292

Reprinted from the Philosophy of Science
Vol. 33 — No. 4 — December 1966

CAUSE AND COUNTERFACTUAL

BY

HERBERT A. SIMON

AND

NICHOLAS RESCHER

Reprinted from the Philosophy of Science
Vol. 33 — No. 4 — December 1966

CAUSE AND COUNTERFACTUAL*

HERBERT A. SIMON

Carnegie Institute of Technology

and

NICHOLAS RESCHER

University of Pittsburgh

It is shown how a causal ordering can be defined in a *complete structure*, and how it is equivalent to identifying the mechanisms of a system. Several techniques are shown that may be useful in actually accomplishing such identification. Finally, it is shown how this explication of causal ordering can be used to analyse causal counterfactual conditionals. First the counterfactual proposition at issue is articulated through the device of a belief-contravening supposition. Then the causal ordering is used to provide modal categories for the factual propositions, and the logical contradiction in the system is resolved by ordering the factual propositions according to these causal categories.

1. Introduction. The problem of the causal counterfactual conditional continues to loom large despite determined efforts to put it to rest. It is the thesis of this paper that the judicious combination of a formulation of the concept of a causal ordering, already available, with a notion of modal categories, also available, provide the clearest means for treating the causal counterfactual conditional.

2. Cause.[1] The causal relation is often described as a relation between events or conditions; e.g.:

The rain caused the wheat to grow.

This mode of expression commonly leads to the misconception that the asymmetry of the causal relation (i.e., the fact that cause and effect cannot be commuted) has something to do with the non-symmetry of implication—that the above statement has something to do with:

If it rains, the wheat grows.

The fatal difficulty in this view is that implication contraposes, so that we are tempted to continue:

If the wheat does not grow, it does not rain.

and thence:

The wheat's not growing causes it not to rain.

Attempts to introduce a modal relation meaning "implies causally" (e.g., Burks's, Angell's)[2] have uniformly foundered on this rock of contraposition. The lack of congruence between causality and implication is forceably indicated by the fact that (1) "If X then Y" is compatible with "If Y then X," whereas "X causes Y" is incompatible with "Y causes X," and (2) "If X then Y" entails "If not-Y then not-X"

* Received August, 1966.

[1] Technical details underlying the following explication of "cause" will be found in [5]. The account here has been generalized, however, to encompass nonlinear as well as linear structures.

[2] See [3] and [2].

whereas "X causes Y" not merely fails to entail "not-Y causes not-X" but is actually incompatible with it. We establish as a regulative guidepost the principle:

Principle 1: The asymmetry of the causal relation is unrelated to the asymmetry of any mode of implication that contraposes.

If we are not bound to relate causality to implication, then we may reconsider, at the outset, what is being related by causal statements. It is usually suggested that the wheat's growing is related to the rain. Let us propose as an alternative that it is rather the size of the wheat crop that is causally related to the weather. That is to say, the following three statements are all part of a single causal relation:

The absence of rain prevents the wheat's growing.
With moderate rain, the wheat crop is good.
With heavy rain, there is a large wheat crop.

or generalizing:

The amount of wheat is a function of the amount of rain.

We draw a second pragmatic conclusion from the example:

Principle 2: A causal relation is not a relation between values of variables, but a function of one variable (the cause) on to another (the effect).

Regarding causality as functional relation eliminates the unwanted asymmetry produced by contraposition, for contraposition does not interchange an independent with a dependent variable. On the other hand, it is not immediately obvious that the asymmetry of functional relation provides a suitable interpretation of the wanted asymmetry between cause and effect; for many, if not most, of the functions that enter in causal discussions possess inverses; and by inverting them, we can interchange a dependent with an independent variable. Thus if ϕ possesses an inverse, ϕ^{-1}, then, from

$$y = \phi(x)$$

we obtain

$$x = \phi^{-1}(y)$$

Therefore the distinction between independent and dependent variables does not explicate, by itself, the asymmetry between cause and effect. Surely we wish to invest the latter distinction with more significance than is accorded by the arbitrary choice of which variable to measure on the abcissa, and which on the ordinate.

3. Complete Structures.

We turn now to the positive task of showing that, given a system of equations—functional relations—and a set of variables appearing in these equations, we can introduce an asymmetric relation among individual equations and variables which corresponds to our commonsense notion of a causal ordering. (When we have occasion to write out functional relations explicitly, we shall generally write them in the canonical form $f_i(x_1 , x_{2_i} ,...) = 0$, where f_i is a name for the function, and the x's are the variables that appear in it.)

Definition 1: A ***structure*** is a set of m functions involving n variables ($n \geqslant m$), such that:

(a) In any subset of k functions of the structure, at least k different variables appear.

(b) In any subset of k functions in which r ($r \geqslant k$) variables appear, if the values of any ($r - k$) variables are chosen arbitrarily, then the values of the remaining k variables are determined uniquely. (Finding these unique values is a matter of solving the equations for them.)

For illustration, we will sometimes consider *linear structures*—i.e., structures in which the functions are linear and non-homogeneous. A linear structure is a set of independent and consistent linear non-homogeneous equations.

Definition 2: A structure is *self-contained* if it has exactly as many functions as variables.

A self-contained structure can be solved for a unique set of values of its variables.

A structure can be represented simply by a matrix of 1's and 0's, the various columns of the matrix being associated with the variables of the structure, and the rows with the functions. Then a 1 in the j^{th} column of the i^{th} row means that the j^{th} variable appears in the i^{th} function, while a zero in that position means that the j^{th} variable does not appear in the i^{th} function.

Consider the following structure matrix:

$$\begin{array}{cccccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\
f_1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
f_2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
f_3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
f_4 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\
f_5 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
f_6 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
f_7 & 0 & 0 & 0 & 0 & 1 & 0 & 1
\end{array}$$

By definition 2, this matrix represents a self-contained structure. Since f_1, f_2, and f_3 each contain only one variable (x_1, x_2, and x_3, respectively) each is also a self-contained structure (and obviously a *minimal* self-contained structure), and each can be solved, by Definition 1, for the value of its variable.

If we now substitute these values of x_1, x_2, and x_3 in f_4 through f_7, we obtain a new *derived structure of first order* with the following matrix (which is simply the lower right-hand 4 × 4 component of the original structure):

$$\begin{array}{ccccc}
 & x_4 & x_5 & x_6 & x_7 \\
f'_4 & 1 & 1 & 0 & 0 \\
f'_5 & 1 & 1 & 0 & 0 \\
f'_6 & 1 & 0 & 1 & 0 \\
f'_7 & 0 & 1 & 0 & 1
\end{array}$$

In this derived structure the set consisting of f'_4 and f'_5 is a minimal self-contained structure, which can be solved for the values of x_4 and x_5. Substituting these values in f'_6 and f'_7, we obtain the *derived structure of second order:*

$$\begin{array}{ccc}
 & x_6 & x_7 \\
f''_6 & 1 & 0 \\
f''_7 & 0 & 1
\end{array}$$

This structure consists of the two minimal self-contained structures, f''_6 and f''_7, which can be solved for x_6 and x_7, respectively.

We see that there was a certain asymmetry in the equations and variables of our original structure, such that subsets of equations could be solved for certain variables without solving for others, but not vice versa. We may depict this ordering:

$$\begin{tikzcd} f_1 \arrow[dr] & \\ f_2 \arrow[r] & \left\{\begin{matrix} f_4 \to f_6 \\ f_5 \to f_7 \end{matrix}\right. \\ f_3 \arrow[ur] & \end{tikzcd}$$

or, alternatively, in terms of variables:

$$\begin{tikzcd} x_1 \arrow[dr] & \\ x_2 \arrow[r] & \left\{\begin{matrix} x_4 \to x_6 \\ x_5 \to x_7 \end{matrix}\right. \\ x_3 \arrow[ur] & \end{tikzcd}$$

It is clear that variables belonging to derived structures of higher order are *dependent* on variables belonging only to the lower-order structures, while the latter variables are *exogenous* to the structures determining the former. We shall interpret the ordering as a causal ordering, so that a variable at the head of an arrow is *directly caused* by the variable or variables at the origin of the arrow. Thus x_4 and x_5 are directly caused by x_1 (and jointly by x_2 and x_3 also), x_6 by x_4 and x_7 by x_5.

By recursion, we can then define the transitive relation, *caused*, so that x_7, for example, is caused by x_1 (jointly with x_2 and x_3, and via x_5).

Let us see to what extent these definitions lead to results that conform to English usage. Consider our example:

The rain causes the wheat to grow.

Add the following statements, which we wish to interpret as simultaneously valid:

Fertilizer causes larger wheat yields.
A large wheat crop causes the wheat price to fall.
An increase in population causes the wheat price to rise.

We define the following variables: R is the rainfall in the given year; W, the size of the wheat crop; F, the amount of fertilizer used; P, the price of wheat; N, the size of the population. We next represent our first two causal sentences by the following functional relation:

$$f_1(R, W, F) = 0$$

Then we represent the last two causal sentences by another functional relation:

$$f_2(W, P, N) = 0.$$

The two functions together do not define a self-contained structure, since they contain five variables. Let us suppose the structure completed by adding three additional functional relations describing (a) a theory of the weather, (b) a theory of how fertilizer applications are decided upon, and (c) a theory of population growth. The function (a) should contain R but none of the other four variables in our system, (b) should contain, of these variables, only F, and (c) should contain only N:

$$\text{(a) } f_3(R) = 0, \text{ (b) } F_4(F) = 0, \text{ (c) } f_5(N) = 0$$

We now have a complete structure, with matrix:

$$\begin{array}{cccccc}
 & R & W & F & P & N \\
f_1 & 1 & 1 & 1 & 0 & 0 \\
f_2 & 0 & 1 & 0 & 1 & 1 \\
f_3 & 1 & 0 & 0 & 0 & 0 \\
f_4 & 0 & 0 & 1 & 0 & 0 \\
f_5 & 0 & 0 & 0 & 0 & 1
\end{array}$$

It is easily seen that this structure determines the causal ordering:

$$\begin{tikzcd}
f_3 \arrow[dr] & & & & R \arrow[dr] & & \\
& f_1 \arrow[dr] & & & & W \arrow[dr] & \\
f_4 \arrow[ur] & & f_2 & & F \arrow[ur] & & P \\
f_5 \arrow[urr] & & & & N \arrow[urr] & &
\end{tikzcd}$$

Reading off the relations in the diagram, we find: The amount of rain (R) and the amount of fertilizer (F) are the causal determinants of the size of the wheat crop (W), while the size of the wheat crop (W) and the population (N) are the causal determinants of the price of wheat (P).

Thus the formalization translates accurately the causal assertions in the original English-language sentences. We are not now asserting that these causal statements are empirically correct (nor have we explained what might be meant by such an assertion); we are merely showing that the formalization captures the common meaning of "cause."

Note that f_1 and f_2, by themselves, are entirely symmetrical in the variables they contain. It is only when they are imbedded in a complete structure, containing f_3, f_4, and f_5, that asymmetry appears. We do not need to designate which is the dependent variable in each of these relations taken singly. Hence the formalization does not rest on the essentially arbitrary distinction between independent and dependent variables.

On the other hand, it is essential that the structure we consider be complete, and if we complete a structure in a different way, we will generally find that we have altered the causal ordering. In the previous example, suppose we replace $f_5(N) = 0$ by $f_6(P) = 0$ ("the wheat price is fixed by the government," say). The reader can easily verify that the causal ordering in this modified structure is:

$$\begin{tikzcd}
f_3 \arrow[dr] & & & & R \arrow[dr] & & \\
& f_1 \arrow[dr] & & & & W \arrow[dr] & \\
f_4 \arrow[ur] & & f_2 & & F \arrow[ur] & & N \\
f_6 \arrow[urr] & & & & P \arrow[urr] & &
\end{tikzcd}$$

This system now asserts that the population is determined by the price of wheat and the size of the wheat crop (that is, the population will reach the size that will just consume the wheat crop at the given price). Now, by common sense, we might suppose that given any particular amount of wheat, the price would be higher the larger the population. From this assumption and the new causal ordering we reach the curious conclusion that, by raising the price of wheat, the government can increase the population (and without increasing the amount of wheat raised!).

For this and other reasons, we would intuitively regard the original structure as the empirically valid one, and the modified one as invalid. Again we postpone the discussion of what is meant by "valid" here and observe simply that different causal orderings obtain when a partial structure is completed in different ways.

4. Invariance. Suppose that we have a complete structure of n functions and n variables. In general, we can replace any one of the functions of the structure by a linear combination of it with one or more of the others, without altering the values of the variables satisfying the entire set of functions. Thus, if we have a complete structure in three variables consisting of $f_1 = 0$, $f_2 = 0$, $f_3 = 0$, the structure consisting of $f_1 = 0$, $f_2 = k$, $f_1 + k_2 f_2 + k_3 f_3 = 0$, $f_3 = 0$ (k_1, k_2, k_3 are non-zero constants) will also (in general) be complete, and will have the same solution as the original structure. However, the two structures will not generally have the same ordering. For example, suppose the matrix of the first structure was:

$$\begin{array}{cccc} & x_1 & x_2 & x_3 \\ f_1 & 1 & 0 & 0 \\ f_2 & 0 & 1 & 0 \\ f_3 & 1 & 1 & 1 \end{array}$$

The ordering would be

$$\begin{tikzcd} x_1 \arrow[dr] & \\ & x_3 \\ x_2 \arrow[ur] & \end{tikzcd}$$

But the second structure would have (in general) the matrix:

$$\begin{array}{cccc} & x_1 & x_2 & x_3 \\ f_1 & 1 & 0 & 0 \\ f'_2 & 1 & 1 & 1 \\ f_3 & 1 & 1 & 1 \end{array}$$

with an ordering:

$$x_1 \rightarrow \begin{cases} x_2 \\ x_3 \end{cases}$$

It might be thought that the effect of replacing individual functions by linear combinations of several would always be to introduce new variables in the modified function, as in the above example. But this is not so, for the linear combination may turn out to be an identity in one (or more) of the variables, which then can be cancelled out. Take $k_1 = -1$, $k_2 = -1$, and $k_3 = 1$ in the last example, and suppose $f_1 = x_1$, $f_2 = x_2$ and $f_3 = x_1 + x_2 + ax_3$. Then $f'_2 = k_1 f_1 + k_2 f_2 + k_3 f_3 = ax_3$, yielding the ordering:

$$\begin{tikzcd} f_1 \arrow[dr] & & & x_1 \arrow[dr] & \\ & f_3 & \text{or} & & x_2 \\ f_2 \arrow[ur] & & & x_3 \arrow[ur] & \end{tikzcd}$$

In algebra, operations of replacing rows (columns, respectively) in a matrix by linear combinations of rows (columns, respectively) are called *elementary row* (column, respectively) *operations*. The application of elementary row operations to a system of equations does not change the set of solutions to the equations. Indeed, a standard technique for solving simultaneous linear equations is to apply elementary row operations to obtain a diagonal matrix — with one variable in each equation.

But if solutions to equations are invariant under elementary row operations, the causal orderings of variables in complete structures are not — as our example has shown. If causal ordering is to have more than conventional or notational significance, we must have some basis for singling out from among a whole class of matrices that are equivalent under the group of row transformations the particular matrix that represents the empirically valid causal ordering. We turn now to this problem.

Perhaps we can get a clue to the answer by considering the analogy of the elementary *column* operations, which are not admissible algebraic operations. Each column of the matrix of a structure corresponds to a variable. A column operation would replace some single variable of the system by a linear combination of variables. In the earlier example, it might replace "price of wheat" by "twice the price of wheat minus four times the population." This operation is inadmissible because it destroys the identity of the variables in terms of which the problem is stated — variables that presumably correspond to empirical observations on the system. Hence, among all matrices equivalent under elementary column transformations, that one alone is uniquely admissible which puts columns and variables in one-to-one correspondence.

Now let us return to row transformations, and assign a label to each function, the label to denote the *mechanism* (a term here introduced informally) which that particular function represents. In the wheat example, f_1 represents the biochemical mechanism involved in the growth of wheat, f_2 represents the economic mechanism relating to wheat buying, f_3 is the meteorological mechanism that determines the weather, f_4 is the producers' decision mechanism with respect to fertilizer, and f_5 is the mechanism of population growth.

An elementary row transformation would replace one of these mechanisms with a linear combination of several of them. For example, replacing f_3 by a combination of f_1 and f_3 would introduce a composite biological-meteorological mechanism pertaining to wheat growth and the weather. Hence, among all matrices equivalent under elementary row transformations, that one alone is uniquely admissible that puts rows and mechanisms in one-to-one correspondence.

It is intuitively clear how we identify the variables of the system (and distinguish them from linear combinations of variables). Our intuitions seem less clear about identifying mechanisms. Unless we regard this identification as intuitively obvious, we have simply substituted a new problem for the original one. The new problem — how to identify mechanisms — may well, however, turn out to be more tractable than the old one.

5. Identifiability of Mechanisms. First, we shall show that no amount of observation of the values of the variables in a complete structure can identify the mechanisms—can distinguish a particular matrix from all those equivalent to it under elementary row transformations. The proof is immediate. Suppose we have a set of k consistent equations in n variables ($n \geqslant k$), and suppose that ($\bar{x}_1$, $\bar{x}_2$,..., $\bar{x}_n$ is an empirically observed set of values for the n variables that satisfies the equations. Then these observed values will also satisfy any set of equations equivalent to the original set under elementary row transformations — all such sets of equations having the same solutions. Thus no number of simultaneous observations of rainfall, fertilizer, wheat crop, wheat price, and population will verify the causal ordering in our example.

Notice that the causal ordering depends on which variables *do not* appear in which mechanisms. Thus, in the wheat example, to introduce the mechanism $f_3(R) = 0$ is equivalent to asserting that a meteorological theory can be constructed that predicts rainfall independently of fertilizer practices, wheat crop, price of wheat, or size of population. With respect to this set of variables, weather is an unmoved mover — an exogenous variable. Similarly, in this structure, population and fertilizer are asserted to be exogenous variables. These assumptions are crucial to the causal ordering.

Cosmology might provide one basis for such assumptions. It might be assumed, for example, that the behavior of any system involving very large quantities of energy (e.g., the atmosphere), is practically autonomous of the behavior of variables involving very much less energy (e.g., wheat growing). We may call this principle the Postulate of Prepotence.

Or, it might be assumed that most variables in the world are not directly connected with most other variables, and that such connections as exist involve a very small number of different kinds of mechanisms. Then, one would include a particular variable in a subsystem only if one could select a mechanism from the list of admitted mechanisms through which that variable could possibly act on that subsystem. We might call this assumption the Postulate of Independence, or, more vividly, the Empty World Postulate.

To see that these cosmological assumptions really correspond to the way we reason about causality, consider the objections that might be raised against the proposed causal ordering of the wheat example. First, it might be argued that the wheat crop influences the weather, since the àcreage planted to wheat affects the rate of evaporation of ground moisture. Notice the objection conforms to the Empty World Postulate, since it does not simply urge that anything may influence anything else, but proposes a specific mechanism of a kind known to be efficacious in other situations. If the proposal to include the wheat crop as a variable affecting the weather were rejected, the Postulate of Prepotence could provide a plausible basis for the rejection.

A similar discourse could examine the plausibility of assuming that the amount of fertilizer applied is independent of the price of wheat, or the population of the size of the wheat crop. To carry out this discussion in detail would call for the static structure considered so far to be expanded into a dynamic model. (We shall postpone questions of dynamics to a later point.)

Having offered the Postulate of Prepotence and the Empty World Postulate as possible sources for the identifying assumptions underlying a causal ordering, we leave this foray into cosmology to consider other possible bases for identification.

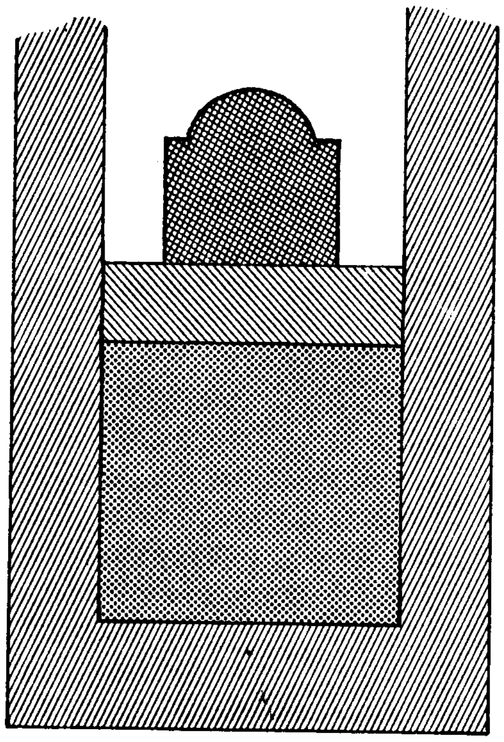
6. Intervention. In many, though not all, contexts where causal language is used, the structure under examination can be altered by intervention. The specific possibilities for intervention then become bases for causal attribution. "Intervention," as the term is used here, has nothing to do with change through time, hence we can illustrate its significance with a wholly static example.

Consider the physical situation depicted in Figure I. A quantity of gas is confined in a chamber by a moveable piston (having unit area of cross-section) on which rests a weight. Assuming the Gas Laws hold in this situation, we have, in equilibrium:

(1) $PV = kT$

where P is pressure per unit area (equal to W, the weight resting on the piston),

V is the volume of the chamber, T is the absolute temperature of the confined gas, and k is a constant that depends on the amount of gas confined. We assume that, under conditions to be specified presently, heat may pass in or out through the walls of the chamber.



THE GAS LAWS

Figure I

Since we have only a single equation, with three variables, we must impose additional constraints to obtain a complete structure and define a causal ordering. We will impose these constraints by assumptions about the possibility of intervention. *Case I.* We assume that the possibility of heat passing in and out of the chamber can be altered. In the first case (constant temperature), we assume that the heat flows so readily that, at equilibrium, the temperature inside the chamber must always equal the temperature outside. Representing the latter by $\bar{T}$, a constant, we obtain:

$$(2) \quad T = \bar{T}$$

Next, we assume that the weight on the piston is also determined exogenously: the "experimenter" may impose any weight, $\bar{W}$, he wishes, just as he may maintain any outside temperature, $\bar{T}$, he wishes. From this new assumption, we get:

$$(3) \quad P = \bar{W}$$

Now, equations (1) through (3) define the complete structure

	P	T	V
(1)	1	1	1
(2)	0	1	0
(3)	1	0	0

with causal ordering

$$\begin{tikzcd} P \arrow[dr] & \\ & V \\ T \arrow[ur] & \end{tikzcd}$$

Thus, we might make the following kind of statement about the system: "In order to decrease the volume of the chamber, increase the weight on the piston, or decrease the temperature of the environment."

Note that we have kept our promise of avoiding dynamics, for these statements do not refer to temporal change, but are statements in comparative statics. They can be put more formally:

> "If, in two situations, $\bar{W}_1 > \bar{W}_2$, then $V_1 < V_2$, *ceteris paribus*, and if $\bar{T}_1 > \bar{T}_2$, then $V_1 > V_2$, *ceteris paribus*."

Case II. We assume (adiabatic case) that the walls of the chamber have been perfectly insulated so that no energy can pass through. The adiabatic assumption imposes on the system a constraint that was absent in Case I — that the total energy of the system must be conserved. This total energy $\bar{E}$, is the sum of the potential energy PV, (since V is equal to the height of the chamber), of the weighted piston, and the heat energy, $Q = qT$, of the gas in the chamber. Hence, the constraint may be written:

$$(2)' \qquad \bar{E} = PV + qT$$

Next, we assume again that the weight on the piston is determined exogenously. From mechanisms (1), (2)′, and (3), we get the new structure:

	P	T	V
(1)	1	1	1
(2)′	1	1	1
(3)	1	0	0

with quite different causal ordering:

$$\begin{tikzcd} & T \\ P \arrow[ur] \arrow[dr] & \\ & V \end{tikzcd}$$

If we regard the interventions themselves, i.e., $\bar{T}$, $\bar{W}$, and $\bar{E}$ as "variables," then the causal diagrams for the two cases can be expanded to:

$$\begin{tikzcd} \bar{W} \arrow[r] & P \arrow[dr] & & & \bar{E} \arrow[r] & E \arrow[dr] & \\ & & V & \text{and} & & & \left\{\begin{matrix}T\\V\end{matrix}\right. & \text{respectively.} \\ \bar{T} \arrow[r] & T \arrow[ur] & & & \bar{W} \arrow[r] & P \arrow[ur] & \end{tikzcd}$$

In the adiabatic case, it is not immediately obvious in what sense the experimenter can intervene to fix $\bar{E}$. This can be explained as follows. Let the system be in equilibrium for $\bar{E}_1$, $\bar{W}_1$, and let P_1, T_1, V_1 be the equilibrium values of the endogenous variables. Now suppose the weight $\bar{W}_1$ is suddenly replaced by a new weight $\bar{W}_2$. The system, now not in equilibrium, has had its energy increased by the amount $V_1(\bar{W}_2 - \bar{W}_1)$ to the total: $\bar{W}_2V_1 + kT_1 = \bar{E}_2$. This is the quantity, $\bar{E}_2$, of equation (2)′, and, it can be seen that the experimenter fixes it by setting V_1, T_1 and $\bar{W}_2$.

In the two different piston cases, how can we specify operationally what variables are being measured and what mechanisms are operating? With respect to the former, the instrumentation required is a thermometer and pressure gauge on the interior of the chamber to measure T and P, respectively, and a scale against which to mark the position of the piston, hence to measure V. Equation (3) derives from the fact that the observed value of P changes if and only if we change $\bar{W}$. The change in the value of the variable is associated, then, with a change in one specific part of the structure, separated, physically and visually, from other parts.

In Case I, Equation (2) derives from the fact that T changes if and only if we change the temperature of the surrounding bath. Both $\bar{T}$ and $\bar{W}$ are observable in the same sense that P, T, and V are observable. Hence, if we can forbid column transformations because they would merge and confuse the operationally distinct measures, P, T, V, we can forbid row transformations because they would merge and confuse the operationally distinct interventions, $\bar{T}$ and $\bar{W}$. Case II is slightly more complicated because of the less transparent status of $\bar{E}$ as an operationally distinct intervention, but its analysis is the same in principle.

In both Cases I and II, Equation (2) also depends on the mechanisms of the boundary between the chamber and its environment. In Case I, the equation implies a law of heat flow that does not admit temperature differentials in equilibrium. In Case II, it implies perfect thermal insulation across the boundary. In a full dynamic treatment of the situation, distinct mechanisms would appear to describe these phenomena across the boundary. Again, particular mechanisms refer to distinct parts — often but not always visually distinct parts — of the total system.

7. A Dynamic Example. Having seen how interventions can be used to define complete structures, and hence causal orderings, we turn next to nonexperimental situations where intervention is not possible. Does the notion of causal ordering apply at all to such situations, and can we identify mechanisms in them?

Let us take a simple example from classical (pre-relativity and pre-Sputnik) celestial mechanics, considering motions in a single dimension to minimize the mathematical complexities. Combining Newton's Second Law with the Inverse Square Law of gravitational attraction, we describe the motion of a set of n mass points by the n equations:

$$(1) \qquad a_i(t) = g \sum_{j \neq i} \frac{m_j}{[x_i(t) - x_j(t)]^2} \ (i = 1, \ldots, n)$$

where $a_i(t)$ is the acceleration of the i^{th} mass point, g the gravitational constant, m_i the (constant) mass of the i^{th} mass point, and $x_i(t)$ the position of the i^{th} mass point. Integration of these equations twice gives the time path of the system. Now, to see more clearly what would be the significance of elementary row operations on this system, we consider the discrete approximation; $a_i(t) \sim x_i(t+2) - 2x_i(t+1) + x_i(t)$, and we rewrite the system

$$(2) \qquad f_i[x_i(t+2),\ x_i(t+1),\ x_i(t),\ \{x_j(t)\}_{j \neq i};\ \{m_j\}_{j \neq i}, g] = 0,\ (i = 1, \ldots, n).$$

Consider now only those terms of the functions that refer to times other than t. In the ith function, these terms involve only positions of the i^{th} mass point. The form in which the structure is defined by (2) may be regarded as canonical in the sense that elementary row transformations will destroy the property just mentioned. That is, after non-trivial transformations, there will be functions in the structure that refer to the positions at times other than t of more than one mass point.

We may restate the matter differently: Any system of differential equations can be reduced, by introducing additional variables to eliminate higher-order derivatives, to a system of the first order. Let us consider a system of first order, and let us introduce a concept of *self-contained dynamic structure* in analogy to our earlier definition of self-contained structure.

Definition 3: A self-contained dynamic structure is a set of n first-order differential equations involving n variables such that:

(a) In any subset of k functions of the structure the first derivatives of at least k different variables appear.

(b) In any subset of k functions in which $r(r \geqslant k)$ first derivatives appear, if the values of any $(r - k)$ first derivatives are chosen arbitrarily, then the remaining k are determined uniquely as functions of the n variables.

By performing elementary row operations on a self-contained dynamic structure, we can solve for the n first derivatives — i.e., replace the structure by an equivalent structure possessing the canonical form described above. Our proposal, then, is that the functions of the structure in this form be interpreted as the mechanisms of the system. The i^{th} function (the function containing dx_i/dt) is then the mechanism determining the time path of x_i. Elementary row operations will be inadmissible since they intermingle first derivatives of variables, just as elementary column operations are inadmissible in intermingling variables.

Notice that a complete dynamic structure is not analogous to a complete structure in the static case. To complete the dynamic structure in the latter sense, we must specify a set of initial conditions, e.g., $x_i(t_0)$ $(i = 1,..., n)$, the values of the n variables for some particular time, t. With this addition a causal ordering is defined: for $t_0 < t$ it will be the normal causal ordering, acting forward in time, but for $t_0 > t$ the directions of the causal arrows will all be reversed. Therefore, we may say that a complete dynamic structure defines a causal ordering up to reversal of directionality. Thus, most features of the ordering (i.e., the forms of the functions) are independent of time precedence.

In particular, note that time reversal is *not* equivalent to the countraposition of implication. For if the sense of time is inverted in Equation (1), the accelerations will still be causally dependent on the gravitational constant and the masses (which are exogenous), and on the instantaneous positions of the mass points, and not *vice versa*. What is reversed is just that, in the originally-stated system accelerations and the present state of the system are the causes of *subsequent* states, in the reversed system, they are causes of *prior* states. In both cases, states cause accelerations, by the gravitational mechanism, while accelerations, by definition, are second derivatives of positions.

8. Equilibrium of Dynamic Systems.

Suppose that we observe the behavior of a dynamic system, described in canonical form, over some period of time. We can divide the variables, by rough criteria, in three classes:

1. Variables that have changed so slowly that they can be replaced by constants for the period under observation, deleting the corresponding mechanisms from the system.

2. Variables that have adjusted so promptly that they are always close to (partial) equilibrium, hence their first derivatives always close to zero. We can replace the first derivatives of these variables by zero in their equations — substituting static equilibrium mechanisms for the original dynamic mechanisms. We will continue to regard variables whose first derivatives have been set equal to zero as the dependent variables in the corresponding equations.

3. All other variables. We will retain their equations in canonical form.

Returning to the wheat crop example, let us complicate the system, first, by assuming that all processes take time, but that the processes determining W and P are relatively rapid, and second, by introducing additional feedbacks:

The amount of fertilizer used will adjust (slowly) to previous levels of the wheat price;
the population will gradually adjust to the amount of wheat available; and
the weather will be slowly changed by the amount of wheat acreage and the size of population.

We might write the full dynamic system schematically in some such form as:

$$dW/dt = f_W(W, R, F)$$

$$dP/dt = f_P(W, P, N)$$

$$\frac{dF}{dt} = f_F(P, F)$$

$$\frac{dN}{dt} = f_N(N, W)$$

$$\frac{dR}{dt} = f_R(N, R, W)$$

The matrix of coefficients on the right-hand sides is given by:

$$\begin{array}{cccccc} & R & W & F & P & N \\ \dot{R} & 1 & \epsilon & 0 & 0 & \epsilon \\ \dot{W} & 1 & 1 & 1 & 0 & 0 \\ \dot{F} & 0 & 0 & 1 & \epsilon & 0 \\ \dot{P} & 0 & 1 & 0 & 1 & 1 \\ \dot{N} & 0 & \epsilon & 0 & 0 & 1 \end{array}$$

where we have introduced ϵ's instead of 1's as the off-diagonal elements in those processes assumed to be "slow" relative to the others. That is to say, we assume R, F, and N can be replaced (approximately) by constants over a period of, say, one year. If we let the ϵ's approach zero, we are back to the same matrix as in the static case. Moreover, if we then assume that W and P adjust rapidly, we can set their derivatives equal to zero in the dynamic system, obtaining precisely our original static structure.

The notion of "nearly ordered" dynamic system illustrated by this example can be appropriately formalized as has been shown by Ando, Fisher, and Simon.[3] Matrices that can be block-triangularized by letting certain elements go to zero at the limit are called ***nearly decomposable***. Systems described by nearly decomposable matrices

[3] See [1]. The formal development of the theory is given in Chapters 4 and 5, and a number of illustrative examples in Chapter 6.

have a number of important special dynamic properties, on one of which the present discussion rests.

We see that the causal ordering in the static case can be interpreted as a set of implicit consequences of assumptions about the relative speeds of various processes in an associated dynamic model. Given these assumptions, the static model represents an approximate short-run equilibrium of the dynamic system.

In moving from the static to the dynamic interpretation of the causal ordering, the exogenous variables, or interventions, in the static system become the "unmoved movers" of the dynamic system — i.e., variables that act strongly on other variables of the system but are only weakly acted upon by other variables. The definite asymmetry in the matrix of the static system corresponds to relative asymmetry in the matrix of the associated dynamic system.

9. Discrete Variables. Our final task is to apply the causal ordering notions to the kinds of standard examples that involve discrete variables:

> Striking a dry match in the presence of oxygen, tinder, and fuel will cause a conflagration.

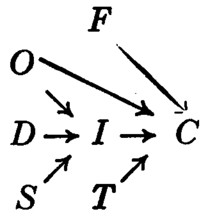
We define the following dichotomous variables: (1), (*S*) struck or unstruck; (2), (*D*) dry or damp; (3), (*O*) oxygen or no oxygen; (4), (*I*) ignited or unignited; (5), (*T*) tinder or no tinder; (6), (*F*) fuel or no fuel; and (7), (*C*) conflagration or no conflagration. The mechanism for lighting matches is specified by the Boolean function.

(1) $I \equiv S \,\&\, D \,\&\, O$

The conflagration mechanism is specified by:

(2) $C \equiv I \,\&\, O \,\&\, T \,\&\, F$

The exogenous variables are *S*, *D*, *O*, *T* and *F*, yielding the obvious causal ordering:



The analysis goes through exactly as in the case of continuous variables, and contraposition creates no problems. The same difficulties as before — but only these difficulties — surround the identification of the individual mechanisms.

10. Transition. We now conclude our analysis of causal ordering. It has been shown how such an ordering can be defined formally in a complete structure, and how it is equivalent to identifying the mechanisms of a system. We have explored several ways in which such identification might be accomplished: through prior assumptions about exogenous variables or interventions in the static case, or by employing a standard canonical form of the equations in the dynamic case. Finally, we showed how a causal ordering in a static structure can be identified by deriving the static structure as the short-run equilibrium of an associated dynamic structure. The next task is to show how this explication of causal ordering can be used to provide a basis for the analysis of causal counterfactual conditionals.

11. The Approach to Counterfactuals Through Modal Categorization. The problem of causal counterfactual conditionals can most effectively be formulated in the framework provided by the concept of ***belief-contravening hypotheses***, that is, suppositions or assumptions which stand in *logical* conflict with accepted beliefs or known facts.[4] Consider, for the sake of illustration, one of the standard examples from the literature of the subject: "If the match had been struck, it would have ignited." The example takes us back to the case discussed in the previous section on "Discrete Variables." The case is as follows:

Accepted Facts: ($\sim S_0$) The match is (in fact) not struck
(D_0) The match is dry
(O_0) Oxygen is present
($\sim I_0$) The match does not (in fact) ignite

Accepted Law: (L) Ignition occurs (in the relevant circumstances) if and only if a dry match is struck in an oxygen-containing medium:

$$I \equiv S \,\&\, D \,\&\, O$$

The counterfactual conditional in question elicits the ostensible consequences of the belief-contravening supposition:

Assume: (S_0) The match is struck

From the standpoint of the abstract logic of the situation, order (i.e., consistency) can be restored in this situation in various ways. We must of course give up the first of the accepted facts, i.e., ($\sim S_0$). But even then inconsistency remains. We might of course give up the law (L), but this course is obviously undesirable. Retaining (L), three alternatives remain if we wish to maintain the maximum of the accepted facts.

Alternative 1		*Alternative* 2		*Alternative* 3	
Retain	***Reject***	***Retain***	***Reject***	***Retain***	***Reject***
D_0	$\sim I_0$	D_0	O_0	O_0	D_0
O_0		$\sim I_0$		$\sim I_0$	

The choice, in other words, is between three candidates for rejection, namely $\sim I_0$, O_0, and D_0. A criterion for selecting one of these for rejection is thus needed to vindicate the plausible counterfactual "If the match had been struck, it would have ignited" over against its implausible competitors

"If the match had been struck it would not have been dry"

and

"If the match had been struck, oxygen would not have been present"

The development of such a ***principle of rejection and acceptance*** in the face of a belief-contravening hypothesis lies at the heart of the problem of counterfactual conditionals. Note that we have already in effect set up a partial criterion of this kind in our aforementioned determination not to regard the law at issue as candidate for rejection in cases such as that now at issue. In general terms, the procedure to be followed is to sort the various propositions in question into ***modal categories*** M_0, M_1, M_2,..., M_n, devised subject to the conception that the lower the characteristic modal category of the proposition, the less its susceptibility to rejection, and correspondingly the greater the modal index, the greater its susceptibility to rejection.[5]

[4] This approach to counterfactuals is developed in considerable detail in [4].

[5] In the present context this summary presentation of the matter will prove sufficient. For a more detailed development of the ideas see [4].

The leading idea of the line of approach to counterfactuals we are concerned to develop here can now be codified in the following:

Procedure: To use the causal ordering of parameters through laws as a basis for allocating the propositions about the facts at issue to modal categories. And specifically to adhere to the: *Rule* — The further down a parameter is in the causal ordering (in terms of its distance from the exogenous variables) the higher the modal category of the proposition about the value of this parameter.

Let us illustrate this procedure in the context of our example.

The Match Ignition Example

Here, the causal ordering of the three parameters is, it will be remembered, as follows:

$$\begin{tikzcd} O \arrow[dr] & \\ D \arrow[r] & I \\ S \arrow[ur] & \end{tikzcd}$$

Using the causal ordering as a basis for modal categorization, we see that O and D will dominate over I, so that, forced to a choice between our three alternatives, we retain D_0 and O_0 and reject $\sim I_0$, thus vindicating the plausible counterfactual "If the match had been struck, it would have ignited" over against its implausible competitors.

The Gas Law Example

Let us begin with Case I of our piston example, where the experimenter can (directly) adjust the temperature and the pressure, with the result that the causal ordering is

$$\begin{tikzcd} P \arrow[dr] & \\ & V \\ T \arrow[ur] & \end{tikzcd}$$

Assuming that in the case under consideration the actual values are P_0, T_0, and V_0. We thus have the following information-base:

L(law): (1) $PV = kT$
Facts: (2) $P = P_0$
(3) $T = T_0$
(4) $V = V_0$

We are now asked to indicate the counterfactual consequences of the (false) hypothesis: Suppose that the temperature is doubled

$T = 2T_0$

Obviously (1) is to be retained and (3) must be rejected: The question is one of a choice between (2) and (4), viz.

(A) If the temperature were doubled, then the pressure would be doubled.

(B) If the temperature were doubled, then the volume would be doubled.

Now given the causal ordering at issue, with V as the structurally dependent variable, relatively more vulnerable to rejection because of its higher modal categorization, we would reject (4), with the result of leading to an endorsement of the conditional (B).

As before, the adiabatic case is more complex. We recall that in this case, we must include the law of conservation of energy:

(2′) $E = PV + qT$

Combining this relation with the gas law, we have:

(4) $\bar{E} = (k + q)T$,

from which we see that "If the temperature were doubled..." implies "If the initital total energy were doubled...." The causal ordering, previously derived was:

$$\begin{tikzcd}
 & T \\
P \arrow[ur] \arrow[dr] & \\
 & V
\end{tikzcd}$$

Here also in the comparison between P and V, V as the structurally dependent variable is relatively more vulnerable because of its higher modal categorization. And thus again, we are led to an endorsement of (B): "If the temperature were doubled, then the volume would be doubled." However, a careful writer might think it more precise and idiomatic to say: "If the energy were increased *so as to double the temperature*, the volume would also be doubled."

This example suggests that, in careful English usage, the counterfactual is meant to imply that its conditional stands in a relation of causal precedence to its consequence. As a final example, we shall return to the wheat growing relations to see whether this generalization is warranted.

The Wheat Growing Example

We follow exactly the same procedure as with the previous examples to see what counterfactual assertions can be made about the wheat-growing relations described earlier. The causal ordering among fertilizer (F), rain (R), wheat crop (W), population (N), and wheat price (P), was:

$$\begin{tikzcd}
F \arrow[dr] & & \\
 & W \arrow[dr] & \\
R \arrow[ur] & & P \\
 & N \arrow[ur] &
\end{tikzcd}$$

Supplementing the two laws from which this ordering derives, we have the facts—the actual amount of fertilizer, rain, wheat crop, population, and wheat price last year, say. What follows from the counterfactual premise: "If the wheat crop had been smaller last year..."?

Retaining the laws, and using the causal ordering to determine which factual premises have modal precedence, we obtain readily:

> If the wheat crop had been smaller last year, the price would have been higher.

However, this counterfactual does not resolve all the contradictions, for the hypothesis of the smaller wheat crop is not consistent with the conjunction of the law determining the wheat crop as a function of rain and fertilizer, and the actual amounts of rain and fertilizer. In the usual treatment of the counterfactual, one or the other of the latter facts—the amount of rain or the amount of fertilizer—would have to yield. But both of these have smaller modal indices than the size of the wheat crop. It is reassuring, under these circumstances, that neither of the corresponding counterfactuals is idiomatic:

> If the wheat crop had been smaller last year, less fertilizer would have been applied to it.
>
> If the wheat crop had been smaller last year, there would have been less rain.

Instead, we would regard it as perfectly idiomatic to say:

> If the wheat crop had been smaller last year, there would have been either less rain or less fertilizer applied.

Even more idiomatic would be:

> (In order) for the wheat crop to have been smaller last year, there would have to have been less rain or less fertilizer.

Thus we see that distinctions are made in English that call attention to the causal ordering among the variables in a counterfactual statement. The ordinary counterfactual corresponds to the case where the cause is the antecedent, and the effect the consequent. Where the causal ordering is reversed, more elaborate locutions, with the modals "must" or "have to," are used.

12. Conclusion. These examples of the preceding section illustrate the proposed line of attack on causal counterfactual conditionals. First the counterfactual proposition at issue is articulated through the device of a belief-contravening supposition. Then the principle of rejection and retention for the resolution of the resultant conflict of logical contradiction is determined through the mechanism of a system of modal categories. Finally, resort is made to the concept of a causal ordering to serve as a guide for the modal categorization of the factual (non-law) propositions at issue.

It is among the merits of this approach that it yields perfectly "natural" results in underwriting as plausible, in all of the standard examples treated in the literature—just those counterfactuals that are consonant with our informal, presystematic intuitions on the matter. Moreover, reliance on causal orderings also solves in a natural way the problem of the otherwise unnatural consequences of the contraposition of laws.

REFERENCES

[1] Ando, *et al.*, *Essays on the Structure of Social Science Models* (Cambridge: M.I.T. Press, 1963).

[2] Angell, R. B., "A Propositional Logic with Subjunctive Conditionals," *Journal of Symbolic Logic*, vol. 27 (1962), pp. 327–343.

[3] Burks, A. W., "The Logic of Causal Propositions," *Mind*, N. S., vol. 60 (1951), pp. 363–382.

[4] Rescher, N., *Hypothetical Reasoning* (Amsterdam: North-Holland Publishing Co., 1964).

[5] Simon, Herbert A., "Causal Ordering and Identifiability," in W. C. Hood and T. C. Koopmans (eds.), *Studies in Econometric Method* (New York: Wiley, 1953), reprinted in *Models of Man* (New York: Wiley, 1957), Chap. 1.

PRINTED IN BRUGES, BELGIUM, BY THE ST CATHERINE PRESS, LTD.