

REASONING WITH CAUSE AND EFFECT

Judea Pearl

Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles 90095
judea@cs.ucla.edu

Abstract

This paper summarizes basic concepts and principles that I have found to be useful in dealing with causal reasoning. The paper is written as a companion to a lecture under the same title, to be presented at IJCAI-99, and is intended to supplement the lecture with technical details and pointers to more elaborate discussions in the literature.

The ruling conception will be to treat causation as a computational schema devised to identify the invariant relationships in the environment, so as to facilitate reliable prediction of the effect of actions. This conception, as well as several of its satellite principles and tools, has been guiding paradigm for several research communities in AI, most notably those connected with causal discovery, troubleshooting, planning under uncertainty and modeling the behavior of physical systems. My hopes are to encourage a broader and more effective usage of causal modeling by explicating these common principles in simple and familiar mathematical form.

After casting the concepts of “causal model,” “actions,” and “counterfactuals” in abstract mathematical terms we will demonstrate by examples how counterfactual questions can be answered from both deterministic and probabilistic causal models (Section 2). In Subsection ??, I will argue that planning and decision making is an exercise in counterfactual reasoning, and will demonstrate this thesis in a simple example taken from econometrics. This will set the stage for our discussion in Subsection 3.1, where I discuss the empirical content of counterfactuals in terms of policy predictions. Subsection 3.2 discusses the role of counterfactuals in the interpretation and generation of causal explanations. We will end with discussions of how causal relationships emerge from actions and mechanisms (Section 3.3) and how causal directionality can be induced by a set of symmetric equations (Section 3.4).

1 Causes and Counterfactuals

In one of his most quoted sentences, David Hume (1748) ties together two aspects of causation: 1. regularity of succession and 2. counterfactual dependency:

“we may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by object similar to the second, Or , in other words, where, if the first object had not been, the second never had existed” [Hume, 1948, Section VII].

The idea of reducing causality to counterfactuals is further echoed by John Stuart Mill (1843), and has reached its fruition in the works of David Lewis (1973b, 1986). But it remains puzzling; how can convoluted expressions of the type “if the first object had not been, the second never had existed” illuminate simple commonplace expressions like “*A* caused *B*”?

Implicit in this proposal lies an intriguing claim that counterfactual expressions are less ambiguous to our mind than causal expressions. But, discerning the truth of counterfactuals requires the generation and examination of possible alternative to the actual situation, and testing whether certain propositions holds in those alternatives—a mental task of non-negligible proportions.

Hume, Mill, and Lewis apparently believed that going through this mental exercise is, nevertheless, simpler than intuiting directly on whether it was *A* that caused *B*. How? What mental representation allows humans to process counterfactuals so swiftly and reliably, and what logic governs that process so as to maintain uniform standards of coherence and plausibility?

Structure vs. similarity

According to Lewis’ account (1973b), the evaluation of counterfactuals involves the notion of *similarity*: one orders possible worlds by some measure of similarity, and the a counterfactual $A \square \rightarrow B$ (read: “*B* if it were *A*”) is declared true in a world *w* just in case *B* is true in all the closest *A*-worlds to *w*.¹

¹Related possible-world semantics were introduced in artificial intelligence to represent actions and database updates [Ginsberg, 1986; Ginsberg and Smith, 1987; Winslett, 1988; Katsuno and Mendelzon, 1991].

This semantics still leaves two questions unsettled and problematic: 1. What choice of similarity measure would make counterfactual reasoning compatible with ordinary conception of cause and effect? 2. What mental representation of worlds ordering would render the computation of counterfactuals manageable and practical (in both man and machine.)

In his initial proposal, Lewis was careful to keep his formalism as general as possible, and, save for the requirement that every world be closest to itself, he did not impose any structure on the similarity measure. However, simple observations tell us that similarity measures cannot be arbitrary. The very fact that people communicate with counterfactuals already suggests that they share a similarity measure, that this measure is encoded parsimoniously in the mind and, hence, that it must be highly structured.

2 Structural Model Semantics

This section presents a structural-model semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998).² Related approaches have been proposed in Simon and Rescher (1966) and Robins (1986).

We start with a deterministic definition of a “causal model” which consists, as we have discussed in earlier chapters, of functional relationships among variables of interest, each relationship representing an autonomous mechanism. Causal and counterfactuals relationships are defined in this model in terms of response to local modifications of those mechanisms. Probabilistic relationships emerge naturally by assigning probabilities to background conditions. After demonstrating, by examples, how this model facilitates the computation of counterfactuals (Section 2.2) and the interpretation of causal utterances in natural discourse (Section 3.1) we then present a general method of computing probabilities of counterfactual expressions using causal diagrams (Section 3.2).

2.1 Definitions: Causal models, actions and counterfactuals

In logics, however, a model is a mathematical object that assigns truth values to sentences in a given language of discourse. A *causal model*, naturally, should encode the truth values of sentences that deal with causal relationships, these include action sentences (e.g., “A will be true if we do B,” counterfactuals (e.g., “A would have been different if it were not for B”) and plain causal utterances (e.g., “A may cause B” or “B occurred because of A”). Because they convey information about the dynamics of changing worlds. Such sentences cannot be interpreted in standard propositional logic, or probability calculus causal models supplement this information,

²Similar models, called “neuron diagrams” [Lewis, 1986, p. 200; Hall, 1998] are used informally by philosophers to illustrate chains of causal processes.

by explicit encoding of the *invariant* relationships in the domain.

Definition 2.1 (Causal model)

A causal model is a triple

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

where

- (i) U is a set of background variables, (also called *exogenous*³), that are determined by factors outside the model.
- (ii) V is a set $\{V_1, V_2, \dots, V_n\}$ of variables, called *endogenous*, that are determined by variables in the model, namely, variables in $U \cup V$.
- (iii) F is a set of functions $\{f_1, f_2, \dots, f_n\}$ where each f_i is a mapping from $U \cup (V \setminus V_i)$ to V_i . In other words, each f_i tells us the value of V_i given the values of all other variables in $U \cup V$. Symbolically, the set of equations F can be represented by writing

$$v_i = f_i(pa_i, u_i) \quad i = 1, \dots, n$$

where pa_i is any realization of the unique minimal set of variables PA_i in V/V_i (connoting parents) that renders f_i nontrivial. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that renders f_i nontrivial.

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i . We call such a graph the *causal digram* associated with M . This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i .

Definition 2.2 (Submodel)

Let M be a causal model, X be a set of variables in V , and x be a particular realization of X . A submodel M_x of M is the causal model

$$M_x = \langle U, V, F_x \rangle$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \quad (1)$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Submodels are useful for representing the effect of local actions and hypothetical changes, including those dictated by counterfactual antecedents. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under

³We will try to refrain from using the term “exogenous” in referring to background conditions, because this term has required more refined technical connotations (see Sections ?? and ??).

any u , then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X . The transformation from M to M_x modifies the algebraic content of F , which is the reason for the name *modifiable structural equations* used in [Galles and Pearl, 1998].⁴

Definition 2.3 (*Effect of action*)

Let M be a causal model, X be a set of variables in V , and x be a particular realization of X . The effect of action $do(X = x)$ on M is given by the submodel M_x .

Definition 2.4 (*Potential response*)

Let Y be a variable in V , and let X be a subset of V . The potential response of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x .⁵

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form “ $do(X = x)$ if $Z = z$ ” can be formalized using the replacement of equations by functions of Z , rather than by constants [Pearl, 1994]. We will not consider disjunctive actions, of the form “ $do(X = x \text{ or } X = x')$ ”, since these complicate the probabilistic treatment of counterfactuals.

Definition 2.5 (*Counterfactual*)

Let Y be a variable in V , and let X a subset of V . The counterfactual sentence “The value that Y would have obtained, had X been x ” is interpreted as denoting the potential response $Y_x(u)$.

Definition 2.5 thus interprets the counterfactual phrase “had X been x ” in terms of a hypothetical external action that modifies the actual course of history and enforces the condition “ $X = x$ ” with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals [Balke and Pearl, 1994], as it permits x to differ from the current value of $X(u)$ without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.⁶

This formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 2.6 (*Probabilistic causal model*)

A probabilistic causal model is a pair

$$\langle M, P(u) \rangle$$

⁴Structural modifications date back to Marschak (1950) and Simon (1953). An explicit translation of interventions into “wiping out” equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995). A similar notion of sub-model is introduced in Fine (1985), though not specifically for representing actions and counterfactuals.

⁵Galles and Pearl (1998) required that F_x has a unique solution, a requirement later relaxed by Halpern (1998). Uniqueness of solution is ensured in recursive systems, i.e., where $G(M)$ is a cyclic.

⁶Simon and Rescher (1966, p. 339) did not include this step in their definition of counterfactuals and ran into difficulties with unwarranted backward inferences triggered by the antecedents.

where M is a causal model and $P(u)$ is a probability function defined over the domain of U .

$P(u)$, together with the fact that each endogenous variable is a function of U , defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \mid Y(u)=y\}} P(u) \quad (2)$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x :

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \quad (3)$$

Likewise a causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W , not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u \mid Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \quad (4)$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u). \quad (5)$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement “ Y would be y if $X = x$ and Y would be y' if $X = x'$.” Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 1997]. The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U , explains away these objections (see Section 3.1) and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F .

Of particular interest to us would be probabilities of counterfactuals, conditional on actual observations. For example, the probability that event $X = x$ “was the cause” of event $Y = y$, may be interpreted as the probability that Y would not be equal to y had X not been x , given that $X = x$ and $Y = y$ have in fact occurred (see Chapter ??, for in depth discussion of probabilities of causation). Such probabilities are well defined in the model described, and require the evaluation of expressions of the form $P(Y_{x'} = y' \mid X = x, Y = y)$ with x' and y' incompatible with x and y , respectively. Eq. (4) allows the evaluation of this quantity using a 3-step procedure that we now summarize in a theorem.

Theorem 2.7 *Given model $\langle M, P(u) \rangle$, the conditional probability $P(B_A|e)$ of a counterfactual sentence “If it were A then B ”, given evidence e , can be evaluated using the following three steps:*

1. **Abduction**—update $P(u)$ by the evidence e , to obtain $P(u|e)$.
2. **Action**—Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A .
3. **Prediction**—Use the modified model $\langle M_A, P(u|e) \rangle$ to compute the probability of B , the consequence of the counterfactual.

2.2 Evaluating counterfactuals: Deterministic analysis

We now apply Theorem 2.7 to demonstrate how counterfactual queries, both deterministic and probabilistic, can be answered formally, using structural-model semantics.

Example-1, The firing squad

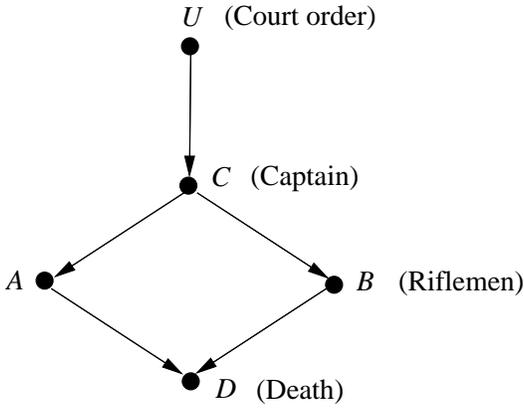


Figure 1: Causal relationships in the 2-man firing squad example.

Consider a 2-man firing squad as depicted in Fig. 1, where A, B, C, D and U stand for the following propositions:

- U = Court orders the execution
- C = Captain gives a signal
- A = Rifleman- A shoots
- B = Rifleman- B shoots
- D = Prisoner dies

Assume that the court’s decision is unknown, that both riflemen are perfectly accurate marksmen who are alert and law abiding, and that the prisoner is not likely to die from fright or other extraneous causes. We wish to construct a formal representation of the story, so that the following sentences can be evaluated mechanically:

- S1. (*prediction*): If rifleman- A did not shoot, the prisoner is alive. Formally, $\neg A \Rightarrow \neg D$
- S2. (*abduction*): If the prisoner is alive, then the Captain did not signal. Formally, $\neg D \Rightarrow \neg C$
- S3. (*transduction*): If rifleman- A shot, then B shot as well. Formally, $A \Rightarrow B$

- S4. (*action*): If the captain gave no signal and rifleman- A decides to shoot, the prisoner will die and B will not shoot. Formally, $\neg C \Rightarrow D_A \ \& \ \neg B_A$
- S5. (*counterfactual*): If the prisoner is dead, then even if A were not to have shot, the prisoner would still be dead. Formally, $D \Rightarrow D_{\neg A}$

Evaluating standard sentences

To prove the first three sentences we need not invoke causal models; these sentences involve standard logical connectives and can be handled therefore using standard logical inference. The story can be captured in any convenient logical theory, T , for example,

$$T : U \Leftrightarrow C, C \Leftrightarrow A, C \Leftrightarrow B, A \vee B \Leftrightarrow D$$

and the validity of S1-S3 can easily be verified by derivation from T .

Two remarks are worth noting before we go to analyze sentences S4 and S5. First, the two-way implications in T are necessary for supporting abduction; if we were to use one-way implications, e.g., $C \Rightarrow A$, we would not be able to conclude C from A . In standard logic, this symmetry removes all distinctions between the tasks of prediction (reasoning forward in time), abduction (reasoning from evidence to explanation) and transduction (reasoning from evidence to explanation, then from explanation to predictions). In non-standard logics (e.g., logic programming), were the implication sign dictates the direction of inference and even contraposition is not supported, meta-logical inference machinery must be invoked to perform abduction [Eshghi and Kowalski, 1989]. Second, the feature that renders S1-S3 manageable in standard logic is that they all deal with *epistemic* inference, that is, inference from beliefs to beliefs about a static world. Sentence S2, for example, can be explicated to state: if we find that the prisoner is alive, we have the license to believe that the captain did not give the signal. The material implication sign \Rightarrow in logic does not extend beyond this narrow meaning, to be contrasted next with the counterfactual implication.

Evaluating action sentences

Sentence S4 invokes a deliberate action “rifleman- A decides to shoot” and, from our representation of actions (Definition 2.3) it must violate some premises, or mechanisms, in the initial theory of the story. In order to formally identify what remains invariant to the action, we must incorporate causal relationships into the theory; logical relationships or even domain constraints alone are not sufficient. The causal model corresponding to our story is as follows:

Model M :

$$\begin{array}{ll}
 & (U) \\
 C = U & (C) \\
 A = C & (A) \\
 B = C & (B) \\
 D = A \vee B & (D)
 \end{array}$$

Here we use equality, rather than implication signs, to stress the fact that each equation represents an

autonomous mechanism, (an “integrity-constraint” in the language of databases); it remains invariant unless specifically violated. We further use parenthetical symbols next to each equation, to explicitly identify the dependent variable (on the left hand side) in the equation, thus emphasizing the causal asymmetry associated with the arrows in Fig. 1.

To evaluate S4, we follow Definition 2.3 and form the submodel M_A , in which the equation $A = C$ is replaced by A (simulating the decision of of rifleman- A to shoot regardless of signals), and obtain

Model M_A :	(U)
	$C = U \quad (C)$
	$A \quad (A)$
	$B = C \quad (B)$
	$D = A \vee B \quad (D)$
Facts: $\neg C$	
Conclusions: $A, D, \neg B, \neg U, \neg C$	

We see that, given the fact $\neg C$, we can easily deduce D and $\neg B$ and thus confirm the validity of S4.

It is important to note that “problematic” sentences like S4, whose antecedent violates one of the basic premises in the story (i.e., that both riflemen are law abiding) are handled naturally in the same deterministic setting as the story is told. Traditional approaches tend to reject sentences like S4 as contradictory, and would insist on re-formulating the problem probabilistically (see next subsection), so as to tolerate exceptions to the law $A = C$. Such reformulations are artificial and unnecessary; the structural approach permits us to draw the intended inferences in the natural, deterministic formulation of the problem.

Evaluating counterfactuals

We are now ready to evaluate the counterfactual sentence S5, using the steps of Theorem 2.7 (though no probabilities are involved). We first add the fact D to the original model, M , evaluate U , then form the submodel $M_{\neg A}$ and, finally, re-evaluate the truth of D in $M_{\neg A}$, using the value of U found in the first step. These steps are explicated as follows:

Step-1:

Model M :	(U)
	$C = U \quad (C)$
	$A = C \quad (A)$
	$B = C \quad (B)$
	$D = A \vee B \quad (D)$
Facts: D	
Conclusions: U, A, B, C, D	

Step-2:

Model $M_{\neg A}$:	(U)
	$C = U \quad (C)$
	$\neg A \quad (A)$
	$B = C \quad (B)$
	$D = A \vee B \quad (D)$

Facts: U

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

Note that it is only the value of U , the background variable, which is carried over from Step-1 to Step-2; all other propositions must be re-evaluated subject to the new modification of the model. This reflects the understanding that background factors U are not affected by either the variables or the mechanisms in the model $\{f_i\}$, hence the counterfactual consequent (in our case D) is to be evaluated under the background conditions prevailing in the actual world. In fact, the background variables are the main carriers of persistence information from the actual world to the hypothetical world.

Note also that this 2-step procedure of evaluating counterfactuals can be combined into one. If we distinguish post-modification variables from pre-modification variables by a star, we can combine M and M_x into one logical theory and prove the validity of S5 by purely logical inference in the combined theory. To illustrate, we write S5 as $D \Rightarrow D^*_{\neg A^*}$ (read: 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:

Combined Theory:

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

Facts: D

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

Note that U is not starred, reflecting the assumptions that background conditions remain unaltered.

It is worth reflecting at this point on the difference between S4 and S5. Syntactically, the two appear to be identical, as both involve a fact implying a counterfactual and, yet, we labeled S4 an “action” sentence and S5 a “counterfactual” sentence. The difference lies in the relationship between the given fact and the antecedent of the counterfactual (i.e., the “action” part). In S4 the fact given (C) is not affected by the antecedent (A) while in S4, the fact given (D) is potentially affected by the antecedent ($\neg A$). The difference between these two situations is fundamental, as can be seen from their methods of evaluation. In evaluating S4, we knew in advance that C would not be affected by the model modification $do(A)$ and, therefore, we were able to add C directly to

the modified model M_A . In evaluating S5, on the other hand, we were contemplating a possible reversal, from D to $\neg D$, due to the modification $do(\neg A)$ and, therefore, we had to first add fact D to the pre-action model M , summarize its impact via U , and reevaluate D once the modification $do(\neg A)$ takes place. Thus, although the causal effect of actions can be expressed syntactically as a counterfactual sentence, this need to route the impact of known facts through U makes counterfactuals a different species than actions (see Section ??).

We should also emphasize that most counterfactuals utterances in natural language presume knowledge of facts that are affected by the antecedent. When we say, for example, “ B would be different if it were not for A ,” we imply knowledge of what the actual value of B is and that B is susceptible to A . It is this sort of sentences that gives counterfactuals their unique character, distinct of action sentences. As we have seen in Section ??, it is this sort of sentences that would require a more detailed specifications for their evaluation.

2.3 Evaluating counterfactuals: Probabilistic analysis

Assume the following modification of the firing-squad story:

1. There is a probability $P(U) = p$ that the court will order an execution within the time period considered.
2. Rifleman- A has a probability q of pulling the trigger out of nervousness, regardless of the Captain’s signal.
3. Rifleman- A nervousness is independent of U .

With these assumptions, we wish to compute the quantity $P(\neg D_{\neg A}|D)$, namely, the probability that the prisoner would be alive if A were not to have shot, given that the prisoner is in fact dead.

Following Theorem 2.7, our first step (abduction) is to compute the posterior probability $P(u, w|D)$, accounting for the fact that the prisoner is found dead. This is easily evaluated to:

$$P(u, w|D) = \begin{cases} \frac{P(u, w)}{1 - (1-p)(1-q)} & \text{if } u = 1 \text{ or } w = 1 \\ 0 & \text{if } u = 0, w = 0 \end{cases} \quad (6)$$

The second step (action) is to form the submodel $M_{\neg A}$, while retaining the posterior probability, giving

$$\langle M_{\neg A}, P(u, w|D) \rangle: \begin{array}{ll} & (U, W) \\ C = U & (C) \\ \neg A & (A) \\ B = C & (B) \\ D = A \vee B & (D) \end{array}$$

The last step (prediction) is to compute $P(\neg D)$ in the probabilistic model above. Noting that $\neg D = \neg U$, the expected result follows:

$$P(\neg D_{\neg A}|D) = P(\neg U|D) = \frac{q(1-p)}{1 - (1-q)(1-p)}.$$

2.4 The twin-networks method

A major practical difficulty in the procedure described above is the need to compute, store and use the posterior distribution $P(u|e)$, where u stand for the set of all background variables in the model. As is illustrated in the last example, even when we start with Markovian model in which the background variables are mutually independent, conditioning on e normally destroys this independence, and makes it necessary to carry over a full description of the joint distribution of U , conditional on e ; such description may be prohibitively large.

A graphical method of overcoming this difficulty is described in Balke and Pearl (1994b), which uses two networks, one to represent the actual world, and one to represent the hypothetical world (Fig. 2).

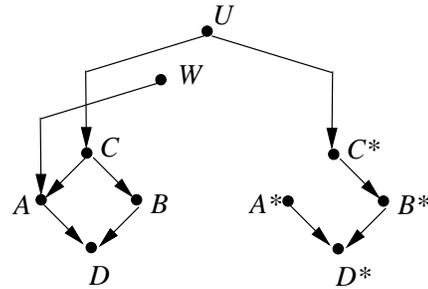


Figure 2:

The two networks are identical in structure, save for the arrows entering A , which have been deleted to mirror the equation deleted from $M_{\neg A}$. Like Siamese twins, the two networks share the background variables (in our case U and W) since those remain invariant under modification. The endogenous variables are replicated and labeled distinctly, because they may obtain different values in the hypothetical vis a vis the actual world. The task of computing $P(\neg D)$ in the model $\langle M_A, P(u, v|D) \rangle$ thus reduces to that of computing $P(\neg D^*|D)$ in the twin network shown.

In general, if we wish to compute the counterfactual probability $P(Y_x = y|z)$, where X, Y and Z are arbitrary sets of variables (not necessarily disjoint), Theorem 2.7 instructs us to compute $P(y)$ in the model $\langle M_x, P(u|z) \rangle$, which reduces to computing an ordinary conditional probability $P(y^*|z)$ in an augmented Bayesian network. Such computation can be performed by standard evidence-propagation techniques in a Bayesian network—the distribution $P(u|z)$ need not be explicated, conditional independencies can be exploited, and local computation methods can be employed such as those summarized in many textbooks (e.g., [Pearl, 1988]).

The twin-network representation also offers a useful way of testing independencies among counterfactual quantities. To illustrate, suppose we have a chain-like causal diagram, $X \rightarrow Z \rightarrow Y$, and we wish to test whether Y_x is independent of X given Z , that is,

$Y_x \perp\!\!\!\perp X|Z$. The twin-network associated with this chain is shown in Fig. 3.

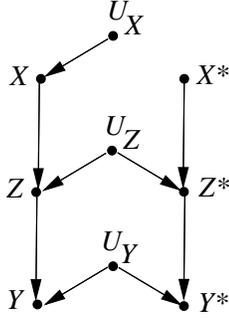


Figure 3: Twin-network representation of the counterfactual Y_x in the model $X \rightarrow Z \rightarrow Y$.

To test whether $Y_x \perp\!\!\!\perp X|Z$ holds in the original model, we test whether Z d -separates X from Y^* in the twin network. As can be easily seen (see Definition ??), conditioning on Z renders the path between X and Y^* d -connected through the collider at Z , hence $Y_x \perp\!\!\!\perp X|Z$ does not hold in the model. This conclusion is not easily discernible from the chain model itself, or from the equations in that model. In the same fashion we can see that whenever we condition on either Y or on $\{Y, Z\}$, we form a connection between Y^* and X , hence, Y_x and X are not independent conditional on those variables. The connection is disrupted, however, if we do not condition on either Y or Z , hence, $Y(x) \perp\!\!\!\perp X$.

The twin-network reveals an interesting interpretation of counterfactuals of the form Z_{pa_z} , where Z is any variable and PA_Z stands for the set of Z 's parents. Consider the question whether Z_x is independent of some given set of variables in the model of Fig. 3. The answer to this question depends on whether Z^* is d -separated from that set of variables. However, any variable that is d -separated from Z^* would also be d -separated from U_Z , hence, the node representing U_Z can serve as a proxy for representing the counterfactual variable Z_x . This is not a coincidence, considering that Z is governed by the equation $z = f_Z(x, u_Z)$. By definition, the statistics of Z_x is equal to the statistics of Z under the condition where X is held fixed at x . Under such condition, Z may vary only if U_Z varies. Therefore, if U_Z obeys a certain independence relationship, Z_x must obey that relationship as well. In general, if U_Z obeys a certain independence relationship, Z_{pa_z} must obey that relationship as well. We thus obtain a simple graphical representation for any counterfactual variable of the form Z_{pa_z} . Using this representation, we can easily verify from Fig. 3 that $(Y(x) \perp\!\!\!\perp X|\{Z, U_Z, Y\})_G$ and $(Y(x) \perp\!\!\!\perp X|\{U_Y, U_Z, Y\})_G$ both hold in the twin-network and, therefore,

$$Y_x \perp\!\!\!\perp X|\{Z, Z_x, Y\} \text{ and } Y_x \perp\!\!\!\perp X|\{Y_z, Z_x, Y\}$$

must hold in the model.

3 Applications and Interpretation of Structural Models

It is worth emphasizing, at this point, that the problem of computing counterfactual expectations is not an academic exercise; it represents in fact the typical case in almost every economic policy analysis. Whenever we undertake to predict the effect of policy, two considerations apply. First, economic policy variables (e.g., taxation or interest rates) are rarely exogenous. Policy variables are endogenous in the data-analysis phase of the study and turn exogenous in the planning phase, when we contemplate their implementation. Second, policies are rarely evaluated in the abstract; rather, they are brought into focus by certain eventualities that demand remedial correction. These two considerations are faithfully represented in the third query above. The current price p_0 represents economic indicators that are available at the time of decision, and that will be affected by the policies considered. The variable P represents endogenous decision variables (as shown in Fig. ??) that turn exogenous in deliberation, as dictated by the submodel $M_{P=p_0}$. The hypothetical mood of Query 3 translates into a practical problem of policy analysis: "Given that the current price is $P = p_0$, find the expected value of the demand (Q) if we change the price today to $P = p_1$." The reasons for using hypothetical phrases in practical decision-making situations is discussed in the next section (3.1).

3.1 The empirical content of counterfactuals

The word "counterfactual" is a misnomer, as it connotes a statement that stands contrary to facts or, at the very least, a statement that escapes empirical verification. Counterfactuals are in neither category; they are fundamental to scientific thought and carry as clear an empirical message as any scientific law.

Consider Ohm's law $V = IR$. The empirical content of this law can be encoded in two alternative forms.

1. **Predictive form:** If at time t_0 we measure current I_0 and voltage V_0 then, *ceteras paribum*, at any future times $t > t_0$, if the current flow will be $I(t)$ the voltage drop will be:

$$V(t) = \frac{V_0}{I_0} I(t).$$

2. **Counterfactual form:** If at time t_0 we measure current I_0 and voltage V_0 then, had the current flow at time t_0 been I' , instead of I_0 , the voltage drop would have been:

$$V' = \frac{V_0 I'}{I_0}$$

On the surface, it seems that the predictive form makes meaningful and testable empirical claims while the counterfactual form merely speculates about events that have not, and could not have occurred; as it is impossible to apply two different currents into the same

resistor at the same time. However, if we interpret the counterfactual form to mean no more nor less than a conversational short hand of the predictive form, the empirical content of the former shines through clearly. Both enable us to make an infinite number of predictions from just one measurement (I_0, V_0) , and both derive their validity from a scientific law (Ohm's law) which ascribes a time-invariant property (the ratio V/I) to any physical object.

We will adapt this predictive interpretation when we speak of counterfactuals. This interpretation is further supported by the observation that counterfactuals, despite their a-temporal appearance, are invariably associated with some law-like, persistent relationships in the world. For example, the statement "had Germany not been punished so severely at the end world-war I, Hitler would not have come to power" would sound bizarre to anyone who does not share our understanding that, as a general rule, "humiliation breeds discontent."

But if counterfactual statements are merely a round-about way of stating sets of predictions, why do we resort to such convoluted modes of expression instead of using the predictive mode directly? The answer, it seems, rests with the qualification "ceteris paribus" that accompanies the predictive claim, which is not entirely free of ambiguities. What should be held constant when we change the current in a resistor? The temperature? the laboratory equipments? the time of day? Certainly not the reading on the voltmeter? Such matters must be carefully specified when we pronounce predictive claims and take them seriously. Many of these specifications are implicit (hence superfluous) when we use counterfactual expressions, especially when we agree over the underlying causal model. For example, we do not need to specify under what temperature and pressure future predictions should hold true; these are implied by the statement "had the current flow at time t_0 been I' , instead of I_0 ." In other words, we are referring to precisely those conditions that prevailed in our laboratory at time t_0 . That statement also implies that we do not really mean for anyone to hold the reading on the voltmeter constant – only variables that, according to our causal model, are not affected by the counterfactual antecedent (I) are expected to remain constant for the predictions to hold true.

To summarize, a counterfactual statement might well be interpreted to convey a set of predictions under well defined set of conditions, those prevailing in the factual part of the statement. For these predictions to be valid, two components must remain invariants: the laws (or mechanisms) and the boundary conditions. Cast in the language of structural models, the laws correspond to the equations $\{f_i\}$ and the boundary conditions correspond to the state of the background variables U . Thus, a precondition for the validity of the predictive interpretation of a counterfactual statement is the assumption that U will remain the same at the time where our predictive claim is to be applied or tested.

This is best illustrated using a betting example. We

must bet heads or tails on the outcome of a fair coin toss; we win a dollar if we guess correctly, lose if we don't. Suppose we bet heads and we win a dollar, without glancing at the outcome of the coin. Consider the counterfactual "Had I bet differently I would have lost a dollar". The predictive interpretation of this sentence translates into the implausible claim: "If my next bet is tails, I will lose a dollar." For this claim to be valid, two invariants must be assumed: the payoff policy and the outcome of the coin. While the former is a plausible assumption in betting context, the latter would be realized in only rare circumstances. It is for this reason that the predictive utility of the statement "Had I bet differently I would have lost a dollar" is rather low, and some would even regard it as hind-sighted nonsense. (It is not hard however to imagine a lottery in which the payoff policy and the outcome of the random device remain constant for a short period of time, during which additional bets are accepted and processed. Most those who play the stock market believe in strategies that allow an investor to quickly recover from a bad move.) At any rate, it is the persistence across time of U and $f(x, u)$ that endows counterfactual expressions with predictive power; take this persistence away, and the counterfactual loses its obvious economical utility.

I said "obvious" because there is an element of utility in counterfactuals that does not translate immediately to predictive payoff, and may explain, nevertheless, the ubiquity of counterfactuals in human discourse. I am thinking of explanatory value. Suppose, in the betting story, coins were tossed afresh for every bet. Is there no value whatsoever to the statement "Had I bet differently I would have lost a dollar?" I believe there is; it tells us that we are not dealing here with a whimsical bookie like the one who decides which way to spin our atoms and electrons, but one who at least glances at the bet, compares it to some standard, and decides a win or a loss using a consistent policy. This information may not be very useful to us as players, but it may be useful to say state inspectors who come every so often to calibrate the gambling machines to ensure the State's take of the profit. More significantly, it may be useful to us players, too, if we venture to cheat slightly, say by manipulating the trajectory of the coin, or by installing a tiny transmitter to tell us which way the coin landed. For such cheating to work, we should know the policy $y = f(x, u)$ and the statement "Had I bet differently I would have lost a dollar?" reveals important aspects of that policy.

Is it far fetched to argue for the merit of counterfactuals by hypothesizing unlikely situations where players cheat and rules are broken? I submit that such unlikely operations are the norm in gauging the explanatory value of sentences. In fact, it is the nature of any explanation, especially causal, that its utility be amortized not over standard situations but, rather, over novel settings which require innovative manipulation of one's environment.

Recapping our discussion, we see that counterfactuals may earn predictive value under two conditions; (1) when the unobserved uncertainty-producing variables

(U) remain constant (until our next prediction or action), (2) when the uncertainty-producing variables offer the potential of being observed sometime in the future (before our next prediction or action.) In both cases we also need to ensure that the outcome-producing mechanism $f(x, u)$ persists unaltered.

These conclusions raise interesting questions on the use of counterfactuals in microscopic phenomena, as none of these conditions holds for the type of uncertainty that we encounter in quantum theory. Heisenberg's dice is rolled afresh billions of times each second, and our measurement of u will never be fine enough to remove all uncertainty from the response equation $y = f(x, u)$. Thus, when we include quantum-level processes in our analysis we face a dilemma; either we disband all talk of counterfactuals (a strategy recommended by some researchers [Dawid, 1997]) or we continue to use counterfactuals but limit their usage to situations where they assume empirical meaning. This amounts to keeping in the analysis only U 's that satisfy conditions (1) and (2) above. Instead of hypothesizing U 's that completely remove all uncertainties, we admit only those U 's that are either (1) persistent or (2) potentially observable.

Naturally, coarsening the granularity of the exogenous variables has its price tag; the mechanism equations $y = f(x, u)$ lose their deterministic character and should be made stochastic. Instead of constructing causal models from a set of deterministic equations $\{f_i\}$ we should consider models made up of stochastic functions $\{f_i^*\}$, where each f_i^* is a mapping from $V \cup U$ to some intrinsic probability distribution $P^*(v_i)$ over the states of V_i . This option lies beyond the scope of the present paper, but its basic character should follow from the three steps of abduction-action-deduction, outlined in Section ??.

3.2 Causal explanations, utterances, and their interpretation

It is a commonplace wisdom that explanation improve understanding, and that he who understands more, can reason and learn more effectively. It is also generally accepted that the notion of explanation cannot be divorced from that of causation; e.g., a symptom may explain our *belief* in a disease, but it does not explain the disease itself. However, the precise relationship between causes and explanations is still a topic of much discussion [Cartwright, 1989; Woodward, 1997]. Having a formal theory of causality and counterfactuals in both deterministic and probabilistic settings, casts new light on the question of what constitutes an adequate explanation, and open new possibilities for automatic generation of explanations by machine.

A natural starting point for generating explanations would be to use a causal Bayesian network (Chapter ??) in which the events to be explained (explanandum) consist of some combination e of instantiated nodes in the network, and the task is to find an instantiation c of a subset of e 's ancestors (i.e., causes) that maximizes some measure of "explanatory power," namely, the degree to which c explains e . However, the proper choice of this measure

is unsettled. Many philosophers and statisticians argue for the likelihood ratio $L = \frac{P(e|c)}{P(e|c')}$ as the proper measure of the degree to which c is a better explanation of e than c' . In Pearl (1988, Chapter 5) and Peng and Reggia (1986), the best explanation is found by maximizing the posterior probability $P(c|e)$. Both measures have their faults and have been criticized by several researchers, including Pearl (1988), Shimony (1991,1993), Suermondt and Cooper (1992), and Chajewska and Halpern (1997). To remedy these faults, more intricate combinations of the probabilistic parameters $[P(e|c), P(e|c'), P(c), P(c')]$ have been suggested, none of which seems to capture well the meaning people attach to the word "explanation".

The problem with probabilistic measures is that they cannot capture the strength of *causal* connection between c and e ; any proposition h whatsoever can, with a small stretch of imagination, be thought of as having some influence on e , however feeble. This would then qualify h as an ancestor of e in the causal network and would enable h to compete and win against genuine explanations by virtue of h having strong spurious association with e .

To rid ourselves of this difficulty, we must go beyond probabilistic measures and concentrate instead causal parameters such as *causal effects* $P(y|do(x))$ and counterfactual probabilities $P(Y_{x'} = y'|x, y)$ as the basis for defining explanatory power. Here x and x' range over the set of alternative explanations, and Y is the set of response variables observed to take on the value y . $P(Y_{x'} = y'|x, y)$ is read as: the probability that Y would take on a different value, y' , had X been x' (instead of the actual values x). (Note that $P(y|do(x)) \triangleq P(Y_x = y)$.) The developments of computational models of representing and evaluating causal effects and counterfactual probabilities, now make it possible to combine these parameters with standard probabilistic parameters and synthesize a more faithful measure of explanatory power, to guide the selection and generation of adequate explanations.

These possibilities trigger an important basic question: Is explanation a concept based on *general causes* (e.g., "Drinking hemlock causes death,") or *singular causes* (e.g., "Socrates' drinking hemlock caused his death,"). Causal effect expressions $P(y|do(x))$ belong to the first category while counterfactual expressions, $P(Y_{x'} = y'|x, y)$ belong to the second, since conditioning on x and y narrows down world scenarios to those compatible with all the specific information at hand.

The classification of causal statements into general and singular categories has been the subject of intensive research in philosophy e.g., see [Kvart, 1986; Good, 1983; Cartwright, 1989; Eells, 1991]. This research has attracted little attention in cognitive science and artificial intelligence, partly because it has not entailed practical inferential procedures, and partly because it was based on problematic probabilistic semantics (see Section ?? for discussion of probabilistic causality). In the context of machine generated explanations, this classi-

fication assumes both cognitive and computational significance. We have discussed in Chapter ?? the sharp demarcation line between two types of causal queries, those that are answerable from the pair $\langle P, G \rangle$ (where G is a causal graph compatible with probability function P), and those that require additional information in the form of functional specification. Generic causal statements (e.g., $P(y|do(x))$) often fall in the first category (as in Chapter ??) while counterfactual expressions (e.g., $P(Y_{x'} = y|x, y)$) fall in the second, thus demanding more detailed specifications and higher computational resources.

The proper classification of explanation into a general or singular category depends on whether the cause c attains its explanatory power relative to its effect e by virtue of c 's general tendency to produce e (as compared with the weaker tendencies of c 's alternatives) or by virtue of c being necessary for triggering a specific chain of events leading to e in the specific situation at hand (as characterized by e and perhaps other facts and observations.) Formally, the difference hinges on whether, in evaluating explanatory powers of various hypotheses, we should condition our beliefs on the events c and e that actually occurred.

Formal analysis of these alternatives is given in chapter ??, when we discuss the necessary and sufficient aspects of causation relative to the notion of probability of causation. In the sequel of this section we will be concerned with the interpretation and generation of explanatory utterances, taking the necessary aspect as a norm.

The following list, taken from [Galles and Pearl, 1997], provides brief examples of utterances used in explanatory discourse and their associate semantics in the modifiable structural model approach described in Section 2.1.

- “ X is a cause of Y ”, if there exist two values x and x' of X and a value u of U such that $Y_x(u) \neq Y_{x'}(u)$.
- “ X is a cause of Y in context $Z = z$ ”, if there exist two values x and x' of X and a value u of U such that $Y_{xz} \neq Y_{x'z}(u)$.
- “ X is a direct cause of Y ”, if there exist two values x and x' of X , and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$ where r is some realization of $V \setminus X$.
- “ X is an indirect cause of Y ”, if X is a cause of Y , and X is not a direct cause of Y .
- “Event $X = x$ may have caused $Y = y$ ” if
 - (i) $X = x$ and $Y = y$ are true, and
 - (ii) There exists a value u of U such that $X(u) = x$, $Y(u) = y$, $Y_x(u) = y$ and $Y_{x'}(u) \neq y$ for some $x' \neq x$.
- “The unobserved event $X = x$ is a likely cause of $Y = y$ ” if
 - (i) $Y = y$ is true, and
 - (ii) $P(Y_x = y, Y_{x'} \neq y | Y = y)$ is high for some $x' \neq x$

- “Event $Y = y$ occurred despite $X = x$ ”, if
 - (i) $X = x$ and $Y = y$ are true, and
 - (ii) $P(Y_x = y)$ is low.

The preceding list demonstrates the flexibility of modifiable structural models in formalizing nuances of causal expressions. Additional nuances, invoking notions such as *enabling*, *preventing*, *maintaining*, and *producing*, etc. will be analyzed in Chapter ?. Related expressions are: “Event A *explains* the occurrence of event B ”, or “ A would *explain* B if C were the case”, or “ B occurred *despite* of A , *because* C was true”. The ability to interpret and generate such explanatory sentences, or to select the expression most appropriate for the context is one of the most intriguing challenges of research in man-machine conversation.

3.3 From mechanisms to actions to causation

The structural model semantics described in Section 2.1 suggests solutions to two problems in cognitive science and artificial intelligence: the representation of actions, and the role of causal ordering. We will discuss these problems in turns, since the second builds on the first.

Action, Mechanisms and surgeries

Whether we take the probabilistic paradigm that actions are transformations from probability distributions to probability distributions, or the deterministic paradigm that actions are transformations from states to states, such transformations could in principle be infinitely complex. Yet, in practice, people teach each other rather quickly what actions normally do to the world, and people predict the consequences of most actions without much hustle. How?

Structural models answer this question by assuming that the actions we normally invoke in common reasoning can be represented as *local surgeries*. The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behavior of a relatively small group of variables. If we understand how the linkages interact with each other (usually they simply share variables) we should also be able to understand what the effect of any given action would be: Simply re-specify those few mechanisms that are perturbed by the action, then let the modified assembly of mechanisms interact with one another, and see what state will evolve at equilibrium. If the specification is complete (i.e., M and U are given), a single state will evolve. If the specification is probabilistic (i.e., $P(u)$ is given) a new probability distribution will emerge and, if the specification is partial (i.e., some f_i 's are not given) a new, partial theory will then be created. In all three cases we should be able to answer queries about post-action states of affair, albeit with decreasing level of precision.

The ingredient that makes this scheme operational is the *locality* of actions. Standing alone, locality is a vague concept because what is local in one space may not be

local in another. A speck of dust, for example, appears extremely diffused in the frequency (or Fourier) representation and, vice versa, a pure musical tone requires a long stretch of time to be appreciated. Structural semantics emphasizes that actions are local in the space of mechanisms and not in the space of variables or sentences or time slots. For example, tipping the left-most object in an array of domino tiles does not appear “local” in physical space, because, in the tradition of domino theories, every tile might be affected by such action. Yet the action is quite local in the mechanism domain: Only one mechanism gets perturbed, the gravitational restoring force that normally keeps the left-most tile in a stable erect position; all other mechanisms remain unaltered, as specified, obedient to the usual equations of physics. It takes no more than a second to describe this action on the phone, without enumerating all its ramifications. The listener, assuming she shares our understanding of domino physics, can figure out for herself the ramifications of this action, or any action of the type: “tip the *i*th domino tile to the right.” By representing the domain in the form of an assembly of stable mechanisms, we have in fact created an oracle capable of answering queries about the effects of a huge set of actions and action combinations, without us having to explicate those effects.

Laws vs. facts

This surgical procedure sounds trivial when expressed in the context of structural equation models. However, it has encountered great difficulties when attempts were made to implement such schemes in classical logic. In order to implement surgical procedures in mechanism space, we need a language in which some sentences are given different status than others; sentences describing mechanisms should be treated differently than those describing other facts of life, such as observations, assumption and conclusions, because the former are presumed stable, while the latter are transitory. Indeed the equations which describe how the domino tiles interact with one another remain unaltered whereas the states of the tiles themselves are free to vary with circumstances.

Admitting the need for this distinction has been a difficult transition in the logical approach to actions and causality, perhaps because much of the power of classical logic stems from its representational uniformity and syntactic invariance, where no sentence commands special status. Probabilists were much less reluctant to embrace the distinction between laws and facts, because this distinction has already been programmed into probability language by Reverend Bayes in 1763: Facts are expressed as ordinary propositions, hence they can obtain probability values and they can be conditioned on; laws, on the other hand, are expressed as conditional-probability sentences (e.g., $P(\textit{accident}|\textit{careless-driving}) = \textit{high}$), hence they should not be assigned probabilities and cannot be conditioned on. It is due to this tradition that probabilists have always attributed nonpropositional character to conditional sentences (e.g., birds fly); refusing to al-

low nested conditionals [Levi, 1988], and insisting on interpreting one’s confidence in a conditional sentence as a conditional probability judgment [Adams, 1975] (see also [Lewis, 1976]). Remarkably, these constraints, which some philosophers view as limitations, are precisely the safeguards that have kept probabilists from confusing laws and facts, and have protected them from some of the traps that have lured logical approaches.⁷

Mechanisms and causal relationships

From our discussion thus far, it may seem that one can construct an effective representation for computing the ramification of actions without appealing to any notion of causation. This is indeed feasible in many areas of physics and engineering. If we have, for instance, a large electric circuit consisting of resistors, voltage sources and current sources, and we are interested in computing the effect of changing one resistor in the circuit, the notion of causality hardly enters the computation. We simply insert the modified value of the resistor into Ohm’s and Kirchoff’s equations, and solve the set of (symmetric) equations for the variable needed. This computation can be performed effectively without committing to any directional causal relationship between the currents and voltages.

To understand the role of causality, we should note that, unlike the resistor-network example, most mechanisms do not have names in common everyday language. In the domino example above I had to struggle hard to name the mechanism which would be perturbed by the action “tip the left-most tile to the right.” This struggle can be saved through causation: instead of telling you the name of the mechanism to be perturbed by the action, I merely describe its net result, in the form of an *event* (or a proposition), e.g., “the left-most tile was tipped to the right.” If your knowledge is organized causally, this specification is sufficient, because you would be able to figure out for yourself which mechanism it is that must have been perturbed in realizing the specified new event, and this should enable you to predict the rest of the scenario.

This linguistic abbreviation defines a new relation among events, a relation we normally call “causation”: Event *A* causes *B*, if the perturbation needed for realizing *A* entails the realization of *B*.⁸ Causal abbreviations of this sort are used very effectively for specifying domain knowledge. Complex descriptions of what relationships are stable and how mechanisms interact with one another are rarely communicated explicitly in terms

⁷The distinction between laws and facts has been proposed by Poole (1985) and Geffner (1992) as a fundamental principle for nonmonotonic reasoning. In database theory, laws are expressed by special sentences called *integrity constraints* [Reiter, 1987]. The distinction seems to be gaining broader support recently as a necessary requirement for formulating actions [Sandewall 1994; Lin 1995].

⁸The word “needed” connotes minimality and can be translated to: “...if every minimal perturbation realizing *A*, entails *B*”.

of mechanisms. Rather, they are communicated in terms of cause-effect relationships between events or variables. We say, for example: “If tile i is tipped to the right, it causes tile $i + 1$ to tip to the right as well”; we do not communicate such knowledge in terms of the tendencies of each domino tile to maintain its physical shape, to respond to gravitational pull and to obey Newtonian mechanics.

3.4 Simon’s causal ordering

Our ability to talk directly in terms of one event causing another, (rather than an action altering a mechanism and the alteration, in turn, producing the effect) is computationally very useful, but, at the same time it requires that the assembly of mechanisms in our domain satisfy certain conditions which lead to causal directionality. Some of these conditions are structural, and others are substantive—invoking relative magnitudes of forces and powers.

The structural requirement leading to causal directionality is formulated in Simon’s notion of “causal ordering” [Simon, 1953]. This notion provides a rationale for designating a variable in each mechanism as the output (or effect), and the other variables, as inputs (or causes). Indeed, the formal definition of causal models given in Section 2.1 assumes that each equation is designated a distinct privileged variable, situated on its left hand side, that is considered “dependent.” In general, however, a mechanism may be specified as a functional constraint

$$G_k(x_1, \dots, x_k; u_1, \dots, u_m) = 0$$

without identifying any so called “dependent” variable. Simon’s causal ordering provides a procedure for deciding whether a collection of such symmetric G functions has a unique way of designating an endogenous “dependent” variable to each mechanisms, while excluding the background variables (since they are determined outside the system).

Simon asked: when can we order the variables (V_1, V_2, \dots, V_n) in such a way that we can solve for each V_i without solving for any of V_i ’s successors? Such an ordering, if it exists, dictates the direction we attribute to causation. This criterion might at first sound artificial, since the order of solving equations is a matter of computational convenience while causal directionality is an objective attribute of physical reality. (See [Iwasaki and Simon, 1986] and [De Kleer and Brown, 1986] for discussion of this topic.) To justify the criterion, we rephrase Simon’s question in terms of actions and mechanism. Assume each mechanism (i.e., equation) can be modified independently of the others and let A_k be the set of actions capable of modifying equation G_k (while leaving other equations unaltered). Imagine that we have chosen an action a_k from A_k , and that we have modified G_k in such a way that the set of solutions $(V_1(u), V_2(u), \dots, V_n(u))$ to the entire system of equations differs from what it was prior to the action. If X is the set of endogenous

variables constrained by G_k , we can ask which members of X would change by the modification. If only one member of X changes, say X_k , and if the identity of that distinct member remains the same for all choices of a_k and u , we designate X_k as the “dependent” variable in G_k . Formally, this property means that the changes in a_k induce a *functional mapping* from the domain of X_k to the domain of $\{V \setminus X_k\}$; all changes in the system (generated by a_k) can be attributed to changes in X_k . It would make sense, in such a case, to designate X_k as a “representative” of the mechanism G_k , and we would be justified in replacing the sentence “action a_k caused event $Y = y$ ” with “Event $X_k = x_k$ caused $Y = y$ ” (Y being any variable in the system). The invariance of X_k to the choice of a_k is the basis for treating an action as a modality $do(X_k = x_k)$ (Definition 2.3). It provides a license for characterizing an action by its immediate consequence(s), independent of the instrument that actually brought about those consequences, and defines in fact the notion of “local action” or “local surgery”.

It can be shown [Nayak 1994, Roizen 1999] that the uniqueness of X_k can be determined by a simple criterion which involves purely topological properties of the equation set (i.e., how variables are grouped into equations). The criterion is that one should be able to form one-to-one correspondence between equations and variables and that the correspondence be unique. This can be decided by solving the *matching problem* [Hall, 1935; Serrano and Gossard, 1987] between equations and variables. If the matching is unique, then the choice of dependent variable in each equation is unique and the directionality induced by that choice defines a directed acyclic graph (DAG). In Figure 1, for example, the directionality of the arrows need not be specified externally, they can be determined mechanically from the set of symmetrical constraints (i.e., logical propositions):

$$S = \{G_1(C, U), G_2(A, C), G_3(B, C), G_4(A, B, D)\} \quad (7)$$

that characterizes the problem. The reader can easily verify that the selection of a privileged variable from each equations is unique and, hence, that the causal directionality of the arrows shown in Figure 1 is inevitable.

Thus, we see that causal directionality, according to Simon, emerges from two assumptions: 1. The partition of variables into background (U) and endogenous (V) sets, and 2. the overall configuration of mechanisms in the model. Accordingly, a variable designated as “dependent” in a given mechanism may well be labeled “independent” when that same mechanism is embedded in a different model. Indeed, the compression of a spring is deemed to cause the acceleration of an object tied to that spring, but when placed in a space ship, the acceleration of the object is considered the cause of the spring’s compression.

Of course, if we have no way of determining the background variables, then several causal orderings may ensue. In Eq. (7), for example, if we were not given the information that U is a background variable, then either one of $\{U, A, B, C\}$ can be chosen as background,

and each such choice would induce a different ordering on the remaining variables. (Some would conflict with commonsense knowledge, e.g., that the Captain’s signal influences the court decision). The directionality of $A \rightarrow D \leftarrow B$ however, would be maintained in all those orderings. The question whether there exists a partition $\{U, V\}$ of the variables that would yield a causal ordering in a system of symmetric constraints can also be solved (in polynomial time) by topological means [Dechter and Pearl, 1991].

Simon’s ordering criterion fails when we are unable to solve the equations one at a time, but must solve a block of k equations simultaneously. In such a case, all the k variables determined by the block would be mutually unordered, though their relationships with other blocks may still be ordered. This occurs, for example, in the economic model of Figure ??, where Eqs. (??) and (??) need to be solved simultaneously for P and Q , and hence the correspondence between equations and variables is not unique; either Q or P could be designated as “independent” in either of the two equations. Indeed, the information needed for classifying Eq. (??) as the “demand equation” (and, respectively, Eq. (??) as the “price equation”) comes not from the way variables are assigned to equations, but from subject-matter considerations. Our understanding that household income directly affects household demand (and not prices), plays a major role in this classification.

In the general case, when we tend to assert categorically that the flow of causation in a feedback loop goes clockwise, rather than counterclockwise, the assertion is normally based on the relative magnitudes of forces. For example, turning the faucet would lower the water level in the water tank but there is practically nothing we can do to the water tank that would turn the faucet. When such information is available, causal directionality is determined by appealing, again, to the notion of hypothetical intervention and asking whether an external control over one variable in the mechanism necessarily affects the others. This consideration then constitutes the operational semantics for identifying the dependent variables V_i in nonrecursive causal models (Definition 2.1).

The asymmetry that characterizes causal relationships in no way conflicts with the symmetry of physical equations (see Chapter ?? for discussion of Russell’s problem with this disparity.) By saying that “ X causes Y and Y does not cause X ” we mean to say that changing a mechanism in which X is normally the dependent variable has a different effect on the world than changing a mechanism in which Y is normally the dependent variable. Since two separate mechanisms are involved, the statement stands in perfect harmony with the symmetry we find in the equations of physics.

Kit Fine has further demonstrated that similarity of appearance is inadequate [Fine, 1975]. Fine considers the counterfactual “Had Nixon pressed the button, a nuclear war would have started,” which is generally accepted as true. However, a world in which the button

happened to be disconnected, is many times more similar to our world, as we know it, than the one yielding a nuclear blast. This example demonstrates not only that similarity measures could not be arbitrary, but also that they must respect our conception of causal laws.⁹ Lewis (1979) has subsequently set up an intricate system of weights and priorities among various dimensions of similarity: size of “miracles” (violations of laws), matching of facts, temporal precedence etc., to bring similarity closer to causal intuition. These priorities are not unproblematic (J. Woodward, personal communication) and are rather post-hoc.

Such difficulties do not enter the structural account. In contrast with Lewis’ theory, counterfactuals are not based on abstract notion of similarity among hypothetical worlds; they rest directly on the mechanisms (or “laws,” to be fancy) that produce those worlds, and on the invariant properties of those mechanisms. Lewis’ elusive “miracles” are replaced by principled mini-surgeries, $do(X = x)$, which represent the minimal change (to a model) necessary for establishing the antecedent $X = x$ (for all u). Thus, similarities and priorities, if they are ever needed, may be read into the $do(*)$ operator as an afterthought (see [Goldszmidt and Pearl, 1992]), but are not basic to the analysis.

The structural account answers the mental representation question by offering a parsimonious encoding of knowledge, from which causes, counterfactuals and probabilities of counterfactuals can be derived by effective algorithms. This parsimony is acquired at the expense of generality though; limiting the counterfactual antecedent to conjunction of elementary propositions prevents us from analyzing disjunctive hypotheticals such as “if Bizet and Verdi were compatriots.”

Axiomatic comparison

If our assessment of inter-world distances comes from causal knowledge, the question arises whether that knowledge does not impose its own structure on distances, a structure that is not captured in Lewis’ logic. Phrased differently, by agreeing to measure closest worlds on the basis of causal relations, do we restrict the set of counterfactual statements we regard as valid? The question is not merely theoretical. For example, Gibbard and Harper (1981) characterize decision-making conditionals, namely, sentences of the form “If we do A , then B ,” using Lewis’s general framework, while our $do(*)$ operator is based directly on causal semantics, and whether the two formalisms are identical is uncertain.¹⁰

We now show that the two formalisms are identical for recursive systems, composition, and effectiveness held in with respect to Lewis’s closest-world framework whenever recursiveness does. We begin by providing a ver-

⁹In this respect, Lewis’ reduction of causes to counterfactuals is somewhat circular.

¹⁰Winslett (1988) and Ginsberg and Smith (1987) have also advanced theories of actions based on closest-world semantics, and have not imposed any special structure for the distance measure, to reflect causal considerations.

sion of Lewis's logic for counterfactual sentences (from [Lewis, 1981]).

References

- [Adams, 1975] E. Adams. *The Logic of Conditionals*, chapter 2. D. Reidel, Dordrecht, Netherlands, 1975.
- [Balke and Pearl, 1994] A. Balke and J. Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume I, pages 230–237. MIT Press, Menlo Park, CA, 1994.
- [Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.
- [Cartwright, 1989] N. Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.
- [Chajewska and Halpern, 1997] U. Chajewska and J.Y. Halpern. Defining explanation in probabilistic systems. In D. Geiger and P.P. Shenoy, editors, *Uncertainty in Artificial Intelligence 13*, pages 62–71. Morgan Kaufmann, San Francisco, CA, 1997.
- [Dawid, 1997] A.P. Dawid. Causal inference without counterfactuals. Technical report, Department of Statistical Science, University College London, UK, 1997.
- [De Kleer and Brown, 1986] J. De Kleer and J.S. Brown. Theories of causal ordering. *Artificial Intelligence*, 29(1):33–62, 1986.
- [Dechter and Pearl, 1991] R. Dechter and J. Pearl. Directed constraint networks: A relational framework for casual modeling. In *Proceedings of the 12th International Joint Conference of Artificial Intelligence (IJCAI-91)*, pages 1164–1170, Sydney, Australia, 1991. Morgan Kaufmann, San Mateo, CA.
- [Eells, 1991] E. Eells. *Probabilistic Causality*. Cambridge University Press, Cambridge, MA, 1991.
- [Eshghi and Kowalski, 1989] K. Eshghi and R.A. Kowalski. Abduction compared with negation as failure. In G. Levi and M. Martelli, editors, *Proceedings of the Sixth International Conference on Logic Programming*, pages 234–254. MIT Press, 1989.
- [Fine, 1975] K. Fine. Review of lewis' counterfactuals. *Mind*, 84:451–458, 1975.
- [Fine, 1985] K. Fine. *Reasoning with Arbitrary Objects*. B. Blackwell, New York, 1985.
- [Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.
- [Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.
- [Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.
- [Geffner, 1992] H. Geffner. *Default Reasoning: Causal and Conditional Theories*. MIT Press, Cambridge, MA, 1992.
- [Gibbard and Harper, 1981] A. Gibbard and L. Harper. Counterfactuals and two kinds of expected utility. In W. L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, pages 153–190. D. Reidel, Dordrecht, 1981.
- [Ginsberg and Smith, 1987] M.L. Ginsberg and D.E. Smith. Reasoning about action I: A possible worlds approach. In Frank M. Brown, editor, *The Frame Problem in Artificial Intelligence*, pages 233–258. Morgan Kaufmann, Los Altos, CA, 1987.
- [Ginsberg, 1986] M.L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(35–79), 1986.
- [Goldszmidt and Pearl, 1992] M. Goldszmidt and J. Pearl. Rank-based systems: A simple approach to belief revision, belief update, and reasoning about evidence and actions. In B. Nebel, C. Rich, and W. Swartout, editors, *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*, pages 661–672. Morgan Kaufmann, San Mateo, CA, 1992.
- [Good, 1983] I.J. Good. A causal calculus. *British Journal for Philosophy of Science*, 11, 12, and 13:305–328, 43–51, and 88, 1983. Reprinted as Ch. 21 in *Good Thinking*, University of Minnesota Press, Minneapolis, MN.
- [Hall, 1935] P. Hall. On representatives of subsets. *Journal of London*, 10:26–30, 1935.
- [Hall, 1998] N. Hall. Two concepts of causation, 1998. In press.
- [Halpern, 1998] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998.
- [Hume, 1948] D. Hume. *An Enquiry concerning Human Understanding*. Open Court Press, LaSalle, 1948. Reprinted 1988.
- [Iwasaki and Simon, 1986] Y. Iwasaki and H.A. Simon. Causality in device behavior. *Artificial Intelligence*, 29(1):3–32, 1986.
- [Katsuno and Mendelzon, 1991] H. Katsuno and A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 387–394, Boston, MA, 1991.
- [Kvart, 1986] I. Kvart. *A Theory of Counterfactuals*. Hackett Publishing, Co., Indianapolis, 1986.
- [Levi, 1988] I. Levi. Iteration of conditionals and the ramsey test. *Synthese*, 76:49–81, 1988.

- [Lewis, 1973a] D. Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973.
- [Lewis, 1973b] D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.
- [Lewis, 1976] D. Lewis. Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85:297–315, 1976.
- [Lewis, 1979] D. Lewis. Counterfactual dependence and time’s arrow. *Nous*, 13:418–446, 1979.
- [Lewis, 1981] D. Lewis. Counterfactuals and comparative possibility. In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*. D. Reidel, Dordrecht, Holland, 1981.
- [Lewis, 1986] D. Lewis. *Philosophical Papers*, volume II. Oxford University Press, New York, 1986.
- [Lin, 1995] F. Lin. Embracing causality in specifying the indeterminate effects of actions. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Quebec, 1995.
- [Marschak, 1950] J. Marschak. Statistical inference in economics. In T. Koopmans, editor, *Statistical Inference in Dynamic Economic Models*, pages 1–50. Wiley, New York, 1950. Cowles Commission for Research in Economics, Monograph 10.
- [Mill, 1843] J.S. Mill. *System of Logic*, volume 1. John W. Parker, London, 1843.
- [Nayak, 1994] P. Nayak. Causal approximations. *Artificial Intelligence*, 70:277–334, 1994.
- [Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.
- [Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.
- [Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, December 1995.
- [Peng and Reggia, 1986] Y. Peng and J.A. Reggia. Plausibility of diagnostic hypotheses. In *Proc., 5th Natl. Conf. on AI (AAAI-86)*, pages 140–45, Philadelphia, 1986.
- [Poole, 1985] D. Poole. On the comparison of theories: Preferring the most specific explanations. In *Proceedings of International Conference on Artificial Intelligence (IJCAI-85)*, pages 144–147, Los Angeles, CA, 1985.
- [Reiter, 1987] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.
- [Robins, 1986] J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.
- [Roizen, 1990] D.I. Roizen. Zigzagging through causal webs. Technical Report R-265, UCLA, Computer Science Department, 1990.
- [Sandewall, 1994] E. Sandewall. *Features and Fluents*, volume 1. Clarendon Press, Oxford, 1994.
- [Serrano and Gossard, 1987] D. Serrano and D.C. Gossard. Constraint management in conceptual design. In D. Sriram and R.A. Adey, editors, *Knowledge Based Expert Systems in Engineering: Planning and Design*, pages 211–224. Computational Mechanics Publications, 1987.
- [Shimony, 1991] S.E. Shimony. Explanation, irrelevance and statistical independence. In *Proceedings of the Ninth Conference on Artificial Intelligence (AAAI’91)*, pages 482–487, 1991.
- [Shimony, 1993] S.E. Shimony. Relevant explanations: Allowing disjunctive assignments. In D. Heckerman and A. Mamdani, editors, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pages 200–207, San Mateo, CA, July 1993. Morgan Kaufmann Publishers.
- [Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.
- [Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.
- [Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.
- [Spirtes et al., 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.
- [Strotz and Wold, 1960] R.H. Strotz and H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.
- [Suermondt and Cooper, 1992] H.J. Suermondt and G.F. Cooper. An evaluation of explanations of probabilistic inference. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, pages 579–585, 1992.
- [Winslett, 1988] M. Winslett. Reasoning about action using a possible worlds approach. In *Proceedings of the Seventh American Association for Artificial Intelligence Conference*, pages 89–93, 1988.
- [Woodward, 1997] J. Woodward. Explanation, invariance and intervention. *Philosophy of Science*, 64(S):26–S41, 1997.