fact the kernel of the physicalistic approach that is exhibited e.g. by behaviorism and by the mechanistic philosophies that apparently dominate the current philosophy of psychology (cf. the Cognitive Science movement, e.g. Simon, 1980). On the second level the image of man implies he is a machine. (Thus on all levels is imbedded in the image the assumption that man is of the same kind as the machine). The effort is accordingly to bring the machine to function within the capacity limits of man, i.e. his machine-like characteristics in terms of perception and motorics.

This approach, however, is doomed to fail. If it could succeed it would mean that all parts of the machine's functions could be automated, since man is regarded as no more as a complex automation. Yet the very need for

the presence of man demonstrates the futility of the approach. Man is necessary because there are essential parts of the machine's functions that cannot be automated. (And this again is because the complexity of the machine defies the language we have for describing it). To describe man by mechanistic principles, as implied by the image on the second level, is thus a blatant contradiction of the fact that man is needed at all. Hence the necessity of going to the third level of the image, where the machine considers man, and conversely itself, as a cognitive system.

5. Self-Reference and Self-Reflectivity

Referring to the definition of a cognitive system given above, it is fairly obvious that using knowledge about oneself implies some kind of self-reflection. But now we have also seen that using knowledge about the environment, in particular about that part of the system which is the 'other', leads to self-reference on the third level of the image. Thus having demonstrated the need for considering self-reference in dealing with the design of control systems, we may now turn to the very problem of self-reference and self-reflectivity.

5.1. Self-Reference

So far we have used the terms self-reference and self-reflectivity rather indiscriminately. But we intend to show that one may assign a precise meaning to each of them, and that this has implications for their use in the design of control systems.

Self-reference may be defined as follows: A system is self-referential when it uses a model of itself as a basis for communication/interaction with other systems.

The details of this have been described in Hollnagel (1978). Put very simply, the system considers itself as a SELF with explicit relations to other systems. It makes the essential distinction between itself and the other and uses knowledge, however rudimentary, about itself vis-a-vis others to structure the communication. Two major aspects of this are the formulation and interpretation of messages. It thus has the essential communicative ability to know when it is referred to in the communication, i.e. to recognize references to itself in the information it receives. This is a fairly basic quality that may be found in men as well as machines.

5.2. Outer-Oriented Self-Reflectivity

We shall make a distinction between two types of self-reflectivity, called outer-oriented and inner-oriented. A system exhibits outer-oriented self-reflectivity:

(1) if it uses knowledge about itself (in the form of an image or a model)
(2) derived from considerations of the reactions from other systems (i.e. feedback)
(3) as a basis for adjusting its pattern of activity.

Or, to put it very simply, a system with outer-oriented self-reflectivity has an adaptive form of self-reference.

The basic for this adaptive form of self-reference is that the system considers the information or messages that it gets. It is thus not sufficient for a system simply to be adaptive and to have a repertoire of responses, however complex. The crucial point is that the activities are not defined in advance, i.e. that the complete input-output relations cannot be prescribed. A thermostat is adaptive, but it does not exhibit outer-oriented self-reflectivity because it has no self-reference. It is unaware of its own existence, even in the most primitive terms. It responds to the feedback, but does not consider it. It treats the feedback according to a predefined set of rules, but is unable to modify these rules. It is this second level of adaptability that is necessary for a system to be characterized as having outer-oriented self-reflectivity.

5.3. Inner-Oriented Self-Reflectivity

Whereas the definition of outer-oriented self-reflectivity was fairly complex, the definition of inner-oriented selfreflectivity is very simple. A system exhibits inner-oriented self-reflectivity if it considers the outer-oriented self-reflectivity, i.e. if it considers how it adapts. The inner-oriented self-reflectivity is thus a thinking about the patterns of response that the system realizes it has. We are thus talking about self-reflectivity on a second level, i.e. self-reflectivity of self-reflectivity. This suggest the possibility of recursion, which is always an interesting but nasty aspect of a system, cf Hofstadter's (1979) tour-de-force on tangled hierarchies and strange loops. we shall, however, refrain from getting mixed up with that on this occassion.

5.4 Control Systems and Self-Reflectivity

We have argued, hopefully convincingly, that control systems need to have self-reference. They furthermore need to be adaptive, hence have self-reflectivity. And we have now seen that the kind of self-reflectivity we talk about is outer-oriented self-reflectivity, but we need not be concerned with this possibility here. Since we cannot expect machines to miraculously develop self-reflectivity, whether it be of one kind or another, it is we who must supply the machines with self-reflectivity. This is not the philosopher's at-

tempt to construct a Calculus Ratiocinator or a revival of Rabbi Loew's attempt to create a Golem. We do not want to make any implications of this endeavour for philosophy and epistemology. It is simply a practical necessity. And it is obvious that if we shall ever design a machine with self-reflectivity, we must know a lot more about the self-reflectivity that we have as human beings. This is why we have presented the problem here.

It would, of course, be nice if we could get a fixed solution to the problem. But judging from the fumblings of Articifial Intelligence and the selected philosophical discussions that penetrate to the technological world, there is little, if any, hope of that. What we rather hope to acomplish is to make other people aware of the problem. It is possible to enter into a long-winded argumentation about the apparent inevitability of an increasing technological society, hence the need for taking these problems seriously. One could also begin to cite statistics on the increase in the number of industrial robots, the proliferation of micro-computers, or the rise in reported accidents in industrial installations (nuclear plants, computer systems, chemical factories, etc.). But for anyone who takes the society in which he lives seriously, there is no need for this type of argumentation. Let us simply acknowledge that the problem is there, that it is growing more and more serious, and that it is up to us to pool our resources and try to find a solution to it. It is thus not only a philosophical challenge but a technological, hence scientific necessity.

6. About the Authors

Erik Hollnagel is a psychologist with a past as a computer programmer. He has a Ph. D. (lic. Psych.) from the University of Aarhus where he was an associate professor for several years. From 1978 to 1982 he worked as a research fellow at the Risø National Laboratories doing research in MMS. He has recently moved to The OECD Halden Reactor Project, Norway, to take part in the establishing of a MMS research laboratory. His current interests include Cognitive Systems Engineering, Artificial Intelligence and psychological theories of human action.

Morten Lind, from the Risø National Laboratory in Denmark, is a systems scientist working with man-machine problems in the control of large complex systems. He has a Ph. D. (lic. Tech.) from the Technical University of Denmark, and his current interests include systems modeling and the analysis and synthesis of adaptive and self-organzing systems which include both men and machines.

REFERENCES

BIBBY, K. S. et al. (1975). Man's role in control systems. Boston: *Proceedings of IFAC Congress.*

FRANCIS, B. A. & WONHAM, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5), 457-465.

HOFSTADTER, D. E. (1979). *Goedel, Escher, Bach: An eternal golden braid.* New York: Basic Books.

HOLLNAGEL, E. (1978). *Qualitative aspects of man-machine communication* (RISØ-M-2114). RISØ National Laboratory, Roskilde, Denmark: Electronics Department, (NKA/KRU-P2(78)5).

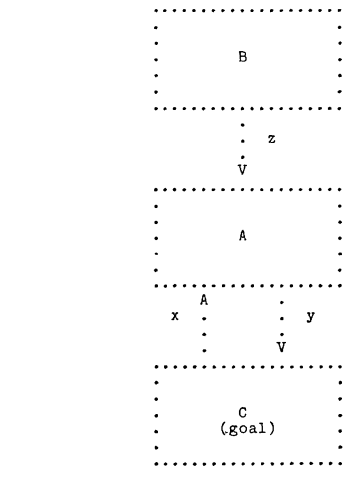
HOLLNAGEL, E. & WOODS, D. D. (1981). *Cognitive systems engineering* (RISØ-M-2330). RISØ National Laboratory, Roskilde, Denmark: Electronics Department.

IHDE, D. (1979). *Technics and praxis.* Dordrecht, Holland: D. Reidel Publishing Company.

RASMUSSEN, J. & LIND, M. (1981). *Coping with complexity* (RISØ-M-2293). RISØ National Laboratory, Roskilde, Denmark: Electronics Department.

SIMON, H. A. (1980). Cognitive science: The newest science of the arfiticial. *Cognitive Science*, 4, 33-46.

Figure 1:
The control theory concerns the analysis and synthesis (or design) of systems or mechanisms that show a goal-directed activity. It makes use of models of A, B, and C.



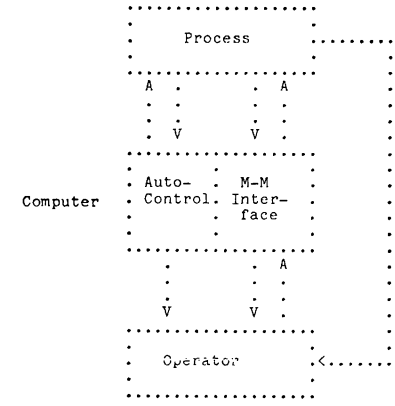


Figure 2:
Man-Machine Systems in Process Control.