Int. J. Systems Sci., 1970, vol. 1, No. 2, 89-97

EVERY GOOD REGULATOR OF A SYSTEM MUST BE A MODEL OF THAT SYSTEM[1]

Roger C. Conant

Department of Information Engineering, University of Illinois, Box 4348, Chicago, Illinois, 60680, U.S.A.

and W. Ross Ashby

Biological Computers Laboratory, University of Illinois, Urbana, Illinois 61801, U.S.A.[2]

[Received 3 June 1970]

The design of a complex regulator often includes the making of a model of the system to be regulated. The making of such a model has hitherto been regarded as optional, as merely one of many possible ways.

m this paper a theorem is presented which shows, under very broad conditions, that any regulator that is maximally both successful and simple *must* be isomorphic with the system being regulated. (The exact assumptions are given.) Making a model is thus necessary.

The theorem has the interesting corollary that the living brain, so far as it is to be successful and efficient as a regulator for survival, *must* proceed, in learning, by the formation of a model (or models) of its environment.

1. INTRODUCTION

Today, as a step towards the control of complex dynamic systems, models are being used ubiquitously. Being modelled, for instance, are the air traffic flow around New york, the endocrine balances of the pregnant sheep, and the flows of money among the banking centres.

So far, these models have been made mostly with the i(lea that the model might help, but the possibility remained that the cybernetician (or the sponsor) might think that

1 Communicated by Dr. W. Ross Ashby. This work was in part supported by the Air Force office of scientific Research under Grant AF-oSR 7o-1865.

2 Now at University College, P.o. Box 78, Cardiff CF1 1XL, Wales.

some other way was better, and that making a model (whether digital, analogue, mathematical, or other) was a waste of time. Recent work (Conant, 1969), however, has suggested that the relation between regulation and modelling might be much closer, that modelling might in fact be a *necessary* part of regulation. In this article we address ourselves to this question.

The answer is likely to be of interest in several ways. First, there is the would-be designer of a regulator (of traffic round an airport say) who is building, as a first stage, a model of the flows and other events around the airport. If making a model is *necessary*, he may proceed relieved of the nagging fear that at any moment his work will be judged useless. Similarly, before any design is started, the question: How shall we start? may be answered by: A model *will* be needed; let's build one.

Quite another way in which the answer would be of interest is in the brain and its relation to behaviour. The suggestion has been made many times that perhaps the brain operates by building a model (or models) of its environment; but the suggestion has (so far as we know) been offered only as a possibility. A proof that model-making is necessary would give neurophysiology a theoretical basis, and would predict modes of brain operation that the experimenter could seek. The proof would tell us what the brain, as a complex regulator for its owner's survival, *must* do. We could have the basis for a theoretical neurology.

The title will already have told this paper's conclusion, but to it some qualifications are essential. To make these clear, and to avoid vagueness and ambiguities (only too ready to occur in a paper with our range of subject) we propose to consider exactly what is required for the proof, and just how the general ideas of regulation, model, and system are to be made both rigorous and objective.

2. REGULATION

Several approaches are possible. Perhaps the most, general is that given by Sommerhoff (195o)) who specifics five variables (each a vector or n-tuple perhaps) that must be identified by the part they play in the whole process.

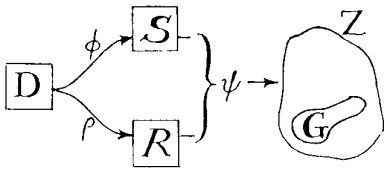


Figure 1

(1) There is the total set Z of events that may occur, the regulated and the unregulated; e.g. all the possible events at an airport, good and bad. (Set Z in Ashby's (1967) reformulation in terms of set theory.)

(2) The set G, a sub-set of Z, consisting of the 'good' events, those ensured by effective regulation.

(3) The set R of events in the regulator *H*; (e.g. in the control tower). [We have found clarity helped by distinguishing the regulator as an object from the set of events, the values of the variables that compose the regulator. Here we use italic and Roman capitals respectively.]

(4) The set S of events in the rest of the system s (e.g. positions of aircraft, amounts of fuel left in their tanks) [with italic and Roman capitals similarly].

(5) The set D of primary disturbers (Sommerhof's 'coenetic variable); those that, by causing the events in the system S, tend to drive the outcomes out of G: (e.g. snow, varying demands, mechanical emergencies).

(Figure 1 may help to clarify the relations, but the arrows are to be understood for the moment as merely suggestive.) A typical act of regulation would be given by a hunter firing at a pheasant that flies past. D would consist of a1l those factors that introduce disturbance by the bird's coming sometimes at one angle, sometimes another; by the hunter being, at the moment, in various postures; by the local wind blowing in various directions; by the lighting being from various directions. S consists of all those variables concerned in the dynamics of bird and gun other than those in the hunter's brain. *H* would be those variables in his brain. G would be the set of events in which shot does hit bird. R is now a 'good regulator' (is achieving 'regulation') if and only if, for all values of D, R is so related to s that their interaction gives an event in G.

This formulation has withstood 2o years' scrutiny and undoubtedly covers the great majority of cases of accepted regulation. That it is also rigorous may be shown (Ashby, 1967) by the fact that if we represent the three mappings by which each value (Figure 1) evokes the next:

$$\phi : D \to S$$

$$\rho : D \to R$$

$$\psi : S \times R \to Z$$

then 'R is a good regulator (for goal G, given D, etc., ϕ and ψ)' is equivalent to

$$\rho \subset \left[\psi^{-1}(G)\right]\phi,$$

to which we must add the obvious condition that

$$\rho\rho^{-1} \subset 1 \subset \rho^{-1}\rho$$

to ensure that ρ is an actual mapping, and not, say, the empty set! (We represent composition by adjacency, by a dot, or by parentheses according to which best gives the meaning.)

It should be noticed that in this formulation there is no restriction to linearity, to continuity, or even to the existence of a metric for the sets, though these are in no way excluded. The variables, too, may be partly functions of earlier real time; so the

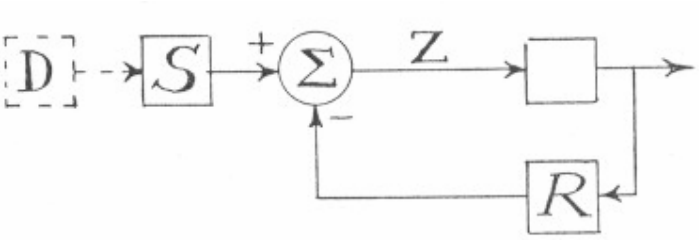
formulation is equally valid for regulations that involve 'memory', provided the sets D, etc., are defined suitably,

Any concept of 'regulation' must include such entities as the regulator R, the regulated system S, and the set of possible outcomes Z. Sometimes, however, the criterion of success is not whether the outcome, after each interaction of S and R, is within a goal-set G, but is whether the outcomes, on some numerical scale, have a root-mean-square sufficiently small.

A third criterion for success is to consider whether the entropy $H(Z)$ is sufficiently small. When Z can be measured on an additive scale they tend to be similar: complete the constancy of outcome $\Leftrightarrow H(Z)$ = r.m.s. = 0, (though the mathematician can devise examples to show that they are essentially independent). But the entropy measure of scatter has the advantage that it can be applied when the outcome can only be classified, not measured (e.g. species of fish caught in trawling, amino-acid chain produced by a ribosome.) In this paper we shall use the last measure, $H(Z)$, and we define 'successful regulation' as equivalent, to '$H(Z)$ is minimal'.

3. ERROR-, AND CAUSE-, CONTROLLED REGULATION

The reader may be wondering why error-controlled regulation has been omitted, but there has been no omission. Everything said so far is equally true of this case; for if the cause-effect linkages are as in fig. 2



R it is still receiving information about D's values, as in fig. 1, but is receiving it after a coding through S. The matter has been discussed fully by Conant (1969). There he showed that the general formulation of fig. 1 (which represents only that H must receive information from D by some route) falls into two essentially distinct classes according to whether the flow of information from D to Z is conserved or lossy. Regulation by error-control is essentially information-conserving, and the entropy of Z cannot fall to zero (there must be some residual variation). When, however, the regulator H draws its information directly from D (the cause of the disturbance) there need be no residual variation: the regulation may, in principle, be made perfect.

The distinction may be illustrated by a simple example. The cow is homeostatic for blood-temperature, and in its brain is an error-controlled centre that, if the blood-temperature falls, increases the generation of heat in the muscles and liver- -but the blood-temperature must fall first. If, however, a sensitive temperature-recorder be inserted in the brain and then a stream of ice-cold air driven past the animal the temperature rises without any preliminary fall. The error-controlled reflex acts, in fad,

only as reserve: ordinarily, the nervous system senses, at the skin, that the cause of a fall has occurred, and reads to regulate before the error actually occurs. Error-controlled regulation is in fact a primitive and demonstrably inferior method of regulation. It is inferior because with it the entropy of the outcomes Z cannot be reduced to zero: its success can only be partial. The regulations used by the higher organisms evolve progressively to types more effective in using information about the causes (at D) as the source and determiner of their regulatory actions. From here on, in this paper, we shall consider 'regulation' of this more advanced, cause-controlled type (though much of what we say will still be true of the error-controlled.)

4. MODELS

Defining 'regulation' as we have seen, is easy in that one is led rapidly to one of a few forms, closely related and easily distinguished in practical use, The attempt to define a 'model', however, leads to no such focus. We shall obtain a definition suitable for this paper, but first let us notice what happens when one attempts precision. We can start with such an unexceptionable 'model' as a table-top replica of Chartres cathedral. The transformation is of the type, in three dimensions:

$$y_1 = kx_1$$
$$y_2 = kx_2$$
$$y_3 = kx_3$$

with k about 10^{-2}. But this example, so clear and simple, can be modified a little at a time to forms that are very different. A model of Switzerland. for instance, might well have the vertical heights exaggerated (so that the three *k*'s are no longer equal). In two dimensions, a (proportional) photograph from the air may be followed by a Mercator's projection with distortion, that no longer leaves the variables separable. So we can go through a map of a subway system, with only the points of connection valid, to 'maps' of a type describable only mathematically.

In dynamic systems, if the transformation converts the real time *t* to a model time *t'* also in real time we have a 'working' model. An unquestionable 'model' here would be a flow of electrons through a net of conducting sheds that accurately models, in real time, the flow of underground water in Arizona. But the model sailing-boat no longer behaves proportionately so that a complex relation is necessary to relate the model and the full-sized boat. Thus, in the working models, as in the static, we can readily obtain examples that deviate more and more from the obvious model to the most extreme types of transformation, without the appearance of any natural boundary dividing model from non-model.

Can we follow the mathematician and use the concept of 'isomorphism'? It seems that we cannot. The reason is that though the concept of isomorphism is unique in the branch where it started (in the finite groups) its extension to other branches leads to so many new meanings that the unicity is lost.

As example. suppose we attempt to apply it to the universe of binary relations. R, a subset of E×E, and S, a subset of F×F, are naturally regarded as 'isomorphic', if there

exists a one-one mapping δ of E onto F such that $S = \delta R \delta^{-1}$ (Riguet 1948, 1951, Bourbaki 1958). But S and R are still closely related, and able to claim some 'model' relationship if the definition is weakened to

$$\exists \delta, \tau : S = \delta R \delta^{-1}$$

(with τ also one-one). Then it can be weakened further by allowing ϕ (and τ) to be a mapping generally or even a binary relation. The sign of equality similarly can be weakened to 'is contained in'. We have now arrived at the relation given earlier (1) under 'regulation'):

$$\rho \subset A \cdot \phi$$

which evidently implies some '-morphic' relation between ρ and ϕ (with *A* assumed given).

In this paper we shall be concerned chiefly with isomorphism between two dynamic systems (*S* and *R* in fig. 1). We can therefore try using the modern abstract definition of 'machine with input' as a rigorous basis.

To discuss iso-, and homo-, morphism of machines, it is convenient first to obtain a standard representation of these ideas in the theory of groups, where they originated. The relation can be stated thus:

Let the two groups be, one of the set E of elements e_1, with group operation (multiplication) δ, so that $\delta(e_i, e_j) = ek$, and other similarly of δ' on elements F. Then the second is a homomorph of the first if and only if there exists a mapping *h*, from E to F, so that, for all **Error! Objects cannot be created from editing field codes.**:

$$\delta'\left[h(e_i), h(e_j)\right] = h\left[\delta(e_i, e_j)\right] \qquad (2)$$

If *h* is one-one onto F, they are isomorphic. This basic equation form will enable us to relate the other possible definitions.

Hartmanis and Stearns (1966) definition of machine *M*' being a homomorphism of *M* follows naturally. Let machine *M* have a set *S* of internal states, a set *I* of input-values (symbols), a set *O* of output-values (symbols), and let it operate according to δ, a mapping of S×I to S, and λ, a mapping of S×I to O. Let machine M' be represented similarly by S', I', O', δ', λ'. Then M' is a homomorphism of M if and only if there exists three mappings:

h_1, of S to S'

h_2, of I to I'

h_3, of O to O'

such that, for all $s \in S$ and $i \in I$

$$h_1[\delta(s,i)] = \delta'[h_1(s), h_2(i)]$$
$$h_3[\lambda(s,i)] = \lambda'[h_1(s), h_2(i)] \qquad (3)$$

This definition corresponds to the natural case in which corresponding inputs (to the two machines) will lead, through corresponding internal states, to corresponding outputs. But, unfortunately for our present purpose, there are many variations, some trivial and some gross, that also represent some sort of 'similarity'. Thus, a more general form, representing a more complex form of relation, would be given if the mappings

h_1 of S to S', and h_2 of I to I'

were replaced by one mapping

h_4 of I×S to I'×S'.

(More general because h_4 may or may not be separable into h_1 and h_2). Then the criterion would be,

$$\forall i, s : \delta'[h_4(s,i)] = h_4[\delta(s,i)] \qquad (4)$$

a form not identical with that at (3).

There are yet more. The 'Black Box' case ignores the internal states S, and treats two Black Boxes as identical if equal inputs give equal outputs. Formally, if μ and μ' are the mappings from input to output, then the second Box is a homomorphism of the first if and only if there exists a mapping *h*, of I to I', such that:

$$\forall i \in I : \mu'[h(i)] = h[\mu(i)] \qquad (5)$$

Here it should be remembered that equality of outputs is only a special case of correspondence. Also closely related are two Black Boxes such that the second is 'de-coder' to the first: the second, given the first's output, will take this as input and emit the original input:

$$\forall i \in I : \mu' \varpi(i) = i \qquad (6)$$

This is an isomorphism. In the homomorphic relation, the input *i* and the final output $\mu'\mu(i)$ would both be mapped by *h* to the same class:

$$\forall i \in I : h\mu'\mu(i) = h(i) \qquad (7)$$

These examples may be sufficient to show the wide range of abstract 'similarities' that might claim to be 'isomorphisms'. There seem, in short, to be as many definitions possible to isomorphism as to model. It might seem that one could make practically any assertion one likes (such as that in our title) and then ensure its truth simply by adjusting the definitions. We believe, however, that we can mark out one case that is sufficiently a whole to be worth special statement.

We consider the regulatory situation described earlier, in which the set of regulatory events R and the set of events S in the rest of the system (i.e. in the 'reguland', S, which we view as R's opponent) jointly determine, through a mapping ψ, the outcome events Z. By all optimal regulator we will mean a regulator which produces regulatory events in such a way that $H(Z)$ is minimal. Then under very broad conditions stated in the proof below, the following theorem holds:

Theorem: The simplest optimal regulator R of a reguland S produces events R which are related to the events S by a mapping $h : S \to R$.

Restated somewhat less rigorously, the theorem says that the best regulator of a system is one which is a model of that system in the sense that the regulator's actions are merely the system's actions as seen through a mapping *h*. The type of isomorphism here is that expressed (in the form used above) by

$$\exists h : \forall i : \rho(i) = h\left[\sigma(i)\right] \qquad (8)$$

where ρ and σ are the mappings that R and S impose on their common input I. This form is essentially that of (5) above,

Proof: The sets R, S, and Z and the mapping $\Psi : R \times S \to Z$ are presumed given. We will assume that over the set S there exists a probability distribution $p(S)$ which gives the relative frequencies of the events in S. We will further assume that the behaviour of any particular regulator R is specified by a conditional distribution $p(R|S)$ giving, for each event in S, a distribution on the regulatory events in R. Now $p(S)$ and $p(R|S)$ jointly determine $p(R, S)$ and hence $p(Z)$ and $H(Z)$, the entropy in the set of outcomes. $(H(Z) \equiv -\sum_{z_k \in Z} p(z_k) \log p(z_k).)$ With $p(S)$ fixed, the class of optimal regulators therefore corresponds to the class of optimal distributions $p(R|S)$ for which (HZ) is minimal. We will call this class of optimal distributions π.

It is possible for there to be very different distributions $p(Z)$ all having the same minimal entropy $H(Z)$. To consider that possibility would merely complicate this proof without affecting it in any essential way, so we will suppose that every $p(R|S)$ in π determines, with $p(S)$ and ψ, the same (unique) $p(Z)$. We now select for examination an arbitrary $p(R|S)$ from π.

The heart of the proof is the following lemma:

Lemma: $\forall S_j \in S$, the set $\left\{\psi(r_i, s_j) : p(r_i, s_j) > 0\right\}$ has only one element. That is, for every s_j in S, $p(R|s_j)$ is such that all r_i with positive probability map, with s_j under ψ to the same z_k in Z.

Proof of lemma: Suppose, to the contrary, that $p(r_1|s_j)>0$, $p(r_2|s_j)>0$, $\psi(r_1, s_j) = z_i$, and $\psi(r_2, s_j) = z_2 \neq z_1$. Now $p(r_1, s_j)$ and $p(r_2, s_j)$ contribute to $p(z_1)$ and $p(z_2)$ respectively, and by varying these probabilities (by subtracting Δ from $p(r_1, s_j)$ and adding Δ to $p(r_2, s_j)$) we could vary $p(z_1)$ and $p(z_2)$ and thereby vary $H(Z)$. We could make Δ either positive or negative, whichever would make $p(z_1)$ and $p(z_2)$ more unequal. One of the

useful and fundamental properties of the entropy function is that any such increase in imbalance in $p(Z)$ necessarily decreases $H(Z)$. Consequently, we could start with a $p(R|S)$ from the class π, which diminishes $H(Z)$, and produce a new $p(R|S)$ resulting in a lower $H(Z)$; this contradiction proves the lemma.

Returning to the proof of the theorem, we see that, for any member of π and any s_j in S, the values of R for which $p(R|S)$ is positive all give the same z_k. Without affecting $H(Z)$ we can arbitrarily select one of those values of R and set its conditional probability to unity and the others to zero. When this process is repeated for all s_j in S, the result must be a member of π with $p(R|S)$ consisting entirely of ones and zeroes. In an obvious sense this is the simplest optimal $p(R|S)$ since it is in fact a mapping h from S into R. Given the correspondence between optimal distributions $p(R|S)$ and optimal regulators R, this proves the theorem.

The Theorem calls for several comments. First, it leaves open the possibility that there are regulators which are just as successful (just as 'optimal') as the simplest optimal regulator(s) but which are unnecessarily complex. In this regard, the theorem can be interpreted as saying that although not all optimal regulators are models of their regulands, the ones which are not are all unnecessarily complex.

Second, it shows clearly that the search for the best regulator is essentially a search among the mappings from S into R; only regulators for which there is such a mapping need be considered.

Third, the proof of the theorem, by avoiding all mention of the inputs to the regulator R and its opponent S, leaves open the question of how R, S, and Z, are interrelated. The theorem applies equally well to the configurations of fig, 1 and fig. 2, the chief difference being that in fig. 2 R is a model of S in the sense that the events R are mapped versions of the events S, whereas in fig. 1 the modelling is stronger; R must be a homo- or isomorphism of S (since it has the same input as S and a mapping-related output).

Last, the assumption that $p(S)$ must exist (and be constant) can be weakened; if the statistics of S change slowly with time, the theorem holds over any period throughout which $p(S)$ is essentially constant. As $p(S)$ changes, the mapping h will change appropriately, so that the best regulator in such a situation will still be a model of the reguland, but a time-varying model will be needed to regulate the time-varying reguland.

5. DISCUSSION

The first effect of this theorem is to change the status of model-making from optional to compulsory. As we said earlier, model-making has hitherto largely been suggested (for regulating complex dynamic systems) as a possibility: the theorem shows that, in a very wide class (specified in the proof of the theorem), success in regulation implies that a sufficiently similar model must have been built, whether it was done explicitly, or simply developed as the regulator was improved. Thus the would-be model-maker now has a rigorous theorem to justify his work.

To those who study the brain, the theorem founds a 'theoretical neurology'. For centuries, the study of the brain has been guided by the idea that as the brain is the organ of thinking, whatever it does is right. But this was the view held two centuries ago about the human heart as a pump; today's hydraulic engineers know too much about pumping to follow the heart's method slavishly: they know what the heart ought to do, and they measure its efficiency. The developing knowledge of regulation, information-processing, and control is building similar criteria for the brain. Now that we know that any regulator (if it conforms to the qualifications given) must model what it regulates, we can proceed to measure how efficiently the brain carries out this process. There can no longer be question about *whether* the brain models its environment: it must.

6. REFERENCES

Ashby, W. Ross, 1967. *Automaton Theory and Learning Systems*, edited by D. J. Stewart (London: Academic Press), p. 23-51

Bourbaki, N., 1958, *Théorie des Ensembles: Fascicule de Résultats*, 3rd edition (Paris : Hermann).

Conant, Roger C., 1969, I.*E.E.E. Trans. Systems Sci.*, 5, 334

Hartmanis, J., and Stearns, R. E., 1966, *Algebraic Structure Theory of Sequential Machines*. (New York: Prentice-Hall).

Riquet, J., 1948, *Bull. Soc. Math. Fr*. 76, 114; Thése de Paris.

Sommerhof, G., 1950, *Analytical Biology* (Oxford University Press).