CMU-CS-85-118

Causality in Device Behavior

Yumi Iwasaki and

Herbert A. Simon

Department of Computer Science

Carnegie-Mellon University

Supported by the Defense Advanced Research Projects Agency, Department of Defense, ARPA Order 3597, monitored by the Air Force Avionics Laboratory under contract F33615-81-K-1539. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

Causality in Device Behavior

Yumi Iwasaki and Herbert A. Simon
Carnegie-Mellon University

Abstract

This paper shows how formal characterizations of causality and of the method of comparative statics, long used in economics, thermodynamics and other domains, can be applied to clarify and make rigorous the qualitative causal calculus recently proposed by de Kleer and Brown (1984). The formalization shows exactly what assumptions are required to carry out causal analysis of a system of interdependent variables in equilibrium and to propagate disturbances through such a system.

List of Figures

Whether causal connections among events can be perceived and verified in the real world is a question much debated in philosophy. Formal treatments of the foundations of the sciences for a long time avoided notions of causation and spoke only of functional relations among variables. Nevertheless, in informal descriptions of real-world phenomena, statements of the form, "A causes B," are exceedingly common. It is the purpose of this paper to show what such statements might mean, and how they can be useful in describing the behavior of physical devices.

The foundations of our formulation have already been laid in a substantial literature that grew, some thirty years ago, out of concerns with the causal relation in the field of econometrics.[1] The formal definition of the causal relation employed in that literature will serve our purpose of describing physical devices in causal terms. In the course of our analysis, we will show that this definition is nearly the same as a recent explication of the causal relation, by de Kleer and Brown (1984); and we will both compare and contrast our own formulation with theirs. We will also show that the procedures proposed by de Kleer and Brown for the propagation of causal disturbances are closely akin to classical methods of comparative statics used widely in thermodynamics and economics.[2]

Introduction

A set of simultaneous equations is often used to describe the behavior of a physical system in terms of functional relations. For example, let s be a binary variable indicating the position of the power switch on an iron (1 if ON, 0 otherwise), and i a binary variable indicating whether or not a current is flowing in the circuit. The functional relations between s and i can be expressed as:

$$s = i.$$

Now, if l is a binary variable indicating whether the pilot lamp on the iron is lit (1 if LIT, 0

[1]The formal basis for the concepts of causality we will explicate here is developed in Simon (1952) and (1953), reprinted as Chapters 2.1 and 2.2 in Simon (1977). See also Chapters 2.3 and 2.4 in the latter volume.

[2]See, for example, Lewis and Randall (1923), pp. 180-184; Lotka (1924), ch. 22; and Samuelson (1947), ch. 2 and 3.

otherwise), the relation between I and i can be expressed as:

$$I = i.$$

Although the functional relations expressed in these two equations are identical in form, we would not treat them symmetrically if we were to express them in causal language. We would say "the position of the switch (on or off) causes the current to flow or halt," but "the flowing (interruption) of the current causes the pilot light to be on (off)." While the equations are wholly symmetric and could be commuted without changing their mathematical content, the causal statements are asymmetric, and could not be commuted without a change in meaning. To say "the current's flowing or interruption causes the switch to be on or off" would mean something quite different from "the position of the switch causes the current to flow or halt." And similarly, we would not ordinarily say that "the pilot light being on or off causes the current to flow or interrupt, respectively." Nevertheless, it is true mathematically that the value of the variable I, if known, determines the value of the variable i, and the value of the variable i determines the value of s.

Hence, we see that the causal relation is directed while the functional relation is not. If we wished to express the causal relation that is implicit in the two equations given above, we would need to use an asymmetric notation like:

$$s \dashrightarrow i \dashrightarrow I$$

Causal Inference

The first of the causal concepts we will consider is *causal ordering* (Simon, 1952), an asymmetric relation among the variables and equations of a set of simultaneous equations. It involves finding subsets of variables whose values can be computed independently of the remaining variables, and using those values to reduce the structure to a smaller set of equations containing only the remaining variables.

The second concept is *mythical causality*, developed by de Kleer and Brown (1984) as part of their construction of qualitative physics theory. They base the notion of causality on the way a disturbance is propagated from one variable to others through a network of equations or constraints.

Both approaches offer a *computational mechanism* for defining a causal dependency structure. It should be emphasized that the structure thus defined by either method only pertains to the *model* being analysed, and only denotes causal relations in the real world to the extent that this model is veridical. The question of how the "real" causal relations in a physical device can be discovered and verified will not be addressed in this paper. We will depend on common intuition to validate our analyses. Saying "current flowing causes the pilot lamp to be lit" clearly appeals more strongly to our intuitions than saying, "the pilot lamp being lit causes the current to flow." Our purpose is not to *validate* these intuitions, but to provide a formal analysis that *explicates* them. Thus, this paper is concerned with how well the operational definitions of causality proposed by the two approaches match our intuitive notions of causality, and what knowledge is required for one to come up with a good causal structure.

For the purpose of illustrating our concepts, the example of an evaporator will be used throughout the first half of this paper. An evaporator is the component of a refrigerator in which the refrigerant evaporates, absorbing heat from the air in the refrigerator chamber. The causal structure underlying the behavior of an evaporator is analyzed using the two approaches in turn. But first we must describe the device.

Behavior of an Evaporator

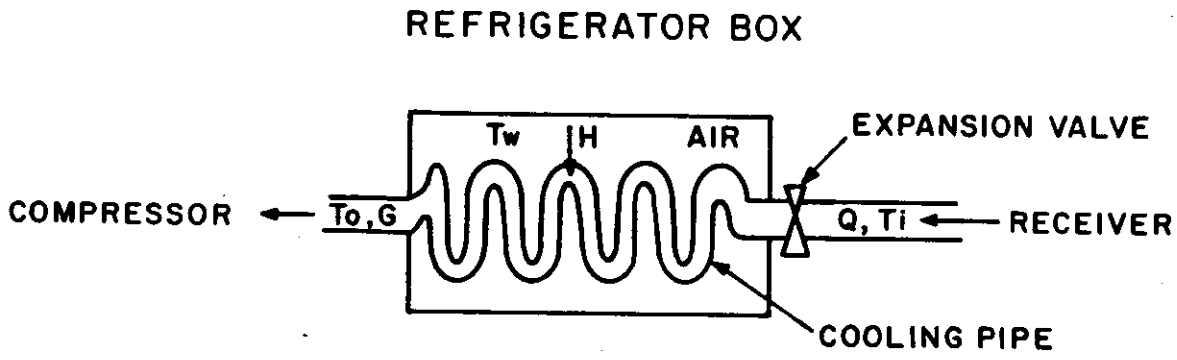

Figure 1: Evaporator

Figure 1 shows an evaporator. Liquid refrigerant flows through an expansion valve from the receiver into the evaporator. When it goes through the valve, it starts to vaporize

because of the sudden pressure drop, which causes the refrigerant's boiling temperature to fall below its current temperature. At first, the vaporization takes place without any heat flowing from the chamber into the refrigerant, because thermal energy in the liquid supplies the requisite latent heat to convert it into vapor, resulting in a sharp decrease in the temperature of the liquid refrigerant.

The refrigerant temperature continues to drop sharply until it becomes equal to the temperature (Tw) of the air within the refrigerator. It continues to decrease, but more slowly, because Tw's now being higher than the refrigerant temperature causes heat to start flowing into the refrigerant from the air in the refrigerator chamber. Eventually, the refrigerant temperature falls to the condensing temperature at the ambient pressure and stabilizes, but the refrigerant continues to boil, the latent heat now being entirely supplied by the heat absorbed through the cooling-pipe wall from the refrigerator chamber.

The refrigerant that passes through the expansion valve is in liquid phase. The refrigerant that leaves the chamber is at the condensing temperature, and is vapor, liquid, or a mixture of both. In terms of equations, the process just described can be symbolized thus:

Variables

Q	Refrigerant flow rate (mass/second)
Ti, To	Temperatures of incoming and outgoing refrigerant
G	Ratio of vapor to total mass of outgoing refrigerant
H	Heat gained by the refrigerant
P	Pressure in refrigerant within chamber
Tc	Condensing temperature of refrigerant
Tw	Temperature of air in refrigerator chamber

Constants

spl	Specific heat of refrigerant in liquid phase
l	Latent heat of refrigerant
k	Heat conduction coefficient of cooling pipe wall

Equations

1. $H = k*(Tw - Tc)*f_1(Q)$

The rate of heat transfer between the air in the chamber and the refrigerant is proportional to the temperature difference, and also is a monotonically increasing function, f_1, of the rate of flow of the fluid.

2. H = G*Q*l - (Ti - To)*Q*spl

Conservation of energy. The energy in the outgoing fluid (G*Q*l + To*Q*spl) is the sum of the heat absorbed (H) and the energy of the incoming fluid (Ti*Q*spl).

3. $Tc = f_2(P)$

The condensing temperature is a monotonically increasing function, f_2, of the pressure.

4. To = Tc

The output temperature of the refrigerant is equal to the condensing temperature.

Note that, by equation (1), the cooling of the refrigerant by evaporation to the condensing temperature is assumed implicitly to be essentially instantaneous. Note also that each equation corresponds to a specific *mechanism* for determining the value of a particular variable. Equation (1) determines the rate of heat flow into the evaporator. Equation (2) conserves the total energy by determining what fraction of the refrigerant is vaporized. Equation (3) determines the condensing temperature of the refrigerant, which is a function of its pressure. Equation (4) determines the outgoing temperature, which is the same as the internal (condensing) temperature. Describing a system in terms of the mechanisms that determine the values of its individual variables is fundamental to causal analysis. In the economics literature, equations describing such mechanisms are called *structural equations*, a term we will find it convenient to borrow.

Causal Ordering

The idea of causal ordering in a system of structural equations (Simon, 1952) can be described roughly as follows. A system of n equations is called self-contained if it has exactly n unknowns. Given a self-contained system, S, if there is a proper subset, s, of S that is also self-contained and that does not contain a proper self-contained subset, s is called a minimal complete subset. Let S_0 be the union of all such minimal complete subsets of S; then S_0 is called the set of minimal complete subsets of zero order. Since S_0 is self-contained, the values of all the variables in S_0 can, in general, be obtained by solving the equations in S_0.

By substituting these values for all the occurrences of these variables in the equations of the set ($S - S_0$), one obtains a new self-contained structure, which is called the derived structure of first order. Let S_1 be the set of minimal complete subsets of this derived structure. It is called the set of complete subsets of 1st order. Repeat the above procedure until the derived structure of the highest order contains no proper subset that is self-contained. If one denotes by V_i the set of variables in the complete subsets of ith order, where $i \geq 0$, then the variables in V_i, $(i > 0)$, are said to be *directly causally dependent* on the elements in V_{i-1}.

Causal Ordering in the Evaporator

Let us apply this procedure for determining a causal ordering to the evaporator. The structure presented in the previous section is not a self-contained structure because it contains only 4 equations in 8 variables. It is necessary to add four assumptions, in the form of additional equations, in order to make it self-contained. In the following equations, q, ti, p and tw all denote positive constants. We will say more about these assumptions later. For the moment, we will simply remark that each one represents a mechanism that is external to the evaporator (exogenous).

5. Ti = ti
6. Q = q
7. P = p
8. Tw = tw

The whole structure with the addition of the new equations is shown in Figure 2 as an 8 x 8 square matrix. The rows represent equations and the columns variables. A 1 appears in the position (i,j) if the jth variable appears in the ith equation with a non-zero coefficient, and a zero otherwise. In the matrices shown below, the zeros are omitted.

Each one of the equations, (5) through (8), above, involving only a single variable, is a minimal complete subset. Together, they constitute the minimal complete subsets of zero order. The derived structure of first order is obtained by substituting the values, tw, q, ti and p for Tw, Q, Ti and P in equations (1) through (4). The derived structures are shown in Appendix 1, together with their respective minimal complete subsets.

	H	G	Q	Ti	To	Tw	Tc	P
1.	1		1			1	1	
2.	1	1	1	1	1			
3.							1	1
4.					1		1	
5.				1				
6.			1					
7.								1
8.						1		

Figure 2: Matrix of Structural Equations for Evaporator

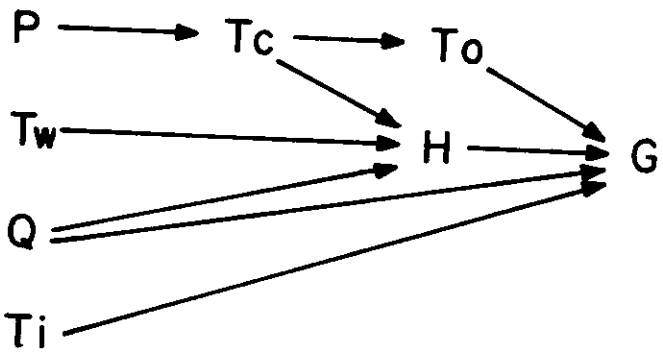

Figure 3: Causal Structure for Evaporator

Figure 3 shows the causal structure that is derived by this process. It can be explained informally as follows:

> The condensing temperature depends on (i.e., is caused by) the pressure (Equation 3), and the output temperature equals (i.e., is caused by) the condensing temperature (Equation 4). The amount of heat absorbed is determined by the condensing temperature, the temperature of the air in the refrigerator and the flow rate of the refrigerant (Equation 1). The percentage of vapor in the outgoing refrigerant depends on the condensing temperature, heat absorbed, flow rate, and input temperature (Equation 2). The pressure inside the evaporator, the flow rate of the refrigerant, the input temperature and the temperature in the chamber are determined exogenously -- that is, independently of the operation of the evaporator itself.

Now the meaning of assuming Ti, Q, P and Tw to be constant becomes clear. This amounts to assuming that their values are causally determined by mechanisms that lie outside

the subsystem under consideration -- the evaporator. They are the "uncaused causes" that drive the subsystem. Whether it is legitimate to thus regard them as exogenous is an empirical question, which could be decided by experimentation with the system. It is not a matter of fiat. It is valid only if there are not feedback mechanisms from the evaporator to the other components of the system in which the input temperature, flow rate, pressure and chamber temperature are determined. In a later section of this paper, we will return to the question of causal ordering in a system that contains such feedbacks.

Mythical Causality

The work by deKleer and Brown on qualitative physics is motivated by the desire to identify the core knowledge that underlies peoples' physical intuitions. Their interest in causality includes both the causal ordering among variables and the direction, positive or negative, of causal effects (whether an increase in variable x causes the value of variable y to increase or decrease). They argue that when people reason about the behavior of the physical environment, they do not calculate precise quantitative values but instead employ a qualitative calculus that seems to capture important aspects of the way values change. For this reason, in qualitative physics behavior is described by a set of qualitative differential equations called *confluences*. Confluences are constraint equations written in terms of qualitative derivatives of variables, whose values are +, – or 0.

In the theory of qualitative physics as developed by de Kleer and Brown, two types of behavior are discussed: Intra-state behavior, or action within a state of the system; and inter-state behavior, or the transformation of the system from one state to another. Since the boundaries between states are characterized by critical values of certain variables, the notion of state might perhaps be translated here as "phase." What de Kleer and Brown call state transformations would be familiar to the physical chemist as phase transformations. In any event, since mythical causality is concerned with intra-state behavior, and undertakes to provide causal explanations of behavior within a single state, we shall give no further attention to inter-state behavior.

Causal action, in this scheme, is brought about by some change that disturbs a system previously lying at equilibrium. Thus, if the system remains at equilibrium indefinitely, no causal action will take place and it will be impossible to discover the dependency relations among the variables. In the scheme of mythical causality, the equilibrium conditions of a static or steady-state system are momentarily relaxed by imposition of a disturbance on the system. Mythical causality "describes the trajectory of non-equilibrium states the device goes through before it re-achieves a situation where the quasi-static models are valid." (de Kleer and Brown, p. 62) In a more traditional vocabulary, "mythical causality" might well be called "causality of virtual movement," or "virtual causality."

Operationally, mythical causality is discovered in the following way: Initially, a device is presumed to be at equilibrium, satisfying the constraints of the model that describes it. This equilibrium is disturbed by a change in the value of one variable.[3] The effect of this initial disturbance is then propagated through the constraint network until all the variables are assigned new values and equilibrium is restored. The propagation is guided by the principle that disturbances are only propagated along topological paths that actually exist in the physical structure, and that are therefore represented in the equations of the model as connections between variables. When sufficient local information is not available to propagate a disturbance at any point, the propagation process comes to a halt. This occurs whenever an equation is encountered in which there remains more than one variable whose values are unknown. Then new premises need to be introduced in order to continue.

The theory of mythical causation presents three heuristic rules to govern the premises that are introduced to permit propagation to continue to completion; they are the component, confluence and conduit heuristics.

The *component heuristic* is described by de Kleer and Brown as follows: "If one pushes or pulls on one side of a component and nothing else is known yet to be acting on the

[3]Presumably, this disturbance is exogenous to the system -- a change in parameter, or in one of the exogenous variables. This restriction is not made explicit by de Kleer and Brown.

component, the component responds as if the unknown actions are negligible." (de Kleer and Brown, p. 69) The *conduit heuristic* requires that "if some component sucks stuff out of a conduit or forces stuff into a conduit the conduit's pressure drops or rises respectively as a consequence of the yet unknown behaviors of the other components attached to the conduit."(de Kleer and Brown, p. 71) The *confluence heuristic* specifies that "if some, but not enough, of the variables of a component confluence are known, propagate as if all but one of the unknown variables is zero."(de Kleer and Brown, p. 72)

De Kleer and Brown observe (p. 73) that: "The three heuristics are all very similar. Each embodies the same intuition: places where the disturbance has not yet reached are not changing." This is essentially, a *ceteris paribus* assumption. In what follows, we will not have occasion to distinguish among them, but will usually refer to the confluence heuristic, which is in a certain sense the most general among them.

Mythical Causality in the Evaporator

Consider, again, the evaporator discussed in the previous section. Several possible cases must be considered, corresponding to different relative magnitudes of various variables, hence to different phases of the system. We will consider the set of confluence equations for the case where the temperature of the refrigerator chamber, Tw, is greater than the condensation temperature, Tc, and the heat absorbed by the refrigerant while it is in the evaporator is also positive. The confluence equations can be obtained from equations (1) through (4) by taking "qualitative differentials," which amounts to replacing each variable in each equation by its differential, and concatenating these as a sum, with the appropriate sign prefixed to each term (de Kleer and Brown, 1984, pp. 21-24):

```
1A.     {dH} - {dTw} + {dTc} - {dQ} = 0
2A.     {dH} - {dG} - {dQ} + {dTi} - {dTo} = 0
3A.     {dTc} - {dP} = 0
4A.     {dTo} - {dTc} = 0.
```

Next, an initial disturbance is introduced, {dP} = +, and the equations used to find the signs of other variables:

```
5A.     {dP} = +            Initial disturbance
```

6A.	{dTc} = +	Substitute (5A) in (3A)
7A.	{dTo} = +	Substitute (6A) in (4A)

Now, there are no more equations containing a single unknown, hence the process halts. A new assumption must be introduced, using one of the heuristics. For example, the confluence heuristic may be used to assume that **{dQ}** = 0, and **{dTw}** = 0.

8A.	{dQ} = 0	Premise (confluence heuristic)
9A.	{dTw} = 0	Premise (confluence heuristic)
10A.	{dH} = -	Substitute (6A), (8A), (9A) in (1A)

Again, further progress is impossible without an assumption. Using the confluence heuristic, we assume that **{dTi}** = 0.

11A.	{dTi} = 0	Premise (confluence heuristic)
12A.	{dG} = -	Substitute (7A), (8A), (10A), (11A) in (2A)

The causal relations among the variables based on mythical causality, with the assumptions that have been made, can be seen to be identical with those in Figure 3, above. The informal explanation of the causal structure is exactly the same as before. The reason is that, by means of the choice of variables for the initial disturbance and the applications of the confluence heuristics, we have treated exactly the same variables, P, Q, Tw and Ti, as exogenous as were assumed exogenous in creating the complete structure of the causal ordering approach.

It is apparent that this causal structure depends strongly on what premises are introduced and on what the initial disturbance is. Even with the three heuristics and the principle that the propagation can only take place along the paths defined by the confluence equations, one is still left with several choices of premises that may be introduced when an assumption is needed to continue. Different choices of premises lead to different causal accounts, which deKleer and Brown call *interpretations*. For example, if we use the confluence heuristic to introduce the premises, **{dH}** = 0, **{dTw}** = 0, and **{dTi}** = 0, we obtain the derivation:

8B.	{dH} = 0	Premise (confluence heuristics)
9B.	{dTw} = 0	Premise (confluence heuristics)
10B.	{dQ} = +	Substitute (6A), (8B), (9B) in (1)
11B.	{dTi} = 0	Premise (confluence heuristics)
12B.	{dG} = -	Substitute (7A), (10B), (11B) in (2)

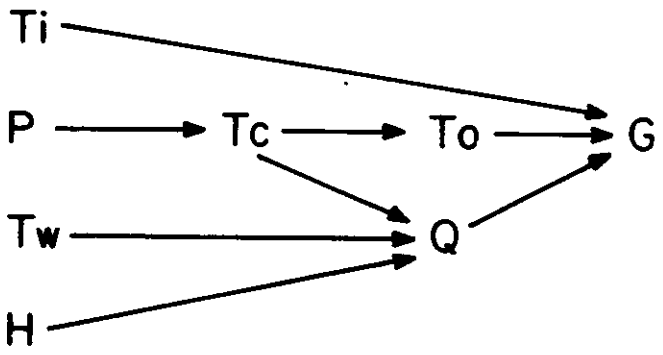

Figure 4: Alternative Causal Structure for Evaporator

This second interpretation, whose causal structure is shown in Figure 4, differs from the first in the relative positions of H and Q. The explanation based on the second interpretation will be the same as that based on the first, except for the second sentence, which must be changed to:

> The flow rate of the refrigerant depends on (is caused by) the condensing temperature, the temperature of the air in the refrigerator chamber, and the heat absorbed by the refrigerant.

The first explanation, that the heat transferred depends on the flow rate, is more compelling than the second explanation which says the fluid flow rate depends on the heat absorbed. Presumably, this is because we do not regard H as exogenously determined, but instead, believe that it is determined by the operation of the evaporator itself. The theory of mythical causality does not provide a basis for choosing between these interpretations, but simply produces all possible interpretations, some of which may match one's intuitive understanding of the physical world better than some others. Notice that each possible interpretation is obtained by annexing to the original equations an additional set of premises that are sufficient in number to produce a self-contained structure.

Comparison of the Two Concepts of Causality

Systems of simultaneous equations provide an essentially acausal representation of the behavior of a device or system. Adding causal notions involves introducing an asymmetric

relation between variables. The causal ordering approach accomplishes this by providing an operational definition of independent and dependent variables in self-contained structures. Mythical causality arrives at a causal interpretation by propagating effects of a disturbance through the network of contraint equations.

At a formal level, the two approaches are essentially indistinguishable. Given the same self-contained structure, both will produce the same causal ordering of the variables. What distinguishes them, if anything, is the rationale provided for the auxiliary assumptions that must be made to achieve self-containment of the structure. These assumptions necessarily refer to the relation of the sub-system under analysis to the larger system of which it is a component. In the case of causal ordering, the auxiliary assumptions are required to be exogenous to the subsystem, a requirement that is patently an empirical matter, and not one for arbitrary decision. In the case of mythical causation, three heuristics are proposed, two of which can be interpreted as rules of thumb for identifying exogenous variables. (The conduit heuristic is perhaps best considered as a heuristic for making explicit a constraint equation that is implicit in the description of a particular system component as a conduit.)

In the method of mythical causality, the causal ordering is discovered dynamically, by propagating the effects of a disturbance. In the method of causal ordering, it is determined by examining the equilibrium equations themselves. Again, this is a distinction without a real difference, for the causal ordering in the latter method does define a temporal order in which the equations can be solved for specific variables, and this temporal order corresponds to the order of propagation of the disturbance in the other method.

Positivity or Negativity of Causal Effects

One genuine difference between the two formulations is that mythical causation does, and causal ordering does not, specify the sign of the effect produced by one variable upon another. If A causes B, does an increase in A cause an increase or a decrease in B? Since the disturbances in the mythical causation schemes are all signed, propagation of the disturbances usually determines the signs of the effects. We have to say "usually," because

the signs may in some cases be indeterminate. Suppose, for example, that we have a qualitative difference equation of the following form:

{dz} - {dx} - {dy} = 0

Suppose, further, that the sign of {dx} has already been determined to be negative, and the sign of {dy} has already been determined to be positive. Then the sign of {dz} will be indeterminate. Indeterminacies of this kind are neither pathological nor rare. They are likely to arise in all sorts of complex systems where countervailing forces are present. For example, an increase in carbon dioxide in the atmosphere affects the atmospheric temperature through at least three different mechanisms. Two of these produce a cooling (more carbon dioxide, lower temperature). The third, the greenhouse effect, produces a heating. Only as knowledge of the relative magnitudes of these effects became available could it be concluded with relative certainty that the greenhouse effect dominates, so that an increase in the carbon dioxide will cause a warming of the atmosphere.

The determination of the signs of the effects of disturbing a system from equilibrium is, in fact, quite independent of the causal ordering of variables in the system. In any self-contained system of simultaneous equations containing one or more parameters, the system can be differentiated with respect to the parameters and individual differentials replaced by their signs, wherever these are known. If sufficient information is available about such signs, these can be "propagated" by using equations in the system to determine the signs of other variables -- just as is done in the scheme of mythical causality. Aside from the assumption that the parameters are determined exogenously, no causal ordering is implied among these variables.

This method of assigning direction to effects of disturbances from equilibrium has been used widely in chemical thermodynamics and in economics, among other fields.[4] It is often called the *method of comparative statics*, because it involves the comparison of two equilibria that differ only in the values of one or more of their parameters. Its successful application in

[4] See footnote (2), above.

any case depends on the availability of sufficient prior information about the signs of variables and of their partial derivatives in the system of equations. If equilibrium corresponds to the maximum or minimum of some function, then some of this prior information will come from the first order conditions for an extremum (derivatives equal to zero), and some may come from the second-order conditions for a maximum or minimum (certain second derivatives negative or positive, respectively). If the equilibrium is derived from a dynamic model, then information about the signs of derivatives can be derived from the conditions for stability of the equilibrium.[5]

Assigning signs to quantities in the movement of a system of interdependent relations from one equilibrium to another is quite independent of determining a causal ordering among these quantities. Knowledge of the signs of effects can be used to determine the signs of causes just as easily as signs of causes can be used to infer signs of effects.

Minimal Complete Subsets of Several Variables

In the example of the evaporator, each of the variables is the sole member of a unique minimal complete subset of some order. Again, this is a special case; in general we may expect some of the minimal complete subsets to contain more than one variable each. Consider a self-contained system consisting of three equations in the three variables, x, y and z, where all three variables appear with non-zero coefficients in all three equations. Suppose we were to try to propagate a disturbance through the confluence equations for this system by setting {dx} = +. This disturbance cannot be propagated without a second assumption, say, {dy} = 0. But now we are left with three equations in two variables, {dx} and {dz}, which, in general, will be inconsistent.

In this example, the three equations together constitute a single minimal complete

[5]For an example of the application of these methods for determining signs of effects, see Simon (1947), reprinted in (1982, ch. 3.1), where it is shown that an increase of productivity in an economy will, under certain plausible conditions, produce a relative increase of urban over rural population. For a simple sociological example, see Simon (1952), reprinted in (1982, ch. 5.4), where the method is explained in some detail. In economics, the method of comparative statics is endemic. Evans (1930) and Samuelson (1947) were among the first to show the power of second-order conditions for providing information about the signs of derivatives. Samuelson's use of the method (1947, pp. 276-280) to analyse the implications of the Keynesian theory is especially well known.

subset of zero order. They stand in no unique causal ordering, but instead there is feedback from each to the other two. In the next main section, we will take a detailed look at systems where feedback is present.

Limitations on Premises

In their development of mythical causation, de Kleer and Brown place certain restrictions on the kinds of information that may legitimately be used as sources of the assumptions needed to propagate disturbances. These restrictions are formulated as the *no-function-in-structure* principle and the *locality* principle. The former of these requires that the laws of the parts of a device (the individual equations that describe it) may not presume the functioning of the whole. The latter principle demands that the laws for a part cannot refer specifically to any other parts. Taken without qualification, these two principles exclude the possibility of basing assumptions on global or non-structural knowledge.

In our view, these principles are too restrictive, and do not correspond to intuitive notions of causality. We will comment on the use of global knowledge, knowledge of function, and knowledge of detailed structure.

Global Context. The evaporator of a refrigerator operates in the context of a larger system that includes a compressor, a condensor, and the food chamber. One can legitimately make a priori assumptions about exogenous variables based on this larger context. For example, in a refrigeration cycle, the compressor acts as a pump, thereby determining the flow rate. The refrigerant flowing into the evaporator comes from the condensor, where the liquid is cooled to room temperature, thereby determining the input temperature. The temperature of the air in the refrigerator chamber is (approximately) only dependent upon external factors, such as how much heat is produced by the objects stored in the chamber, how often the door is opened, the outside temperature, and so on. Knowing these facts, one can make the assumptions, represented as equations (5) through (8), that P, Q, Ti, Tw are, to a first approximation, exogenous independent variables, that is, determined only by factors external to the local context of the evaporator. The stricture against using global knowledge

should apply only to global knowledge that presumes, explicitly or implicitly, the functioning of the particular component that is being analysed.

Teleology and Function. Knowledge of the purpose of a component, or the way it is supposed to function, is another source of assumptions not wholly independent of global context. Conditions that are supposed to be satisfied when a component, other than the one being analysed, is operating normally may be used as additional constraints. Of course, the assumption that a component is operating normally can fail in the presence of a malfunction, but in that case the system must be described by a new and different set of equations. As long as faulty components are the exception and not the norm, knowledge of normal functioning can supply constraints. In our evaporator example, the assumption of equation (7), $P = p$, is based on the knowledge that the expansion valve is designed to maintain the pressure at a particular desired value.

Detailed Structure. Assumptions derived from global context and normal functioning are simply a substitute for more detailed description of the structure of components and the system in which they are embedded. The reason for the equivalence is that the structure of a device is designed so that it will function as it is supposed to. In the example of the evaporator, knowledge of the function of the expansion valve tells us that the pressure will be approximately constant when the component is operating normally. A detailed structural specification of the expansion valve should also show that the way it is constructed guarantees this result.

Causal analysis for trouble-shooting must take into account that components of a system may not always function as designed. Hence such analysis must allow for relaxation of the system constraints corresponding to the faulty components, and replacement of those constraints by the relations that determine behavior when the fault condition holds. Neither causal ordering nor mythical causality address directly the problem of causal analysis of a defective component -- i.e., a component that does not obey the equations by which it has been described. That can be accomplished only by redescribing the component in its faulty

state.

Assumptions derived from global context and normal functioning allow one to terminate the infinite regress toward describing a structure at more and more detailed levels. After sufficient details have been explored, one can always stop by saying, "that's the way it is designed," or "that's the way it works." This is exactly what we do when we define a computer instruction in a higher level programming language by specifying the function from its inputs to its outputs. Of course the definition is only valid if the instruction has been properly debugged -- that is, if it has been designed correctly to perform its function.

Systems with Feedback

The previous section showed that mythical causality and causal ordering are nearly equivalent so long as there is no feedback. In this section, we examine how systems involving feedback are handled by the two methods. Although the evaporator used as our example in the previous sections does not, as described, involve any feedback, feedback is present as a control element in many devices, and offers an interesting problem for causal analysis. Feedback is explained by de Kleer and Brown (1984, p. 73) in these terms:

> In the words of Norbert Wiener "feedback is a method of controlling a system by reinserting into it the results of its past performance." More technically, feedback is defined as the transmission of a signal from a later to an earlier stage. We define feedback as occurring when a sequence of cause-effect interactions produces an effect on antecedents in the sequence.

Hence, the causal diagram of a system with feedback must include a loop, showing a path by which a variable, v, affects another variable that is an antecedent in the causal chain leading to v. The feedback loop need not appear as a physical loop in the topological structure of the device.[6] Consider Figure 5, which represents a simple conduit. Liquid, at the rate of Q, flows through a pipe with input pressure, Pi, and output pressure, Po. The behavior of the system is given by the following equations:

---

[6] de Kleer and Brown, (p. XX) use the term "feedback" only in reference to a physical loop. The feedback present in the system described here, involving only a simple conduit, they call "reflection." We do not find this distinction useful, hence use "feedback" in a broader sense, both in connection with physical loops and "reflections."

1. Q = a*(Pi - Po)
2. Po = b*Q
3. Pi = p

According to the first equation, the flow rate is proportional to the pressure drop, with **a** a positive constant. By the second equation, the output pressure is proportional to the flow rate, with **b** a positive constant. By the third equation, the input pressure is exogenous.

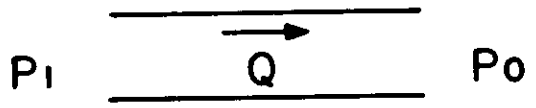

Figure 5: Simple Conduit

Causality in the Conduit

Now, if we apply the method of mythical causation to determine the causal structure of this simple system, we run into a difficulty. Suppose we introduce a disturbance by setting **{dPi}** = +. The disturbance cannot be propagated because there are two additional variables in equation (1); we must make another assumption. Using the conduit heuristic, we assume that **{dPo}** = +. Now, by the second equation, **{dQ}** = +, from which, by the first equation, it also follows that **{dPi} > {dPo}**. The result is consistent but curious, because it was obtained by assuming (through equation (2)) that **{dPo}** was the causal determinant of **{dQ}**. But surely, from the physics of the situation, a disturbance in **Pi** can be propagated to **Po** only through a change in **Q**, and not vice versa. Hence, we have not provided a plausible account of the causal mechanisms operating in this system.

It will not help matters if we use the confluence heuristic for our second assumption, setting **{dPo}** = 0. That assumption would give us **{dQ}** = + in the first equation, but propagating that disturbance to the second equation would require **{dPo}** = +, a contradiction. The technique of propagating disturbances simply does not give us a clear picture of how the system will behave. To obtain such a picture, we must use a more sophisticated process of analysis; and the method of comparative statics, especially with the use of second-order conditions for the stability of equilibrium, provides us with such a

process.

Let us suppose that the process of adjustment of Q to changes in input and output pressure is not instantaneous, but requires time, and that the same is true of the process of adjustment of Po to changes in Q. Then in place of the equilibrium equations, (1) through (3), we may introduce differential equations that, in first approximation, could be:

```
4.  dQ/dt = a*(Pi - Po) - Q
5.  dPo/dt = bQ - Po
6.  Pi = pi
```

The first two of these equations assert that, when Q and Po are disturbed from equilibrium, they tend to return to their equilibrium values at rates proportional to their deviations from them. The disturbance, as before, will take the form of a perturbation in Pi which will send the system to new equilibrium values. It is well known (see any elementary treatment of linear differential equations) that the solution to equations (4) and (5) takes the form of an exponential, with exponent equal to $-1 \pm <-(ab)>^{1/2}$. Since a and b are both positive, this means that the square-root of $-ab$ will be imaginary, and the equilibrium will be stable, being approached in an oscillatory manner. Now we can determine the signs of all of the perturbations caused by a positive disturbance in Pi. From equations (1) through (3) we will have:

```
{dQ} = a{dPi} - a{dPo}
{dPo} = b{dQ}
{dPi} = +, whence
{dQ} = a{dPi} - ab{dQ}
{dQ} = [a/(1+ab)]{dPi} = + and {dPo} = +.
```

Second Example: A Pressure Regulator

A slightly more complex feedback system, a pressure regulator (Figure 6), is analysed by de Kleer and Brown (1984). We now show that the method of comparative statics can be used, in a very similar manner to that illustrated above, to analyze the causal relations in this device. We will simplify slightly the notation used by de Kleer and Brown and their description of the regulator without omitting any essential elements.

The constraint equations governing the behavior of the pressure regulator are these:

```
1.   Xs = s - a*Po
```

PRESSURE REGULATOR

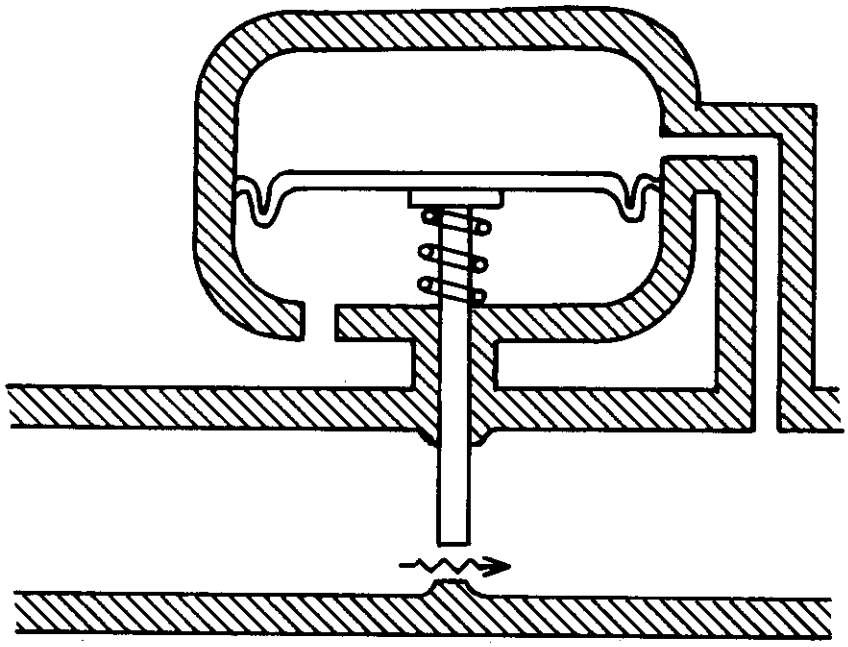

Figure 6: Pressure Regulator

This equation describes the behavior of the pressure sensor controlling the valve opening. s and a are positive constants. The pressure sensor senses the pressure at the outlet of the valve and adjusts the opening of the valve, Xs, increasing it when the pressure drops, and decreasing it when the pressure increases.

2. `Po = e*Q`

The pressure at the outlet, Po, is proportional to the flow throught the valve, Q. e is a positive constant.

3. `Q = b*(Pi - Po) + c*Xs`

The flow through the valve increases with the pressure drop from inlet to outlet and with the opening of the valve.

4. `Pi = p`

The inlet pressure, Pi, is exogenous; p is a positive constant.

As in the case of the conduit, if we undertake to propagate a disturbance through the pressure regulator, we will be faced with indeterminacies and/or contradictions. Setting {dPi} = +, the disturbance cannot be propagated without additional assumptions. Again, if

we set {dPo} = 0, we arrive at an inconsistency between equations (2) and (3), the former implying that {dQ} = 0, the latter that {dQ} = + (since, by equation (1), {dXs} = 0).

On the other hand, if we set {dPo} = +, we arrive at a consistent result, but at the cost of making the counterintuitive assumption that the disturbance in Pi is propagated to Po, and thence to Q.

These difficulties can be avoided by recognizing that the variables Q, Po, and Xs are all mutually dependent, forming a single minimal complete subset of order 1, while Pi is the sole member of the minimal complete subset of order 0. Hence, there is no causal ordering among the first-mentioned variables. Both Po and Xs modify the rate of flow, Q; Po controls Xs; and Po also modifies Q. The behavior of the system can be determined, however, by applying the method of comparative statics.

We will simplify matters slightly by assuming that the effect of the flow rate, Q, on the outlet pressure, Po (the "reflection"), is instantaneous, while the adjustment of Q and of the valve setting, Xs, take time. Then, we can write, for the dynamic behavior, the equations:

```
1.   dXs/dt = s - aPo - Xs
2.   Po = eQ
3.   dQ/dt = b(Pi - Po) + cXs - Q
4.   Pi = p
```

The time path of this system is also exponential, with exponent $[-(2 + be) + (b^2e^2 - 4ace)^{1/2}]/2$. Since the first term is negative, and the square-root term is less than the first term, the system is dynamically stable. If $b^2e - 4ac < 0$, the system oscillates on its path to equilibrium; otherwise, it approaches equilibrium exponentially.

With the assurance that the system is stable, we can now perturb the equilibrium equations, and solve them simultaneously, obtaining:

```
{dQ}/{dPi} = b/[1 + (b + ac)e] = +
{dPo} = e{dQ} = +
{dXs} = -a{dPo} = -
```

From these examples, we see that when there is feedback in a system, the effects of disturbances cannot generally be determined by propagation, but if the system is stable, they can often be determined by the method of comparative statics, which, in this case, will usually

require the solution of some simultaneous equations.

It may be objected that in these cases where the methods of mythical causality encounter difficulties or fail intuition will also fail. This may well be a psychological fact -- people may not be skillful in tracing causality through systems with feedback. And it certainly cannot be claimed that people generally see intuitively the solutions of simultaneous equations. Independently of the psychology of the matter, however, the methods of causal ordering and of comparative statics provide us with an effective classical tool, involving relatively simple mathematics, for analysing the causal relations and dynamic behavior of systems of these kinds.

Conclusion

In this paper we have shown how classical methods of causal ordering and of comparative statics can be used to determine the causal relations among the variables and mechanisms that describe a device, and to assess the qualitative effects of disturbances in the system caused by exogenous variables or parameters. These procedures, which are widely used in several fields of science, are generally consistent with, but somewhat more general than, the methods for determining mythical causation and for propagating disturbances that have recently been proposed by de Kleer and Brown.

The methods of causal ordering and comparative statics provide a rationale for the auxiliary assumptions that de Kleer and Brown use to guaranty propagation of disturbances. They are also capable of elucidating causal relations and qualitative effects in devices with feedback loops, which are handled only with some difficulty by the newer methods.

We concur fully with de Kleer and Brown that these methods (both the classical and theirs) capture some of the "physical intuition" that human beings are able to apply in reasoning about physical devices. However, the methods of comparative statics, with their ability to handle simultaneous relations, probably go beyond the limits of unaided human intuition, which seems most successful when the components of the system being analysed can be dealt with sequentially, as they are in the propagation method.

When the equilibrium of the system being examined represents the maximum or minimum of some function, the first-order and second-order conditions for extrema provide essential information about the signs of derivatives that is invaluable in inferring the qualitative effects of disturbances. When the equilibrium relations of the system can be derived from dynamic equations, the same kinds of information are provided by the conditions for dynamic stability.

An inference engine capable of the kinds of reasoning described in this paper would probably be a useful adjunct to many kinds of expert systems. Building such a system from the specifications we have sketched here does not appear to be a difficult task, and it is our intent to undertake such a construction as a next step toward understanding qualitative reasoning.

Appendix:
Derived Structures of Higher Orders for the Evaporator Model

Derived structure of the first order

	H	G	To	Tc
1.1	1			1
2.1	1	1	1	
3.1				1
4.1			1	1

minimal complete subset : 3.1
variable in the minimal complete subset : Tc

Derived structure of the second order

	H	G	To
1.2	1		
2.2	1	1	1
4.2			1

minimal complete subset : 1.2, 4.2
variable in the minimal complete subset : H, To

Derived structure of the third order

	G
2.3	1

minimal complete subset : 2.3
variable in the minimal complete subset : G

References

De Kleer, J., & Brown, J. S. A qualitative physics based on confluences. *Artificial Intelligence*, 24:7-83 (1984)

Evans, G.C. *Mathematical Introduction to Economics*, N.Y.: McGraw-Hill, 1930

Lewis, G.N., & Randall, M. *Thermodynamics and the Free Energy of Chemical Substances*, N.Y.: McGraw-Hill, 1923

Lotka, A.J. *Elements of Mathematical Biology*, N.Y.: Dover, [1924] 1951

Samuelson, P.A. *Foundations of Economic Analysis*, Cambridge, Mass.: Harvard University Press, 1947

Simon, H.A. Effects of increased productivity upon the ratio of urban to rural population. *Econometrica*, 15:31-42 (1947)

-----. A formal theory of interaction in social groups. *American Sociological Review*, 17:202-211 (1952)

-----. On the definition of the causal relation. *The Journal of Philosophy*, 49:517-528 (1952)

-----. Causal ordering and identifiability, *in* W. Hood & T. Koopmans (eds.), *Studies in Econometric Methods*, N.Y.: Wiley,1953, pp. 49-74

-----. *Models of Discovery*, Dordrecht, Holland: D. Reidel, 1977

-----. *Models of Bounded Rationality*, Cambridge, Mass: M.I.T. Press, 2 vols., 1982