AI Magazine Volume 23 Number 1 (2002) (© AAAI)

Reasoning with Cause and Effect

Judea Pearl

■ This article is an edited transcript of a lecture given at IJCAI-99, Stockholm, Sweden, on 4 August 1999. The article summarizes concepts, principles, and tools that were found useful in applications involving causal modeling. The principles are based on structural-model semantics in which functional (or counterfactual) relationships representing autonomous physical processes are the fundamental building blocks. The article presents the conceptual basis of this semantics, illustrates its application in simple problems, and discusses its ramifications to computational and cognitive problems concerning causation.

The subject of my lecture this evening is causality.[1] It is not an easy topic to speak about, but it is a fun topic to speak about. It is not easy because like religion, sex, and intelligence, causality was meant to be practiced, not analyzed. It is fun because, like religion, sex, and intelligence, emotions run high; examples are plenty; there are plenty of interesting people to talk to; and above all, the experience of watching our private thoughts magnified under the microscope of formal analysis is exhilarating.

From Hume to AI

The modern study of causation begins with the Scottish philosopher David Hume (figure 1). Hume has introduced to philosophy three revolutionary ideas that, today, are taken for granted by almost everybody, not only philosophers: First, he made a sharp distinction between analytic and empirical claims—analytic claims are the product of thoughts; empirical claims are matters of fact. Second, he classified causal claims as empirical rather than analytic. Third, he identified the source of all empirical claims with human experience, namely, sensory input.

Putting ideas 2 and 3 together has left philosophers baffled for over two centuries over two major riddles: First, what empirical evidence legitimizes a cause-effect connection? Second, what inferences can be drawn from causal information and how?

We in AI have the audacity to hope that today after two centuries of philosophical debate, we can say something useful on this topic because for us, the question of causation is not purely academic. We must build machines that make sense of what goes on in their environment so they can recover when things do not turn out exactly as expected. Likewise, we must build machines that understand causal talk when we have the time to teach them what we know about the world because the way we communicate about the world is through this strange language called *causation*.

This pressure to build machines that both learn about, and reason with, cause and effect, something that David Hume did not experience, now casts new light on the riddles of causation, colored with an engineering flavor: How should a robot acquire causal information from the environment? How should a robot process causal information received from its creator-programmer?

I do not touch on the first riddle because David Heckerman (1999) covered this topic earlier, both eloquently and comprehensively. I want to discuss primarily the second problem—how we go from facts coupled with causal premises to conclusions that we could not obtain from either component alone.

On the surface, the second problem sounds

Figure 1. David Hume, 1711–1776.

trivial: One can simply take the causal rules, apply them to the facts, and derive the conclusions by standard logical deduction. However, it is not as trivial as it sounds. The exercise of drawing the proper conclusions from causal input has met with traumatic experiences in AI. One of my favorite examples is described in the following hypothetical man-machine dialogue:

Input:
1. If the grass is wet, then it rained.
2. If we break this bottle, the grass will get wet.
Output:
If we break this bottle, then it rained.

Another troublesome example (Lin 1995) is illustrated in the following dialogue:

Input:
1. A suitcase will open iff both locks are open.
2. The right lock is open.
3. The suitcase is closed.
Query:
What if we open the left lock?
Output:
The right lock might get closed.

In these two examples, the strange output is derived from solid logical principles, chaining in the first, constraint satisfaction in the second, yet we feel that there is a missing ingredient that the computer did not quite grasp, something to do with causality. Evidently there is some valuable information conveyed by causal vocabulary that is essential for correct understanding of the input. What is this information, and what is that magic logic that should permit a computer to select the right information, and what is the semantics behind such logic? It is this sort of question that I would like to address in this talk because I know that many in this audience are dealing with such questions and have made promising proposals for answering them. Most notable are people working in qualitative physics, troubleshooting, planning under uncertainty, modeling behavior of physical systems, constructing theories of action and change, and perhaps even those working in natural language understanding because our language is loaded with causal expressions. Since 1990, I have examined many (though not all) of these proposals, together with others that have been suggested by philosophers and economists, and I have extracted from them a set of basic principles that I would like to share with you tonight.

The Basic Principles

I am now convinced that the entire story of causality unfolds from just three basic principles: (1) causation encodes behavior under interventions, (2) interventions are surgeries on mechanisms, and (3) mechanisms are stable functional relationships.

The central theme is to view causality as a computational scheme devised to facilitate prediction of the effects of actions. I use the term *intervention* here, instead of action, to emphasize that the role of causality can best be understood if we view actions as external entities, originating from outside our theory not as a mode of behavior within the theory.

To understand the three principles it is better to start from the end and go backwards: (3) The world is modeled as an assembly of stable mechanisms, or physical laws, that are sufficient for determining all events that are of interest to the modeler. The mechanisms are autonomous, like mechanical linkages in a machine or logic gates in electronic circuits—we can change one without changing the others. (2) Interventions always involve the breakdown of mechanisms. I call this breakdown a *surgery* to emphasize its dual painful-remedial character. (1) Causal relationships tell us which mechanism is to be surgically modified by any given action.

Causal Models

These principles can be encapsulated neatly and organized in a mathematical object called a *causal model*. In general, the purpose of a model is to assign truth values to sentences in a given language. If models in standard logic assign truth values to logical formulas, causal models embrace a wider class of sentences, including phrases that we normally classify as "causal." These include:

Actions:
B will be true if we do A.
Counterfactuals:
B would be different if A were true.
Explanation:
B occurred because of A.

There could be more, but I concentrate on these three because they are commonly used and because I believe that all other causal sentences can be reduced to these three. The difference between actions and counterfactuals is merely that in counterfactuals, the clash between the antecedent and the current state of affairs is explicit.

To allay any fear that a causal model is some formidable mathematical object, let me exemplify the beast with two familiar examples. Figure 2 shows a causal model we all remember from high school—a circuit diagram. There are four interesting points to notice in this example:

First, it qualifies as a causal model because it contains the information to confirm or refute all action, counterfactual, and explanatory sentences concerning the operation of the circuit. For example, anyone can figure out what the output would be if we set Y to zero, if we replace a certain OR gate with a NOR gate, or if we perform any of the billions of combinations of such actions.

Second, Boolean formulas are insufficient for answering such action queries. For example, the Boolean formula associated with the circuit $\{Y = \mathrm{OR}(X, Z), W = NOT(Y)\}$ is equivalent to the formula associated with $\{W = NOR(X, Z), Y = NOT(W)\}$. However, setting Y to zero in the first circuit entails $W = 1$; not so in the second circuit, where Y has no effect on W.

Third, the actions about which we can reason, given the circuit, were not specified in advance; they do not have special names, and they do not show up in the diagram. In fact, the great majority of the action queries that this circuit can answer have never been considered by the designer of this circuit.

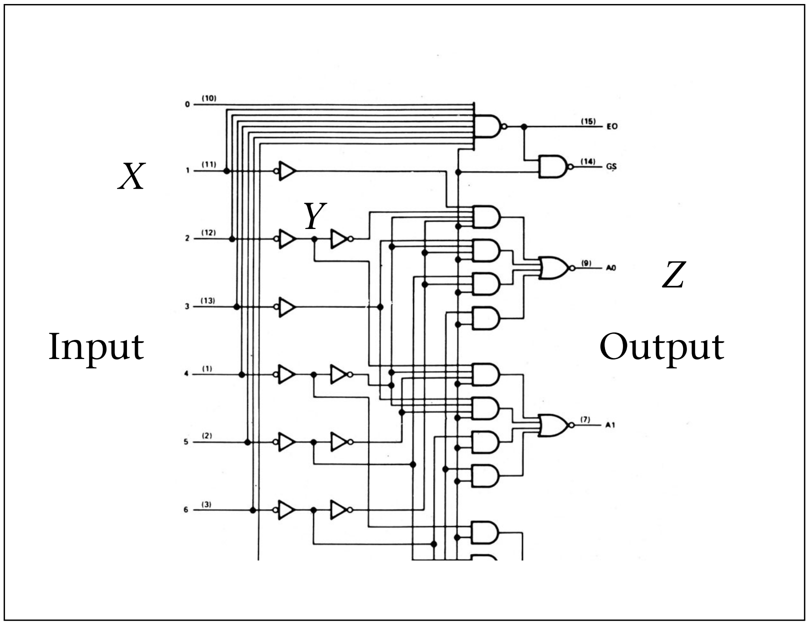


Figure 2. Causal Models: Why They Are Needed.

Fourth, how does the circuit convey this extra information? It is done through two encoding tricks: (1) The symbolic units correspond to stable physical mechanisms (that is, the logical gates). Second, each variable has precisely one mechanism that determines its value. In our example, the electric voltage in each junction of the circuit is the output of one and only one logic gate (otherwise, a value conflict can arise under certain input combinations).

As another example, figure 3 displays the first causal model that was put down on paper: Sewal Wright's (1921) path diagram. It shows how the fur pattern of the guinea pigs in a litter is determined by various genetic and environmental factors.

Again, (1) it qualifies as a causal model, (2) the algebraic equations in themselves do not qualify as a causal model, and (3) the extra information comes from having each variable determined by a stable functional relationship connecting it to its parents in the diagram. (As in $O = eD + hH + eE$, where O stands for the percentage of black area on the guinea pig fur.)

Now that we are on familiar ground, let us observe more closely the way a causal model encodes the information needed for answering causal queries. Instead of a formal definition that is given shortly (definition 1), I illustrate the working of a causal model through a vivid example. It describes a tense moment in the life of a gentleman facing a firing squad (figure 4). The captain (C) awaits the court order (U);

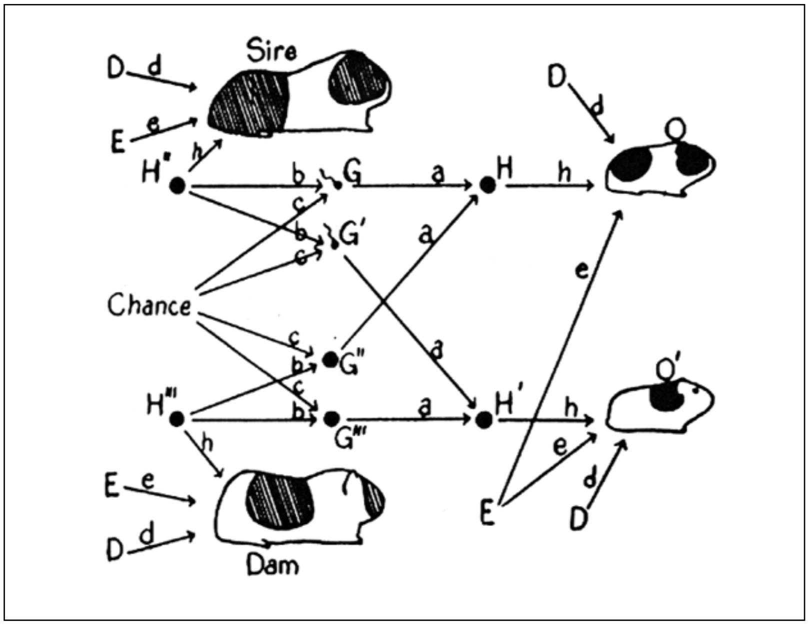


Figure 3. Genetic Models (S. Wright, 1920).

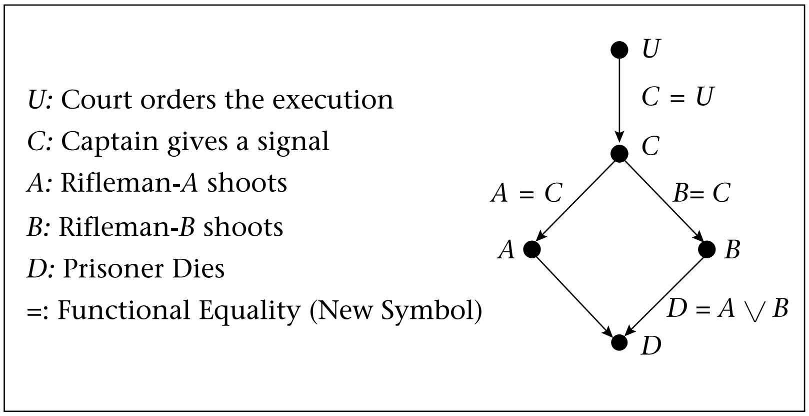

U: Court orders the execution
C: Captain gives a signal
A: Rifleman-A shoots
B: Rifleman-B shoots
D: Prisoner Dies
=: Functional Equality (New Symbol)
U
C = U
C
A = C
B= C
A
B
D = A ∨ B
D

Figure 4. Causal Models at Work (the Impatient Firing Squad).

riflemen A and B obey the captain's signal; the prisoner dies iff any of the riflemen shoot. The meaning of the symbols is obvious from the story: the only new symbol is the functional equality =, which is borrowed here from Euler (around 1730), meaning that the left-hand side is determined by the right-hand side and not the other way around.

Answering Queries with Causal Models

Assume that we want to evaluate (mechanically) the following sentences:

S1(Prediction)—If rifleman A did not shoot, then the prisoner is alive:
$\neg A \Rightarrow \neg D$.

S2(Abduction)—If the prisoner is alive, then the captain did not signal:
$\neg D \Rightarrow \neg C$.

S3(Transduction)—If rifleman A shot, then B shot as well:
$A \Rightarrow B$.

S4(Action)—If the captain gave no signal, and rifleman A decides to shoot, then the prisoner will die, and B will not shoot:
$\neg C \Rightarrow D_A$ and $\neg B_A$.

S5(Counterfactual)—If the prisoner is dead, then the prisoner would still be dead even if rifleman A had not shot:
$D \Rightarrow D_{\neg A}$.

S6(Explanation)— The prisoner died because rifleman A shot.
Caused(A, D).

The simplest sentences are S1 to S3, which can be evaluated by standard deduction; they involve standard logical connectives because they deal with inferences from beliefs to beliefs about a static world.

Next in difficulty is action sentence S4, requiring some causal information; then comes counterfactual S5, requiring more detailed causal information; and the hardest is the explanation sentence (S6) whose semantics is still not completely settled (to be discussed later).

Sentence S4 offers us the first chance to witness what information a causal model provides on top of a logical model.

Shooting with no signal constitutes a blatant violation of one mechanism in the story: rifleman A's commitment to follow the captain's signal. Violation renders this mechanism inactive; hence, we must excise the corresponding equation from the model, using a surgeon's knife, and replace it with a new mechanism: A = TRUE (figure 5). To indicate that the consequent part of our query (the prisoner's death, D) is to be evaluated in a modified model, with A = TRUE overriding $A = C$, we use the subscript notation D_A. Now we can easily verify that D holds true in the modified model; hence, S4 evaluates to true. Note that the surgery suppresses abduction; from seeing A shoot, we can infer that B shot as well (recall $A \Rightarrow B$), but from making A shoot, we can no longer infer what B does, unless we know whether the captain signaled.

Everything we do with graphs, we can, of course, do with symbols. We need to be careful, however, in distinguishing facts from rules (domain constraints) and mark the privileged element in each rule (the left-hand side), as in figure 6.

Here we see for the first time the role of causal order: Which mechanism should be excised by an action whose direct effect is A?

(Note that the symbol A appears in two equations.) The answer is, Excise the equation in which A is the privileged variable ($A = C$ in figure 6). Once we create the mutilated model MA, we draw the conclusions by standard deduction and easily confirm the truth of S4: The prisoner will be dead, denoted DA, because D is true in MA.

Consider now our counterfactual sentence S5: If the prisoner is dead, he would still be dead if A were not to have shot, $D \Rightarrow D_{\neg A}$. The antecedent $\neg A$ should still be treated as interventional surgery but only after we fully account for the evidence given: D.

This calls for three inferential steps, as shown in figure 7: (1) *abduction:* interpret the past in light of the evidence; (2) *action*: bend the course of history (minimally) to account for the hypothetical antecedent ($\neg A$); and (3) *prediction:* project the consequences to the future.

Note that this three-step procedure can be combined into one. If we use an asterisk to distinguish postmodification from premodification variables, we can combine M and M_A into one logical theory and prove the validity of S5 by purely logical deduction in the combined theory. To illustrate, we write S5 as $D \Rightarrow D^*_{\neg A^*}$ (thus, if D is true in the actual world, then D would also be true in the hypothetical world created by the modification $\neg A^*$) and prove the validity of D^* in the combined theory, as shown in figure 8.

Suppose now that we are not entirely ignorant of U but can assess the degree of belief $P(u)$. The same three steps apply to the computation of the counterfactual probability (that the prisoner be dead if A were not to have shot). The only difference is that we now use the evidence to update $P(u)$ into $P(u \mid D)$ and draw probabilistic, instead of logical, conclusions (figure 9).

Graphically, the combined theories of figure 8 can be represented by two graphs sharing the U variables (called *twin network*) (figure 10). The twin network is particularly useful in probabilistic calculations because we can simply propagate evidence (using Bayes's network techniques) from the actual to the hypothetical network.

Let us now summarize the formal elements involved in this causal exercise.

S4. (action):

If the captain gave no signal and Mr. A decides to shoot, the prisoner will die:

$\neg C \Rightarrow D_A$

and B will not shoot:

$\neg C \Rightarrow \neg B_A$

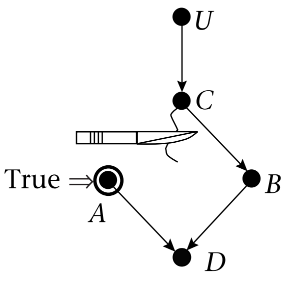
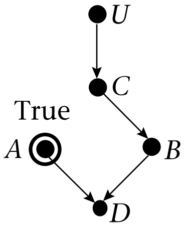


Figure 5. Why Causal Models? Guide for Surgery.

Model M_A (Modify $A = C$):

$\cancel{A = C}$

	(U)
$C = U$	(C)
A	(A)
$B = C$	(B)
$D = A \vee B$	(D)



Facts: $\neg C$

Conclusions: $A, D, \neg B, \neg U, \neg C$

S4. (action): If the captain gave no signal and A decides to shoot, the prisoner will die and B will not shoot, $\neg C \Rightarrow D_A$ & $\neg B_A$

Figure 6. Mutilation in Symbolic Causal Models.

S5. If the prisoner is dead, he would still be dead if A were not to have shot. $D \Rightarrow D_{\neg A}$

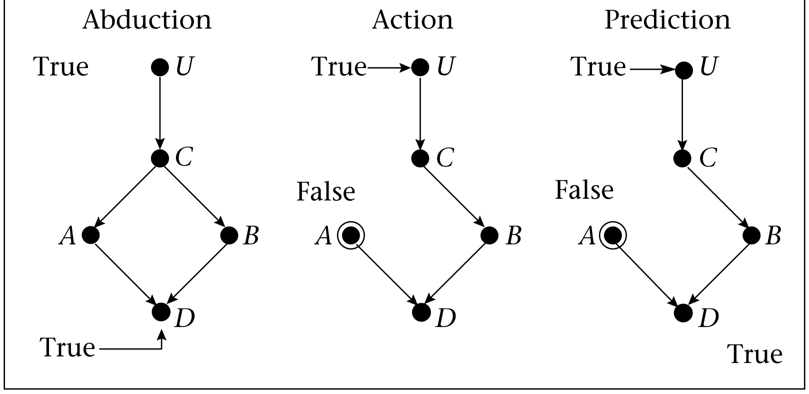


Figure 7. Three Steps to Computing Counterfactuals.

Definition 1: Causal Models

A causal model is a 3-tuple

$\langle M = V, U, F \rangle$

with a mutilation operator $do(x)$: $M \to M_x$, where

(1) $V = \{V_1, \ldots, V_n\}$ endogenous variables,

(2) $U = \{U_1, \ldots, U_m\}$ background variables

(3) F = set of n functions, f_i: $V \setminus V_i \times U \to V_i$, each of the form $v_i = f_i(pa_i, u_i)$ $PA_i \subseteq V \setminus V_i$ U_i U

(4) $M_x = \langle U, V, F_x \rangle$, $X \subseteq V$, $x \in X$, where $Fx = \{f_i$: $V_i \notin X\} \cup \{X = x\}$

Prove: $D \Rightarrow D_{\neg A}$

Combined Theory:

		(U)
$C^* = U$	$C = U$	(C)
$\neg A^*$	$A = C$	(A)
$B^* = C^*$	$B = C$	(B)
$D^* = A^* \vee B^*$	$D = A \vee B$	(D)

Facts: D

Conclusions: $U, A, B, C, D, \neg A^*, C^*, B^*, D^*$

Figure 8. Symbolic Evaluation of Counterfactuals.

P(S5). The prisoner is dead. How likely is it that he would be dead if A were not to have shot. $P(D_{\neg A} \mid D) = ?$

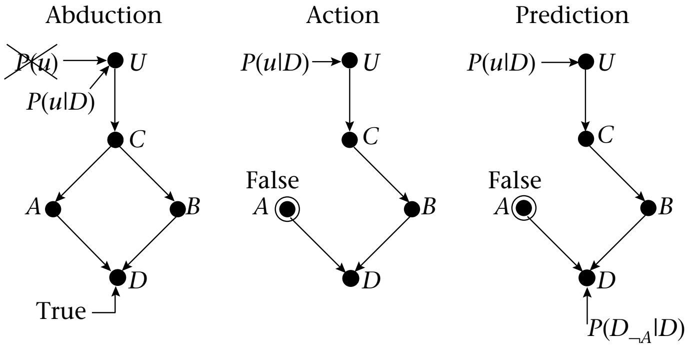
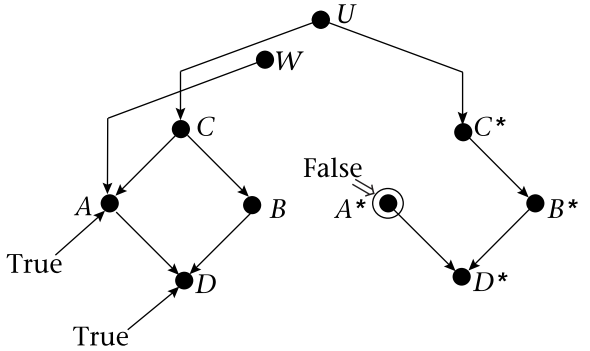


Figure 9. Computing Probabilities of Counterfactuals.



P(Alive had A not shot | A shot, Dead) =

$P(\neg D)$ in model $\langle M_{\neg A}, P(u, w \mid A, D)\rangle$ =

$P(\neg D^* \mid D)$ in twin-network

Figure 10. Probability of Counterfactuals: The Twin Network.

(Replace all functions f_i corresponding to X with the constant functions $X = x$).

Definition 2: Counterfactual

The sentence, Y would be y had X been x," denoted $Y_x(u) = y$, is interpreted to mean "the solution for Y in M_x is equal to y under the current conditions $U = u$."

The Role of Causality

We have seen that action queries can be answered in one step: standard deduction on a mutilated submodel. Counterfactual queries, however, required a preparatory stage of abduction. The questions naturally arise: Who needs counterfactuals? Why spend time on computing such convoluted sentences? It turns out that counterfactuals are commonplace, and abduction-free action sentences are a fiction. Action queries are brought into focus by certain undesired observations, potentially modifiable by the actions. The step of abduction, which is characteristic of counterfactual queries, cannot be disposed of and must therefore precede the surgery step. Thus, most action queries are semantically identical to counterfactual queries.

Consider an example from troubleshooting, where we observe a low output and we ask:

Action query:
Will the output get higher if we replace the transistor?

Counterfactual query:
Would the output be higher had the transistor been replaced?

The two sentences in this example are equivalent: Both demand an abductive step to account for the observation, which complicates things a bit. In probabilistic analysis, a functional specification is needed; conditional probabilities alone are not sufficient for answering observation-triggered action queries. In symbolic analysis, abnormalities must be explicated in functional details; the catch-all symbol $\neg ab(p)$ (standing for "p is not abnormal") is not sufficient.

Thus, we come to the million dollar question: Why causality?

To this point, we have discussed actions, counterfactuals, surgeries, mechanism, abduction, and so on, but is causality really necessary? Indeed, if we know which mechanisms each action modifies, and the nature of the modification, we can avoid all talk of causation—the ramification of each action can be obtained by simply mutilating the appropriate mechanisms, then simulating the natural course of events. The price we pay is that we need to specify an action not by its direct effects but, rather, by the mechanisms that the action modifies.

For example, instead of saying "this action moves the coffee cup to location *x,*" I would need to say, "This action neutralizes the static friction of the coffee cup and replaces it with a forward acceleration *a* for a period of 1 second, followed by deceleration for a period of 2 seconds...."

This statement is awfully clumsy: Most mechanisms do not have names in nontechnical languages, and when they do, the names do not match the granularity of ordinary language. Causality enables us to reason correctly about actions while we keep the mechanism implicit. All we need to specify is the action's direct effects; the rest follows by mutilation and simulation.

However, to figure out which mechanism deserves mutilation, there must be one-to-one correspondence between variables and mechanisms. Is this a realistic requirement? In general, no. An arbitrary collection of *n* equations on *n* variables would not normally enjoy this property. Even a typical resistive network (for example, a voltage divider) does not enjoy it. But because causal thinking is so pervasive in our language, we can conclude that our conceptualization of the world is more structured and that it does enjoy the one-to-one correspondence. We say "raise taxes," "clean your face," "make him laugh," or, in general, *do*(*p*), and miraculously, people understand us without asking for a mechanism name.[2]

Perhaps the best "AI proof" of the ubiquity of the modality *do*(*p*) is the existence of the language STRIPS (Fikes and Nilsson 1971), in which actions are specified by their effects—the ADD-LIST and DELETE-LIST. Let us compare causal surgeries to STRIPS surgeries. Both accept actions as modalities, both perform surgeries, but STRIPS performs the surgery on propositions, and causal theories first identify the mechanism to be excised and then perform the surgery on mechanisms, not on propositions. The result is that direct effects suffice, and indirect ramifications of an action need not be specified; they can be inferred from the direct effects using the mutilate-simulate cycle of figure 5.

Applications

Thus, we arrive at the midpoint of our story. I have talked about the story of causation from Hume to robotics, I have discussed the semantics of causal utterances and the principles behind the interpretation of action and counterfactual sentences, and now it is time to ask about the applications of these principles. I talk about two types of applications: The first relates to the evaluation of actions and the second to finding explanations.

Inferring Effects of Actions

Let us start with the evaluation of actions. We saw that if we have a causal model *M*, then predicting the ramifications of an action is trivial—mutilate and solve. If instead of a complete model we only have a probabilistic model, it is again trivial: We mutilate and propagate probabilities in the resultant causal network. The important point is that we can specify knowledge using causal vocabulary and can handle actions that are specified as modalities.

However, what if we do not have even a probabilistic model? This is where data come in. In certain applications, we are lucky to have data that can supplement missing fragments of the model, and the question is whether the data available are sufficient for computing the effect of actions.

Let us illustrate this possibility in a simple example taken from economics (figure 11). Economic policies are made in a manner similar to the way actions were taken in the firing squad story: Viewed from the outside, they are taken in response to economic indicators or political pressure, but viewed from the policy maker's perspective, the policies are decided under the pretense of free will.

Like rifleman *A* in figure 5, the policy maker considers the ramifications of nonroutine actions that do not conform to the dictates of the model, as shown in the mutilated graph of figure 11. If we knew the model, there would be no problem calculating the ramifications of each pending decision—mutilate and predict—but being ignorant of the functional relationships among the variables, and having only the skeleton of the causal graph in our hands, we hope to supplement this information with what we can learn from economic data.

Unfortunately, economic data are taken under a "wholesome" model, with tax levels responding to economic conditions, and we need to predict ramifications under a mutilated model, with all links influencing tax levels removed. Can we still extract useful information from such data? The answer is yes. As long as we can measure every variable that is a common cause of two or more other measured variables, it is possible to infer the probabilities of the mutilated model directly from those of the nonmutilated model, regardless of the underlying functions. The transformation is given by

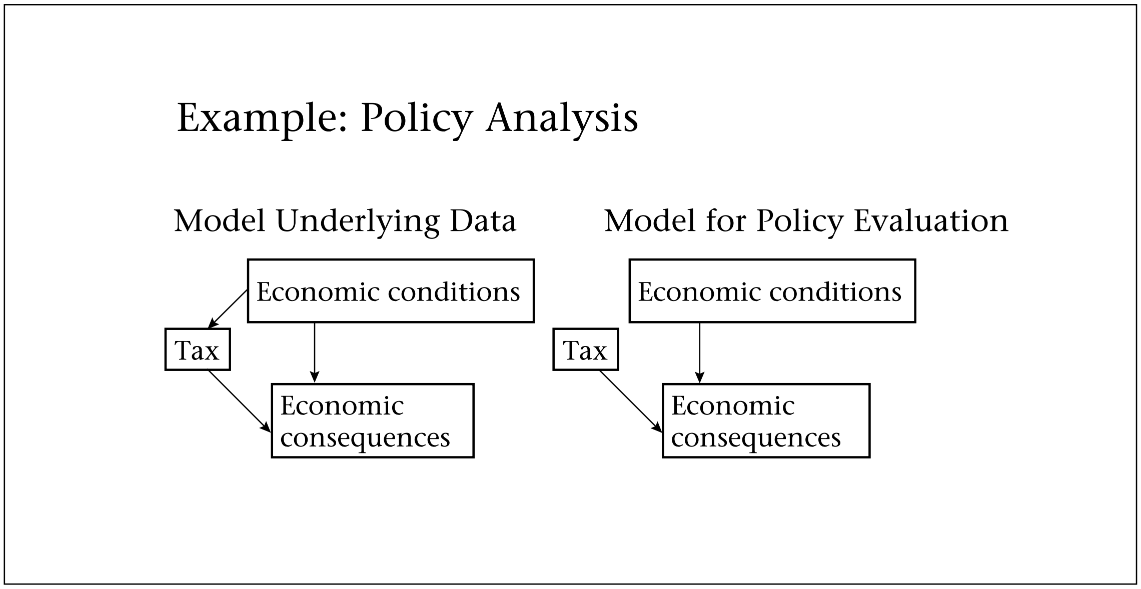


Figure 11. Intervention as Surgery.

the manipulation theorem described in the book by Spirtes, Glymour, and Schienes (1993) and further developed by Pearl (2000, 1995).

Remarkably, the effect of certain policies can be inferred even when some common factors are not observable, as is illustrated in the next example (Pearl 2000).

Smoking and the Genotype Theory

In 1964, the surgeon general issued a report linking cigarette smoking to death, cancer, and most particularly, lung cancer. The report was based on nonexperimental studies in which a strong correlation was found between smoking and lung cancer, and the claim was that the correlation found is causal; namely, if we ban smoking, the rate of cancer cases will be roughly the same as the rate we find today among nonsmokers in the population (model A, figure 12). These studies came under severe attacks from the tobacco industry, backed by some very prominent statisticians, among them Sir Ronald Fisher. The claim was that the observed correlations can also be explained by a model in which there is no causal connection between smoking and lung cancer. Instead, an unobserved genotype might exist that simultaneously causes cancer and produces an inborn craving for nicotine (see model B, figure 12).

Formally, this claim would be written in our notation as

$$
\begin{aligned}
&P(cancer \mid do(smoke)) \\
&\quad = P(cancer \mid do(not\_smoke)) \\
&\quad = P(cancer)
\end{aligned}
$$

stating that making the population smoke or stop smoking would have no effect on the rate of cancer cases. Controlled experiment could decide between the two models, but these are impossible, and now also illegal, to conduct.

This is all history. Now we enter a hypothetical era where representatives of both sides decide to meet and iron out their differences. The tobacco industry concedes that there might be some weak causal link between smoking and cancer, and representatives of the health group concede that there might be some weak links to genetic factors. Accordingly, they draw this combined model (model C in figure 12), and the question boils down to assessing, from the data, the strengths of the various links. In mutilation language, the question boils down to assessing the effect of smoking in the mutilated model shown here (model D) from data taken under the wholesome model shown before (model C). They submit the query to a statistician, and the answer comes back immediately: impossible. The statistician means that there is no way to estimate the strength for the causal links from the data because any data whatsoever can perfectly fit either one of the extreme models shown in model A and model B; so, they give up and decide to continue the political battle as usual.

Before parting, a suggestion comes up: Perhaps we can resolve our differences if we measure some auxiliary factors. For example, because the causal link model is based on the understanding that smoking affects lung cancer through the accumulation of tar deposits in the lungs, perhaps we can measure the amount of tar deposits in the lungs of sampled individuals, which might provide the necessary information for quantifying the links? Both sides agree that this suggestion is reasonable, so they submit a new query to the statistician: Can we find the effect of smoking on cancer assuming that an intermediate measurement of tar deposits is available?

Sure enough, the statistician comes back with good news: It is computable! In other words, it is possible now to infer the effect of smoking in the mutilated model shown here (model B) from data taken under the original wholesome model (model C). This inference is valid as long as the data contain measurements of all three variables: smoking, tar, and cancer. Moreover, the solution can be derived in closed mathematical form, using symbolic manipulations that mimic logical derivation but are governed by the surgery semantics. How can this derivation be accomplished?

Causal Calculus

To reduce an expression involving *do*(*x*) to those involving ordinary probabilities, we need a calculus for *doing,* a calculus that enables us to deduce behavior under intervention from behavior under passive observations. Do we have such a calculus?

If we look at the history of science, we find to our astonishment that such a calculus does not in fact exist. It is true that science rests on two components: one consisting of passive observations (epitomized by astronomy) and the other consisting of deliberate intervention (represented by engineering and craftsmanship). However, algebra was not equally fair to these two components. Mathematical techniques were developed exclusively to support the former (seeing), not the latter (doing). Even in the laboratory, a place where the two components combine, the seeing part enjoys the benefits of algebra, whereas the doing part is at the mercy of the scientist's judgment. True, when a chemist pours the content of one test tube into another, a new set of equations becomes applicable, and algebraic techniques can be used to solve the new equations. However, there is no algebraic operation to represent the transfer from one test tube to another and no algebra for selecting the correct set of equations when conditions change. Such selection has thus far relied on unaided scientific judgment.

Let me convince you of this imbalance using a simple example. If we want to find the chance it rained, given that we see wet grass, we can express our question in a formal sentence, *P*(*rain* | *wet*) and use the machinery of probability theory to transform the sentence into other expressions that are more convenient or informative. However, suppose we ask a different question: What is the chance it rained if we make the grass wet? We cannot even express our query in the syntax of probability because the vertical bar is already taken to mean "given that we see." We know intuitively what the answer should be—*P*(*rain* | *do*(*wet*)) = *P*(*rain*)—because making the grass wet does not change the chance of rain. However, can this intuitive answer, and others like it, be derived mechanically to comfort our thoughts when intuition fails?

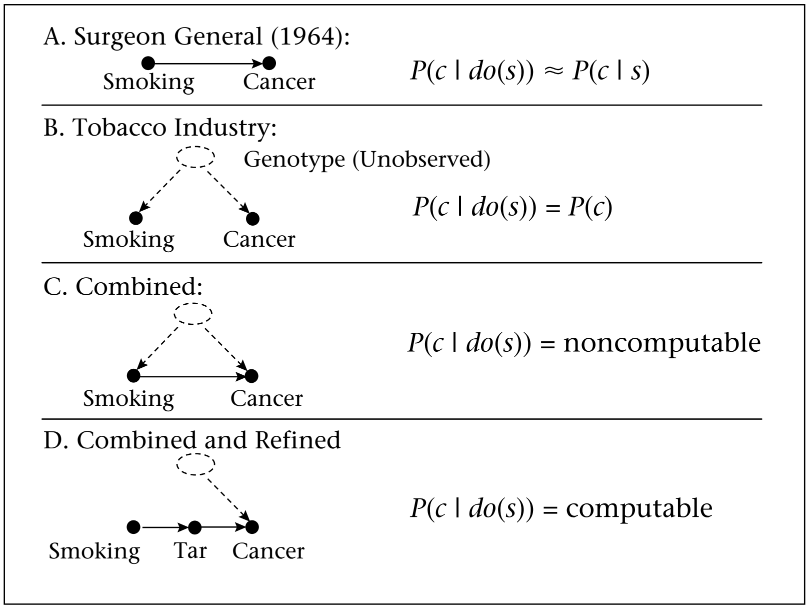


Figure 12. Predicting the Effects of Banning Cigarette Smoking.

A. Model proposed by the surgeon general. B. Model proposed by the tobacco industry. C. Combined model. D. Mutilated combined model, with one additional observation.

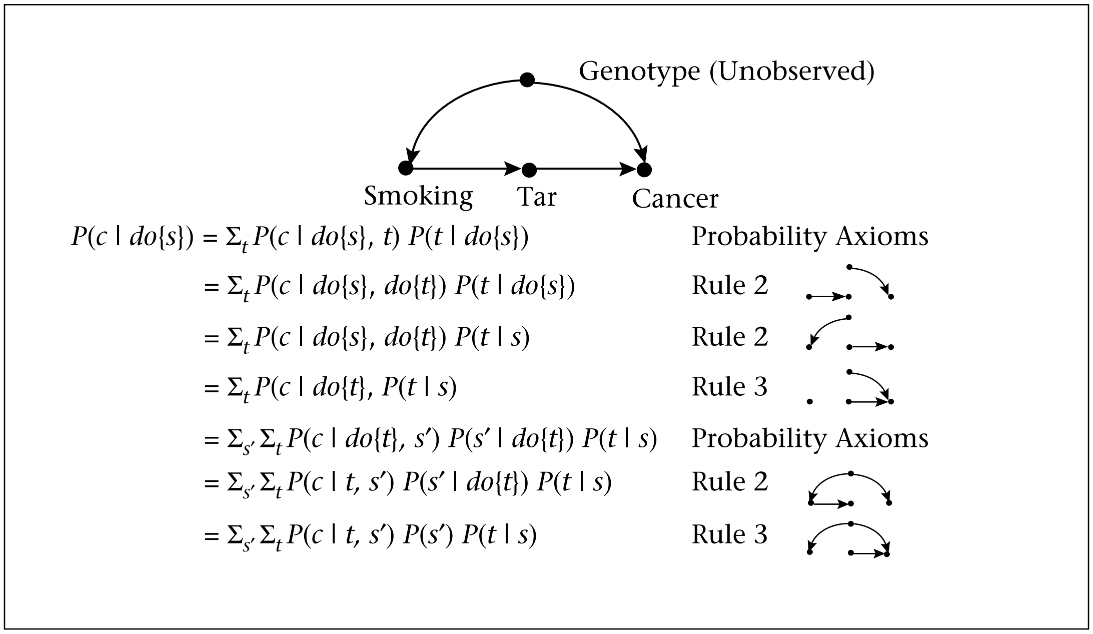
The answer is yes, but it takes a new algebra to manage the *do*(*x*) operator. To make it into a genuine calculus, we need to translate the surgery semantics into rules of inference, and these rules consist of the following (Pearl 1995):

Rule 1: Ignoring observations

$$P\big(y \big| do(x), z, w\big) = P\big(y \big| do(x), w\big) \text{ if } \big(\mathrm{Y} \perp\!\!\!\perp Z \big| X, W\big)_{G_{\overline{X}}}$$

Rule 2: Action-observation exchange

$$P\big(y \big| do(x), do(z), w\big) = P\big(y \big| do(x), z, w\big) \text{ if } (Y \perp\!\!\!\perp Z \big| X, W)_{G_{\overline{X}\underline{Z}}}$$



$$
\begin{aligned}
P(c \mid do\{s\}) &= \Sigma_t P(c \mid do\{s\}, t)\ P(t \mid do\{s\}) && \text{Probability Axioms}\\
&= \Sigma_t P(c \mid do\{s\}, do\{t\})\ P(t \mid do\{s\}) && \text{Rule 2}\\
&= \Sigma_t P(c \mid do\{s\}, do\{t\})\ P(t \mid s) && \text{Rule 2}\\
&= \Sigma_t P(c \mid do\{t\}, P(t \mid s) && \text{Rule 3}\\
&= \Sigma_{s'} \Sigma_t P(c \mid do\{t\}, s')\ P(s' \mid do\{t\})\ P(t \mid s) && \text{Probability Axioms}\\
&= \Sigma_{s'} \Sigma_t P(c \mid t, s')\ P(s' \mid do\{t\})\ P(t \mid s) && \text{Rule 2}\\
&= \Sigma_{s'} \Sigma_t P(c \mid t, s')\ P(s')\ P(t \mid s) && \text{Rule 3}
\end{aligned}
$$

Figure 13. Derivation in Causal Calculus.

Rule 3: Ignoring actions

$$
\begin{aligned}
&P(y \mid do(x), do(z), w)\\
&= P\big(y \mid do(x), w\big) \text{ if } \big(Y \perp\!\!\!\perp Z \mid X, W\big)_{G_{\overline{X, Z(W)}}}
\end{aligned}
$$

These three rules permit us to transform expressions involving actions and observations into other expressions of this type. Rule 1 allows us to ignore an irrelevant observation, the third to ignore an irrelevant action, the second to exchange an action with an observation of the same fact. The if statements on the right are separation conditions in various subgraphs of the diagram that indicate when the transformation is legal.[3] We next see them in action in the smoking-cancer example that was discussed earlier.

Figure 13 shows how one can prove that the effect of smoking on cancer can be determined from data on three variables: (1) smoking, (2) tar, and (3) cancer.

The question boils down to computing the expression *P*(*cancer* | *do*(*smoke*)) from nonexperimental data, namely, expressions involving no actions. Therefore, we need to eliminate the *do* symbol from the initial expression. The elimination proceeds like an ordinary solution of an algebraic equation—in each stage, a new rule is applied, licensed by some subgraph of the diagram, until eventually we achieve a formula involving no *do* symbols, meaning an expression computable from nonexperimental data.

If I were not a modest person, I would say that this result is amazing. Behold, we are not given any information whatsoever on the hidden genotype: It might be continuous or discrete, unidimensional or multidimensional, yet measuring an auxiliary variable (for example, tar) someplace else in the system enables us to discount the effect of this hidden genotype and predict what the world would be like if policy makers were to ban cigarette smoking.

Thus, data from the visible allow us to account for the invisible. Moreover, using the same technique, a person can even answer such intricate and personal questions as, "I am about to start smoking—should I?"

I think this trick is amazing because I cannot do such calculations in my head. It demonstrates the immense power of having formal semantics and symbolic machinery in an area that many respectable scientists have surrendered to unaided judgment.

Learning to Act by Watching Other Actors

The common theme in the past two examples was the need to predict the effect of our actions by watching the behavior of other actors (past policy makers in the case of economic decisions and past smokers-nonsmokers in the smoking-cancer example). This problem recurs in many applications, and here are a couple of additional examples.

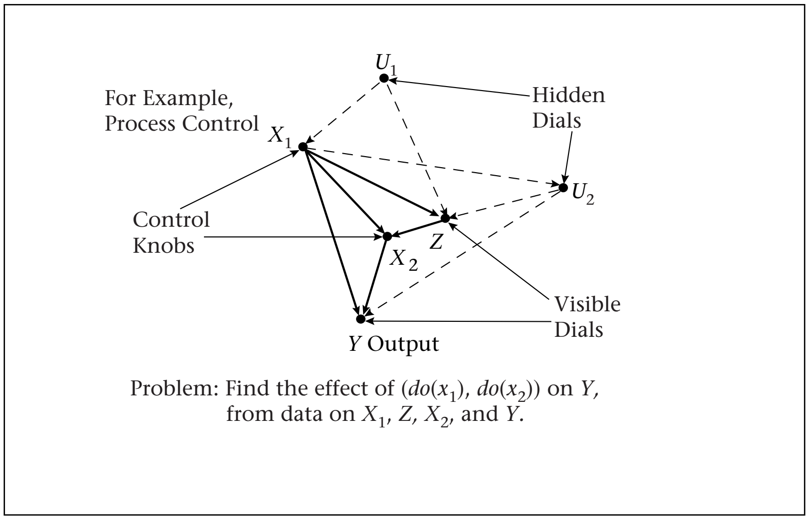
In the example of figure 14, we need to predict the effect of a plan (sequence of actions) after watching an expert control a production process. The expert observes dials that we cannot observe, although we know what quantities these dials represent.

The example in figure 15 (owed to J. Robins) comes from sequential treatment of AIDS patients.

The variables X_1 and X_2 stand for treatments that physicians prescribe to a patient at two different times: Z represents observations that the second physician consults to determine X_2, and Y represents the patient's survival. The hidden variables U_1 and U_2 represent, respectively, part of the patient history and the patient disposition to recover. Doctors used the patient's earlier PCP history (U_1) to prescribe X_1, but its value was not recorded for data analysis.

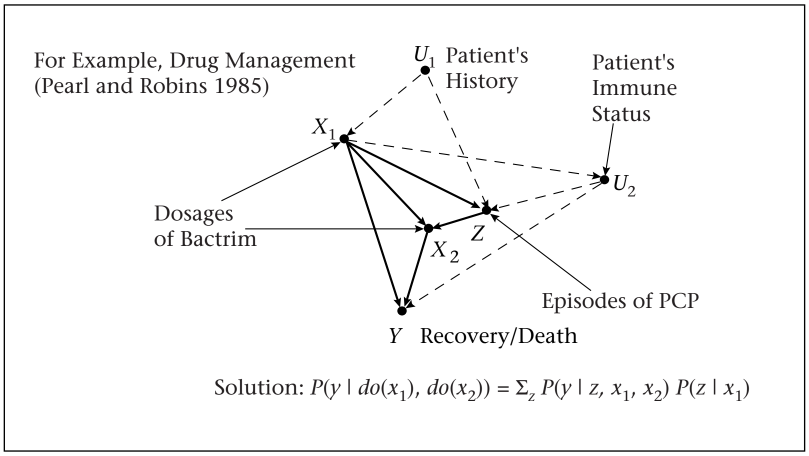
The problem we face is as follows: Assume we have collected a large amount of data on the behavior of many patients and physicians, which is summarized in the form of (an estimated) joint distribution P of the observed four variables (X_1, Z, X_2, Y). A new patient comes in, and we want to determine the impact of the plan ($do(x_1)$, $do(x_2)$) on survival (Y), where x_1 and x_2 are two predetermined dosages of bactrim, to be administered at two prespecified times.

Many of you have probably noticed the similarity of this problem to Markov decision processes, where it is required to find an optimal sequence of actions to bring about a certain response. The problem here is both simpler and harder. It is simpler because we are only required to evaluate a given strategy, and it is harder because we are not given the transition probabilities associated with the elementary actions—these need to be learned from data, and the data are confounded by spurious correlations. As you can see on the bottom line of figure 15, this task is feasible—the calculus reveals that our modeling assumptions are sufficient for estimating the desired quantity from the data.



Problem: Find the effect of ($do(x_1)$, $do(x_2)$) on Y, from data on X_1, Z, X_2, and Y.

Figure 14. Learning to Act by Watching Other Actors (Process Control).



Solution: $P(y \mid do(x_1), do(x_2)) = \Sigma_z P(y \mid z, x_1, x_2)\, P(z \mid x_1)$

Figure 15. Learning to Act by Watching Other Actors (Treatment Management).

Deciding Attribution

I now demonstrate how causal calculus can answer questions of attribution, namely, finding causes of effects rather than effects of causes. The U.S. Army conducted many nuclear experiments in Nevada in the period from 1940 to 1955. Data taken over a period of 12 years indicate that fallout radiation apparently has resulted in a high number of deaths from leukemia in children residing in southern Utah. A lawsuit was filed. Is the Army liable for these deaths?

According to a fairly common judicial standard, damage should be paid if and only if it is more probable than not that death would not have occurred but for the action. Can we calculate this probability *PN* that event *y* would not have occurred but for event *x*? (*PN* stands for probability of necessity.) The answer is yes; *PN* is given by the formula

$$PN = \frac{\left[P(y|x) - P(y'|x')\right]}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$

where *x*′ stands for the complement of *x*. However, to get this formula, we must assume a condition called *monotonicity*; that is, radiation cannot prevent leukemia. In the absence of monotonicity, the formula provides a lower bound on the probability of necessity (Pearl 2000; Tian and Pearl 2000).

This result, although it is not mentioned explicitly in any textbooks on epidemiology,

Figure 16. Causal Explanation.

"She handed me the fruit and I ate."
"The serpent deceived me, and I ate."

statistics, or law, is rather startling. First, it shows that the first term, which generations of epidemiologists took to be the formal interpretation *PN*, by the power of which millions of dollars were awarded (or denied) to plaintiffs in lawsuits, is merely a crude approximation of what lawmakers had in mind by the legal term *but for* (Greenland 1999). Second, it tells us what assumptions must be ascertained before this traditional criterion coincides with lawmakers' intent. Third, it shows us precisely how to account for confounding bias $P(y \mid x') - P(y \mid do(x'))$ or nonmonotonicity. Finally, it demonstrates (for the first time, to my knowledge) that data from both experimental and nonexperimental studies can be combined to yield information that neither study alone can provide.

AI and Peripheral Sciences

Before I go to the topic of explanation, I would like to say a few words on the role of AI in such applications as statistics, public health, and social science. One of the reasons that I find these areas to be fertile ground for trying out new ideas in causal reasoning is that unlike AI, tangible rewards can be reaped from solving relatively small problems. Problems involving barely four to five variables, which we in AI regard as toy problems, carry tremendous payoffs in public health and social science. Billions of dollars are invested each year on various public health studies: Is chocolate ice cream good for you or bad for you? Would red wine increase or decrease your heart rate? The same applies to the social sciences. Would increasing police budget decrease or increase crime rates? Is the Columbine school incident owed to TV violence or failure of public education? The Interuniversity Consortium for Political and Social Research distributed about 800 gigabytes worth of such studies in 1993 alone.

Unfortunately, the causal-analytic methodology currently available to researchers in these fields is rather primitive, and every innovation can make a tremendous difference. Moreover, the major stumbling block has not been statistical but, rather, conceptual—lack of semantics and lack of formal machinery for handling causal knowledge and causal queries—perfect for AI involvement. Indeed, several hurdles have recently been removed by techniques that emerged from AI laboratories. I predict that a quiet revolution will take place in the next decade in the way causality is handled in statistics, epidemiology, social science, economics, and business. Although news of this revolution will never make it to the Defense Advanced Research Projects Agency newsletter, and even the National Science Foundation might not be equipped to manage it, it will nevertheless have an enormous intellectual and technological impact on our society.

Causes and Explanations

We now come to one of the grand problems in AI: generating meaningful explanations. It is the hardest of all causal tasks considered thus far because the semantics of explanation is still debatable, but some promising solutions are currently in the making (see Pearl [2000] and Halpern and Pearl [2001a, 2001b]).

The art of generating explanations is as old as mankind (figure 16).

According to the Bible, it was Adam who first discovered the ubiquitous nature of causal explanation when he answered God's question with "she handed me the fruit and I ate." Eve was quick to catch on: "The serpent deceived me, and I ate." Explanations here are used for exonerating one from blame, passing on the responsibility to others. The interpretation therefore is counterfactual: "Had she not given me the fruit, I would not have eaten."

Counterfactuals and Actual Causes

The modern formulation of this concept starts again with David Hume, who stated (1748): "We may define a cause to be an object followed by another, …, where, if the first object has not been, the second never had existed."

This counterfactual definition was given a possible-world semantics by David Lewis (1973) and an even simpler semantics using our structural interpretation of definition 2. Note how we write, in surgery language, Hume's sentence: "If the first object (x) had not been, the second (y) never had existed":

(1) $X(u) = x$

(2) $Y(u) = y$

(3) $Y_{x'}(u) \neq y$ for $x' \neq x$

meaning that given situation u, the solution for Y in a model mutilated by the operator $do(X = x')$ is not equal to y.

However, this definition of *cause* is known to be ridden with problems: It ignores aspects of sufficiency, it fails in the presence of alternative causes, and it ignores the structure of intervening mechanisms. I first demonstrate these problems by examples and then provide a solution using a notion called *sustenance* that is easy to formulate in our structural-model semantics.

Let us first look at the aspect of sufficiency (or production), namely, the capacity of a cause to produce the effect in situations where the effect is absent. For example, both *match* and *oxygen* are necessary for fire, and neither is sufficient alone. Why is *match* considered an adequate explanation and *oxygen* an awkward explanation? The asymmetry surfaces when we compute the probability of sufficiency:

$P(x$ is sufficient for $y)$
$= P(Y_x = y \mid X \neq x, Y \neq y)$

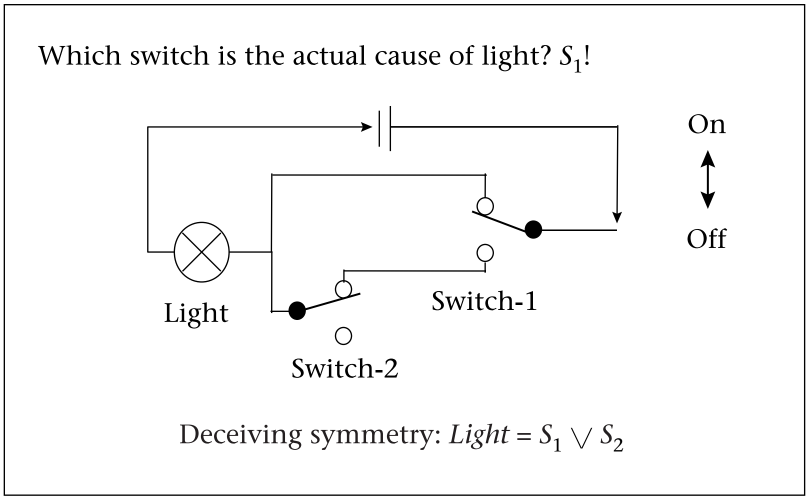
for which we obtain:

$P(oxygen$ is sufficient for $fire) = P(match) = low$

$P(match$ is sufficient for $fire) = P(oxygen)$
$= high$

Thus, we see that human judgment of explanatory adequacy takes into account not merely how necessary a factor was for the effect but also how sufficient it was.

Another manifestation of sufficiency occurs in a phenomenon known as *overdetermination*. Suppose in the firing squad example (figure 4), rifleman B is instructed to shoot if and only if rifleman A does not shoot by 12 noon. If rifleman A shoots before noon and kills, we would naturally consider his shot to be the cause of death. Why? The prisoner would have died without A's shot. The answer lies in an argument that goes as follows: We cannot exonerate an action that actually took place by appealing to hypothetical actions that might or might not materialize. If for some strange reason B's rifle gets stuck, then death would not have occurred were it not for A's shot. This argument instructs us to imagine a new world, contrary to the scenario at hand, in which some structural contingencies are introduced, and in this contingency-inflicted world, we are to perform Hume's counterfactual test. I call this type of argument *sustenance* (Pearl 2000) formulated as follows:



Which switch is the actual cause of light? S_1!
On
Off
Light
Switch-1
Switch-2
Deceiving symmetry: $Light = S_1 \vee S_2$

Figure 17. Preemption: How the Counterfactual Test Fails.

Definition 3: Sustenance

Let W be a set of variables, and let w, w' be specific realizations of these variables. We say that x causally sustains y in u relative to contingencies in W if and only if

(1) $X(u) = x$;

(2) $Y(u) = y$;

(3) $Y_{xw}(u) = y$ *for all* w; *and*

(4) $Y_{x'w'}(u) = y' \neq y$ *for some* $x' \neq x$ *and some* w'.

The last condition $Y_{x'w}(u) = y'$ weakens necessity by allowing Y to differ from y (under $x' \neq x$) under a special contingency, when W is set to some w'. However, the third condition, $Y_{xw}(u) = y$ carefully screens the set of permitted contingencies by insisting that Y retain its value y (under x) for every setting of $W = w$.

We now come to the second difficulty with the counterfactual test—its failure to incorporate structural information.

Consider the circuit in figure 17. If someone were to ask us which switch causes the light to be on in this circuit, we would point to switch 1. After all, switch 1 (S1) causes the current to flow through the light bulb, but switch 2 (S2) is totally out of the game. However, the overall functional relationship between the switches and the light is deceptively symmetric:

Light = $S1 \vee S2$

Turning switch 1 off merely redirects the current but keeps the light on, but turning

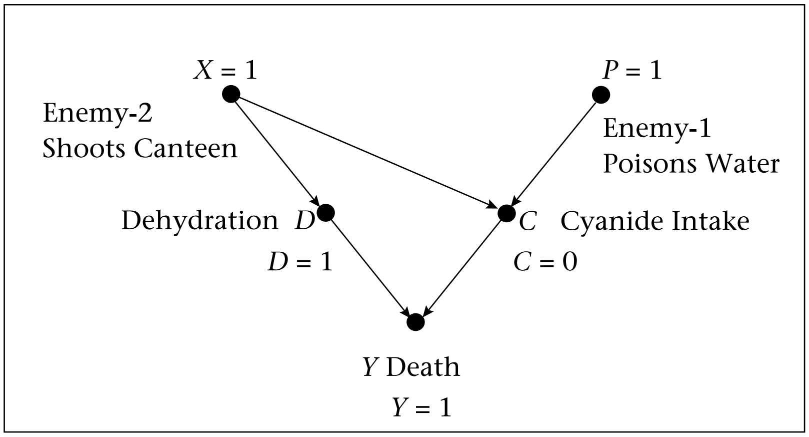


Figure 18. The Desert Traveler (The Actual Scenario).

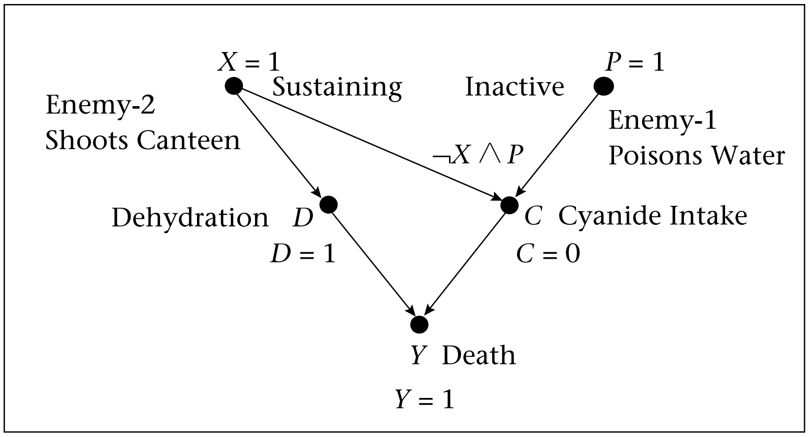


Figure 19. The Desert Traveler (Constructing a Causal Beam-1).

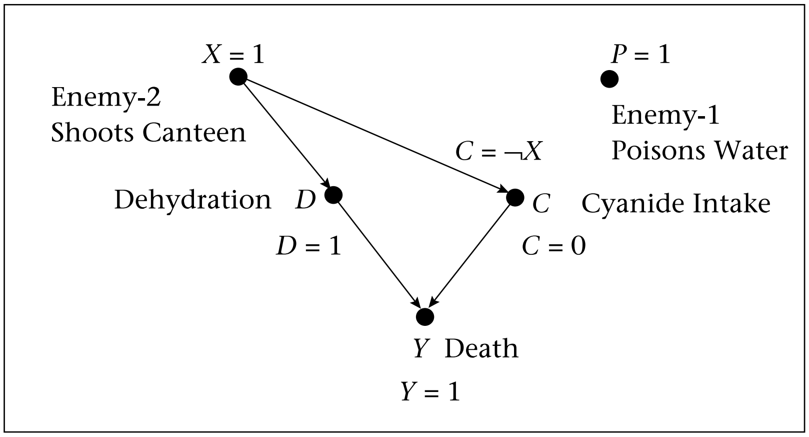


Figure 20. The Desert Traveler (Constructing a Causal Beam-2).

switch 2 off has no effect whatsoever—the light turns off if and only if both switches are off. Thus, we see that the internal structure of a process affects our perception of actual causation. Evidently, our mind takes into consideration not merely input-output relationships but also the inner structure of the process leading from causes to effects. How?

Causal Beam

The solution I would like to propose here is based on local sustenance relationships. Given a causal model, and a specific scenario in this model, we construct a new model by pruning away, from every family, all parents except those that minimally sustain the value of the child. I call the new model *a causal beam* (Pearl 2000). In this new model, we conduct the counterfactual test, and we proclaim an event $X = x$ the actual cause of $Y = y$ if y depends on x in the new model. I now demonstrate this construction using a classical example owed to P. Suppes (figure 18). It is isomorphic to the two-switch problem but more bloodthirsty.

A desert traveler T has two enemies. Enemy 1 poisons T's canteen, and enemy 2, unaware of enemy 1's action, shoots and empties the canteen. A week later, T is found dead, and the two enemies confess to action and intention. A jury must decide whose action was the actual cause of T's death. Enemy 1 claims T died of thirst, and enemy 2 claims to have only prolonged T's life.

Now let us construct the causal beam associated with the natural scenario in which we have death ($Y = 1$), dehydration ($D = 1$), and no poisoning ($C = 0$). Consider the C-family (figure 19).

Because emptying the canteen is sufficient for sustaining no cyanide intake, regardless of poisoning, we label the link $P \rightarrow C$ *inactive* and the link $X \rightarrow C$ *sustaining*. The link $P \rightarrow C$ is inactive in the current scenario, which allows us to retain just one parent of C, with the functional relationship $C = \neg X$ (figure 20).

Next, consider the Y-family (in the situation $D = 1$, $C = 0$) (figure 21). Because dehydration would sustain death regardless of cyanide intake, we label the link $C \rightarrow Y$ *inactive* and the link $D \rightarrow Y$ *sustaining*.

We drop the link $C \rightarrow Y$, and we end up with a causal beam leading from shooting to death through dehydration. In this final model we conduct the counterfactual test and find that the test is satisfied because $Y = X$. Thus, we have the license to classify the shooter as the cause of death, not the poisoner, though none meets the counterfactual test for necessity on a global scale—the asymmetry emanates from structural information.

Things will change of course if we do not know whether the traveler craved for water before or after the shot. Our uncertainty can be modeled by introducing a background variable, U, to represent the time when the traveler first reached for drink (figure 22).

If the canteen was emptied before T needed to drink, we have the dehydration scenario, as before (figure 21). However, if T drank before the canteen was emptied, we have a new causal beam in which enemy 1 is classified as the cause of death. If U is uncertain, we can use $P(u)$ to compute the probability $P(x$ caused $y)$ because the sentence "x was the actual cause of y" receives a definite truth value for every value u of U. Thus,

$$P(x \text{ caused } y) = \sum_{\{u \mid x \text{ caused } y \text{ in } u\}} P(u)$$

Temporal Preemption

We come now to resolve a third objection against the counterfactual test—temporal preemption. Consider two fires advancing toward a house. If fire 1 burned the house before fire 2, we (and many juries nationwide) would consider fire 1 the actual cause of the damage, although fire 2 would have done the same if it were not for fire 1. If we simply write the structural model as

$$H = F1 \vee F2,$$

where H stands for "house burns down," the beam method would classify each fire equally as a contributory cause, which is counterintuitive. Here, the second cause becomes ineffective only because the effect has already happened—a temporal notion that cannot be expressed in the static causal model we have used thus far. Remarkably, the idea of a causal beam still gives us the correct result if we use a dynamic model of the story, as shown in figure 23.

Dynamic structural equations are obtained when we index variables by time and ask for the mechanisms that determine their values. For example, we can designate by $S(x, t)$ the state of the fire in location x and time t and describe each variable $S(x, t)$ as dependent on three other variables: (1) the previous state of the adjacent region to the north, (2) the previous state of the adjacent region to the south, and (3) the previous state at the same location (figure 24).

To test which fire was the cause of the damage, we simulate the two actions at their corresponding times and locations and compute the scenario that unfolds from these actions. Applying the process equations recursively, from left to right, simulates the propagation of the two fires and gives us the actual value for each variable in this spatiotemporal domain. In figure 24, white represents unconsumed regions, black represents regions on fire, and grey represents burned regions.

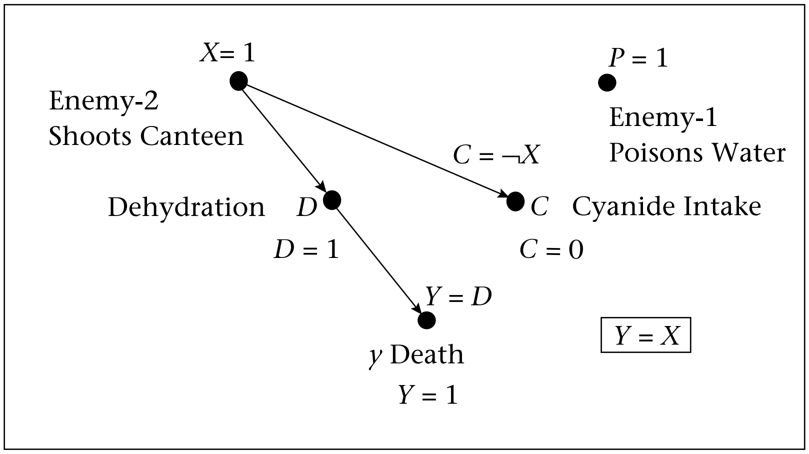


Figure 21. The Desert Traveler (The Final Beam).

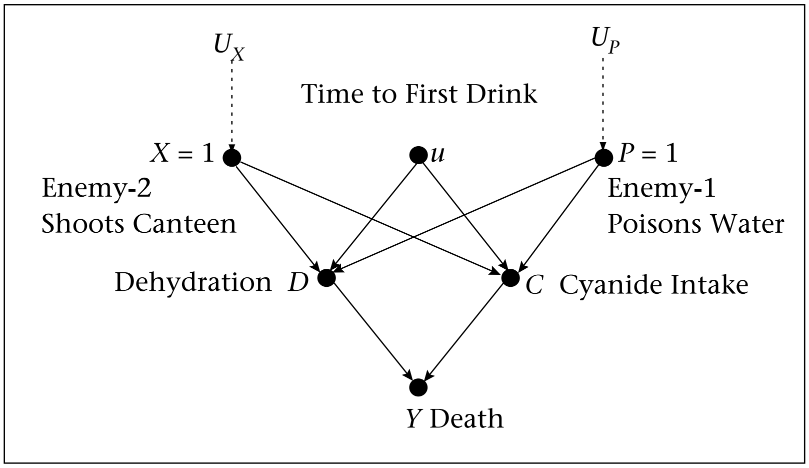


Figure 22. The Enigmatic Desert Traveler (Uncertain Scenario).

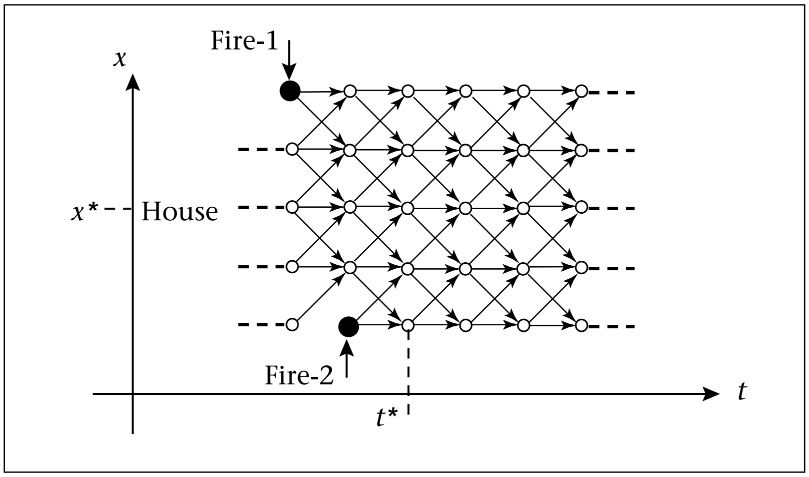


Figure 23. Dynamic Model under Action: do*(Fire-1),* do*(Fire-2).*

We are now ready to construct the beam and conduct the test for causation. The resulting beam is unique and is shown in figure 25.

The symmetry is clearly broken—there is a dependence between fire 1 and the state of the

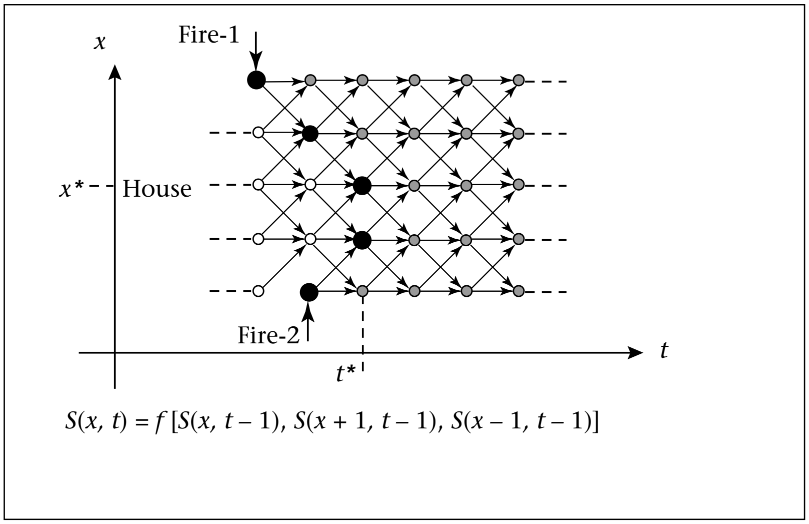


Figure 24. The Resulting Scenario.

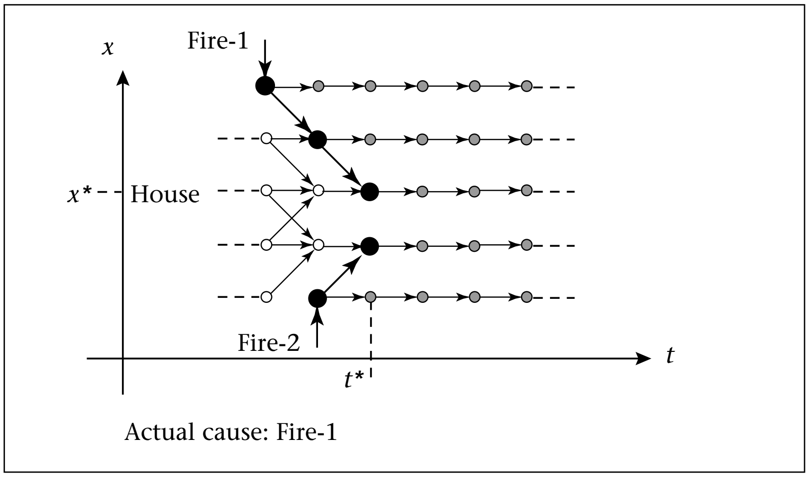


Figure 25. The Dynamic Beam.

house (in location x^*) at all times $t \geq t^*$; no such dependence exists for fire 2. Thus, the earlier fire is proclaimed the actual cause of the house burning.

Conclusions

I would like to conclude this article by quoting two great scientists. The first is Democritus (460–370 B.C.), the father of the atomic theory of matter, who said: "I would rather discover one causal relation than be King of Persia." Although the political situation in Persia has changed somewhat from the time he made this statement, I believe Democritus had a valid point in reminding us of the many application areas that could benefit from the discovery of even one causal relation, namely, from the solution of one toy problem, on an AI scale. As I discussed earlier, these applications include medicine, biology, economics, and social science, and I believe AI is in a unique position to help these areas because only AI enjoys the combined strength of model searching, learning, and automated reasoning within the logic of causation.

The second quote is from Albert Einstein who, a year before his death, attributed the progress of Western science to two fortunate events: "Development of Western science is based on two great achievements: The invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)."

I have tried to convince you that experimental science has not fully benefited from the power of formal methods—formal mathematics was used primarily for analyzing passive observations under fixed boundary conditions, but the choice of actions, the design of new experiments, and the transitions between boundary conditions have been managed by the unaided human intellect. The development of autonomous agents and intelligent robots requires a new type of analysis in which the doing component of science enjoys the benefit of formal mathematics side by side with its observational component. A short glimpse at these benefits was presented here. I am convinced that the meeting of these two components will eventually bring about another scientific revolution, perhaps equal in impact to the one that took place during the Renaissance. AI will be the major player in this revolution, and I hope each of you take part in seeing it off the ground.

Acknowledgments

I want to acknowledge the contributions of a wonderful team of colleagues and students with whom I was fortunate to collaborate: Alex Balke, David Chickering, Adnan Darwiche, Rina Dechter, Hector Geffner, David Galles, Moises Goldszmidt, Sander Greenland, Joseph Halpern, David Heckerman, Jin Kim, Jamie Robins, and Tom Verma. Most of these names should be familiar to you from other stages and other shows, except perhaps Greenland and Robins, two epidemiologists who are carrying the banner of causal analysis in epidemiology (Greenland et al. 1999).

I have borrowed many ideas from other authors, the most influential ones are listed in the references, but others are cited in Pearl (2000). I only mention that the fundamental idea that actions be conceived as modifiers of mechanisms goes back to Jacob Marschak and

Herbert Simon. Strotz and Wold (1960) were the first to represent actions by "wiping out" equations, and I would never have taken seriously the writings of these "early" economists if it were not for Peter Spirtes's lecture,[4] where I first learned about manipulations and manipulated graphs.

Craig Boutilier deserves much credit for turning my unruly lecture into a readable article.

Notes

[1] The lecture was delivered as part of the IJCAI Research Excellence Award for 1999. Color slides and the original transcript can be viewed at bayes.cs.ucla.edu/IJCAI99/. Detailed technical discussion of this material can be found in Pearl (2000).

[2] H. Simon (1953) devised a test for deciding when a one-to-one correspondence exists; see Pearl (2000).

[3] The notation is defined in Pearl (2000). For example,

$$\left(Y \perp\!\!\!\perp Z \middle| X\right)_{G_{\overline{X}\,\underline{Z}}}$$

states that in the subgraph formed by deleting all arrows entering X and leaving Z, the nodes of X d-separate those of Y from those of Z.

[4] Given at the International Congress of Philosophy of Science, Uppsala, Sweden, 1991.

Bibliography

Druzdzel, M. J., and Simon, H. A. 1993. Causality in Bayesian Belief Networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence,* eds. D. Heckerman and A. Mamdani, 3–11. San Francisco, Calif.: Morgan Kaufmann.

Fikes, R. E., and Nilsson, N. J. 1971. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence* 2(3–4): 189–208.

Good, J. 1961. A Causal Calculus, I. *British Journal for the Philosophy of Science* 11:305–318.

Greenland, S. 1999. Relation of the Probability of Causation to the Relative Risk and the Doubling Dose: A Methodologic Error That Has Become a Social Problem. *American Journal of Public Health* 89:1166–1169.

Greenland, S.; Pearl, J.; and Robins, J. M. 1999. Causal Diagrams for Epidemiologic Research. *Epidemiology* 10(1): 37–48.

Haavelmo, T. 1943. The Statistical Implications of a System of Simultaneous Equations. *Econometrica* 11:1–12. Reprinted in 1995. *The Foundations of Econometric Analysis,* eds. D. F. Hendry and M. S. Morgan, 477–490. Cambridge, U.K.: Cambridge University Press.

Hall, N. 2002. Two Concepts of Causation. In *Causation and Counterfactuals,* eds. J. Collins, N. Hall, and L. A. Paul. Cambridge, Mass.: MIT Press. Forthcoming.

Halpern, J. Y. 1998. Axiomatizing Causal Reasoning. In *Uncertainty in Artificial Intelligence,* eds. G. F. Cooper and S. Moral, 202–210. San Francisco, Calif.: Morgan Kaufmann.

Halpern, J. Y., and Pearl, J. 2001a. Causes and Explanations: A Structural-Model Approach—Part 1: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* 194–202. San Francisco, Calif.: Morgan Kaufmann.

Halpern, J. Y., and Pearl, J. 2001b. Causes and Explanations: A Structural-Model Approach—Part 2: Explanations. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 27–34. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Heckerman, D. 1999. Learning Bayesian Networks. Invited Talk presented at the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), 3 August, Stockholm, Sweden.

Heckerman, D., and Shachter, R. 1995. Decision-Theoretic Foundations for Causal Reasoning. *Journal of Artificial Intelligence Research* 3:405–430.

Hume, D. 1748 (rep. 1988). *An Enquiry Concerning Human Understanding*. Chicago: Open Court.

Lewis, D. 1973. *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Lin, F. 1995. Embracing Causality in Specifying the Indirect Effects of Actions. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1985–1991. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Michie, D. 1999. Adapting Good's Theory to the Causation of Individual Events. In *Machine Intelligence 15,* eds. D. Michie, K. Furukawa, and S. Muggleton, 60–86. Oxford, U.K.: Oxford University Press.

Nayak, P. P. 1994. Causal Approximations. *Artificial Intelligence* 70(1–2): 277–334.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearl, J. 1995. Causal Diagrams for Empirical Research. *Biometrika* 82(4): 669–710.

Reiter, R. 1987. A Theory of Diagnosis from First Principles. *Artificial Intelligence* 32(1): 57–95.

Shoham, Y. 1988. *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence.* Cambridge, Mass.: MIT Press.

Simon, H. A. 1953. Causal Ordering and Identifiability. In *Studies in Econometric Method,* eds. W. C. Hood and T. C. Koopmans, 49–74. New York: Wiley.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. Causation, Prediction, and Search. New York: Springer-Verlag.

Strotz, R. H., and Wold, H. O. A. 1960. Recursive versus Nonrecursive Systems: An Attempt at Synthesis. *Econometrica* 28(2): 417–427.

Tian, J., and Pearl, J. 2000. Probabilities of Causation: Bounds and Identification. *Annals of Mathematics and Artificial Intelligence* 28:287–313.

Wright, S. 1921. Correlation and Causation. *Journal of Agricultural Research* 20:557–585.

Wright, S. 1920. The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea Pigs. *Proceedings of the National Academy of Science* 6:320–332.

Judea Pearl received a B.S. in electrical engineering from the Technion Haifa Israel in 1960, an M.S. in physics from Rutgers University, and a Ph.D. in electrical engineering, from the Polytechnic Institute of Brooklyn. He joined the faculty at the University of California at Los Angeles in 1970 and is currently director of the Cognitive Systems Laboratory, where he conducts research in knowledge representation, probabilistic and causal reasoning, constraint processing, and learning. A member of the National Academy of Engineering, a Fellow of the Institute of Electrical and Electronics Engineers, and a Fellow of the American Association for Artificial Intelligence, Pearl is a recipient of the IJCAI Research Excellence Award (1999), the AAAI Classic Paper Award (2000), and the Lakatos Award of Outstanding Contribution to the Philosophy of Science (2001).

Call for Proposals:

2003 AAAI Spring Symposium Series

March 24-26, 2003

Sponsored by the American Association for Artificial Intelligence

AAAI invites proposals for the 2003 Spring Symposium Series, to be held March 24-26, 2003 at Stanford University, California.

The Spring Symposium Series is an annual set of meetings run in parallel at a common site. It is designed to bring colleagues together in an intimate forum while at the same time providing a significant gathering point for the AI community. The two and a half day format of the series allows participants to devote considerably more time to feedback and discussion than typical one-day workshops. It is an ideal venue for bringing together new communities in emerging fields.

The symposia are intended to encourage presentation of speculative work and work in progress, as well as completed work. Ample time should be scheduled for discussion. Novel programming, including the use of target problems, open-format panels, working groups, or breakout sessions, is encouraged. Working notes will be prepared, and distributed to the participants. At the discretion of the individual symposium chairs, these working notes may also be made available as AAAI Technical Reports following the meeting. Most participants of the symposia will be selected on the basis of statements of interest or abstracts submitted to the symposia chairs; some open registration will be allowed. All symposia are limited in size, and participants will be expected to attend a single symposium.

Proposals for symposia should be between two and five pages in length, and should contain:

A title for the symposium.

A description of the symposium, identifying specific areas of interest, and, optionally, general symposium format.

The names and addresses (physical and electronic) of the organizing committee: preferably three or more people at different sites, all of whom have agreed to serve on the committee.

A list of potential participants that have been contacted and that have expressed interest in participating. A common way of gathering potential participants is to send email messages to email lists related to the topic(s)of the symposium. Note that potential participants need not commit to participating, only state that they are interested.

Ideally, the entire organizing committee should collaborate in producing the proposal. If possible, a draft proposal should be sent out to a few of the potential participants and their comments solicited.

Approximately eight symposia on a broad range of topics within and around AI will be selected for the 2003 Spring Symposium Series. All proposals will be reviewed by the AAAI Symposium Committee, chaired by Holly Yanco, University of Massachusetts Lowell. The criteria for acceptance of proposals include:

Perceived interest to the AAAI community. Although AAAI encourages symposia that cross disciplinary boundaries, a symposium must be of interest to some subcommunity of the AAAI membership. Symposia that are of interest to a broad range of AAAI members are also preferred.

Appropriate number of potential participants. Although the series supports a range of symposium sizes, the target size is around 40-60 participants.

Lack of a long-term ongoing series of activities on the topic. The Spring Symposium Series is intended to nurture emerging communities and topics, so topics that already have yearly conferences or workshops are inappropriate.

An appropriate organizing committee. The organizing committee should have (1) good technical knowledge of the topic, (2) good organizational skills, and (3) connections to the various communities from which they intend to draw participants. Committees for cross-disciplinary symposia must adequately represent all the disciplines to be covered by the symposium.

Accepted proposals will be distributed as widely as possible over the subfields of AI, and balanced between theoretical and applied topics. Symposia bridging theory and practice and those combining AI and related fields are particularly solicited.

Symposium proposals should be submitted as soon as possible, but no later than April 22, 2002. Proposals that are submitted significantly before this deadline can be in draft form. Comments on how to improve and complete the proposal will be returned to the submitter in time for revisions to be made before the deadline. Notifications of acceptance or rejection will be sent to submitters around May 6, 2002. The submitters of accepted proposals will become the chair of the symposium, unless alternative arrangements are made. The symposium organizing committees will be responsible for:

Producing, in conjunction with the general chair, a Call for Participation and Registration Brochure for the symposium, which will be distributed to the AAAI membership

Additional publicity of the symposium, especially to potential audiences from outside the AAAI community

Reviewing requests to participate in the symposium and determining symposium participants

Preparing working notes for the symposium

Scheduling the activities of the symposium

Preparing a short review of the symposium, to be printed in *AI Magazine*

AAAI will provide logistical support, will take care of all local arrangements, and will arrange for reproducing and distributing the working notes. Please submit (preferably by electronic mail) your symposium proposals, and inquiries concerning symposia, to:

Holly Yanco
Computer Science Department
University of Massachusetts Lowell
Olsen Hall, 1 University Avenue
Lowell, MA 01854
Voice: 978-934-3642
Fax: 978-934-3551
E-mail: holly@cs.uml.edu