CAUSATION, ACTION, AND COUNTERFACTUALS

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

1 INTRODUCTION

The central aim of many empirical studies in the physical, behavioral, social, and biological sciences is the elucidation of cause-effect relationships among variables. It is through cause-effect relationships that we obtain a sense of a "deep understanding" of a given phenomenon, and it is through such relationships that we obtain a sense of being "in control," namely, that we are able to shape the course of events by deliberate actions or policies. It is for these two reasons, understanding and control, that causal thinking is so pervasive, popping up in everything from everyday activities to high-level decision-making: For example, every car owner wonders why an engine won't start; a cigarette smoker would like to know, given his/her specific characteristics, to what degree his/her health would be affected by refraining from further smoking; a policy maker would like to know to what degree anti-smoking advertising would reduce costs of health care; and so on. Although a plethora of data has been collected on cars and on smoking and health, the appropriate methodology for extracting answers to such questions from the data has been fiercely debated, partly because some fundamental questions of causality have not been given fully satisfactory answers.

The two fundamental questions of causality are:

1. What empirical evidence is required for legitimate inference of cause-effect relationships?

2. Given that we are willing to accept causal information about a certain phenomenon, what inferences can we draw from such information, and how?

The primary difficulty is that we do not have a clear empirical semantics for causality; statistics teaches us that causation cannot be defined in terms of statistical associations, while any philosophical analysis of causation in terms of deliberate control quickly reaches metaphysical dead-ends over the meaning of free will. Indeed, Bertrand Russell noted that causation plays no role in physics proper and offered to purge the word from the language of science.

Philosophical difficulties notwithstanding, scientific disciplines that must depend on causal thinking have developed paradigms and methodologies that successfully bypass the unsettled questions of causation and that provide acceptable answers to pressing problems of experimentation and inference. Social scientists, for example, have adopted path analysis and structural equations, and programs such as LISREL have become common tools in social science research. Econometricians, likewise, have settled for stochastic simultaneous equations models as carriers of causal information and have focused most of their efforts on developing statistical techniques for estimating the parameters of these models. Statisticians, in contrast, have adopted Fisher's randomized experiment as the ruling paradigm for causal inference, with occasional excursions into its precursor – the Neyman-Rubin model of potential response.

None of these paradigms and methodologies can serve as an adequate substitute for a comprehensive theory of causation, suitable for AI purposes. The structural equations model is based largely on informal modeling assumptions and has hardly been applied beyond the boundaries of linear equations with Gaussian noise. The statisticians' paradigm, on the other hand, is too restrictive, as it does not allow for the integration of large body of substantive knowledge with statistical data. And philosophers have essentially abandoned the quest for the empirical basis of causation, focusing instead on semantical analysis of counterfactual and subjunctive conditionals, with little attention to issues of representation.

A new perspective on the problem of causation has recently emerged from a rather unexpected direction – computer science. When encoding and processing causal relationships on digital machines became necessary, the problems and assumptions that other disciplines could keep dormant and implicit had to be explicated in great detail, so as to meet the levels of precision necessary in programming.

The need to explicate cause-effect relationships arose in several areas of AI: automated diagnosis, robot planning, qualitative physics, and database updates. In the area of robotics, for example, the two fundamental problems of causation got translated into concrete and practical questions:

1. How should a robot acquire causal information through interaction with its environment?

2. How should a robot process the causal information it receives from its creator-programmer?

Attempts to gloss over previous difficulties with causation quickly resulted in a programmer's nightmare. For example, when given the information: "If the grass is wet, then the sprinkler must have been on" and "If I break this bottle, the grass will get wet", the computer will conclude: "If I break this bottle, the sprinkler must have been on". The swiftness and concreteness with which such bugs surface has enabled computer scientists to pinpoint loosely stated assumptions and then to assemble new and more coherent theories of actions, causation, and change.

The purpose of this paper is to summarize recent advances in causal reasoning, especially those that use causal graphs. Section 2 will survey some of the difficulties connected with the formalization of actions in AI. Building on the paradigm that actions are a form of surgeries performed on theories of mechanisms, we will describe the emergence of the causal relation as an abbreviation for the surgery process. Finally, we connect actions to Lewis' theory of counterfactuals using imaging to replace conditioning.

Section 3 will provide the formal underpinning for the discussion of Section 2 and will demonstrate how causal graphs can resolve many of the confusions and contradictions that have prevented a workable theory of actions from evolving. Specifically, in Section 3.1 it will be shown that if qualitative causal information is encoded in the form of a graph, then it is very simple to assess, from non-experimental data, both the strength with which causal influences operate among variables and how probabilities will change as a result of external interventions. Using graph encoding, it is also possible to specify conditions under which manipulative experiments are not necessary and where passive observations suffice. The graphs can also be queried to produce mathematical expressions for causal effects, or to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.

Finally, Section 3.2 will present a symbolic machinery that admits both probabilistic and causal information about a given domain, and produces probabilistic statements about the effect of actions and the impact of observations. The calculus admits two types of conditioning operators: ordinary Bayes conditioning, $P(y|X = x)$, which represents the observation $X = x$, and causal conditioning, $P(y|do(X = x))$, read: the probability of $Y = y$ conditioned on holding X constant (at x) by deliberate action. Given a mixture of such observational and causal sentences, together with the topology of the causal graph, the calculus derives new conditional probabilities of both types, thus enabling one to quantify the effects of actions and observations.

2 ACTION AS A LOCAL SURGERY

What gives us the audacity to expect that actions should have neat and compact representations? Why did the authors of STRIPS [Fikes & Nilsson, 1971] and BURIDAN [Kushmerick et al., 1993] believe they could get away with such short specification for actions?

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, people predict the consequences of any given action without much hustle, and AI researchers are writing languages for actions as if it is a God given truth that action representation should be compact, elegant and meaningful. Why?

The paradigm I wish to explore in this paper is that these expectations are not only justified but, mainly, that once we understand the justification, we will be in better shape to craft effective representations for actions.

2.1 Mechanisms and surgeries

Why are the expectations justified? Because the actions we normally invoke in common reasoning tasks are *local surgeries*. The world consists of a huge number of autonomous and invariant linkages or mechanisms (to use Simon's word), each corresponding to a physical process that constrains the behavior of a relatively small groups of variables. In principle, then, the formalization of actions should not be difficult. 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 an action would be: Simply re-specify those few mechanisms that are perturbed by the action, then let the modified population of mechanisms interact with one another, and see what state will evolve at equilibrium. If the new specification is complete, a single state will evolve. If the specification is probabilistic, a new probability distribution will emerge and, if the specification is logical (possibly incomplete) a new, mutilated logical theory will then be created, capable of answering queries about post-action states of affair.

If this sounds so easy, why did AI ever get into trouble in the arena of action representation? The first answer I wish to explore is that 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. It is important therefore to emphasize 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 the spatial representation, 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 which normally keeps the left-most tile in a stable erect position. 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 *ith* 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.

2.2 Laws vs. facts

This surgical procedure still sounds easy and does not explain why AI got into trouble with action representation. The trouble begins with the realization that 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 mechanism which couples the state of the $(i+1)th$ domino tile to that of the *ith* domino tile remains unaltered (unless we set them apart by some action) whereas the states of the tiles themselves are free to vary with circumstances.

Admitting the need for this distinction has been a difficult cultural transition in the logical approach to actions, 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., $P(accident|careless\text{-}driving) = high$), 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 allow 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.[1]

2.3 Mechanisms and causal relationships

The next issue worth discussing is how causality enters into this surgical representation of actions. To understand the role of causality, we should note that 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". And there is really no need for the struggle; instead of telling you the name of the mechanism to be perturbed by the action, I might as well gloss over the details of the perturbation process and summarize its net result in the form of an *event*, e.g., "the left-most tile is tipped to right", which yields equivalent consequences as the perturbation summarized. After all, if you and I share the same understanding of physics, you should be able to figure out for yourself which mechanism it is that must be perturbed in order to realize 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.[2] 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 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.

A formulation of action as a local surgery on causal theories has been developed in a number of recent papers [Goldszmidt & Pearl, 1992; Pearl, 1993a; Pearl, 1993b; Darwiche & Pearl, 1994; Pearl, 1994a; Goldszmidt & Darwiche, 1994]. Section 3.1 provides a brief summary of this formulation, together with a simple example that illustrates how the surgery semantics generalizes to nonprobabilistic formalisms.

2.4 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, having an effect) is computationally very

[1]The distinction between laws and facts has been proposed by Poole (1985) and Geffner (1992) as a fundamental principle for nonmonotonic reasoning. It seems to be gaining broader support recently as a necessary requirement for formulating actions.

[2]The word "needed" connotes minimality and can be translated to: "...if every minimal perturbation realizing A, entails B".

useful, but, at the same time it requires that the assembly of mechanisms in our domain satisfy certain conditions. Some of these conditions are structural, nicely formulated in Simon's "causal ordering" [Simon, 1953], and others are substantive – – invoking relative magnitudes of forces and powers.

The structural requirement is that there be a one-to-one correspondence between mechanisms and variables – a unique variable in each mechanism is designated as the output (or effect), and the other variables, as inputs (or causes). Indeed, the formal definition of causal theories given in Section 3.1 assumes that each equation is associated with a unique variable, situated on its left hand side. In general, a mechanism may be specified as a function

$$G_i(X_1, ..., X_n; U_1, \ldots, U_m) = 0$$

without identifying any so called "dependent" variable X_i. Simon's causal ordering provides a procedure for deciding whether a collection of such G functions has a unique preferred way of associating variables with mechanisms, based on the requirement that we should be able to solve for the *ith* variable without solving for its successors in the ordering.

In certain structures, called *webs* [Dalkey, 1994, Dechter & Pearl, 1991], Simon's causal ordering determines a unique one-to-one correspondence, but in others, such as those involving feedback, the correspondence is not unique. Yet in examining feedback circuits, for example, people can assert categorically that the flow of causation goes clockwise, rather than counterclockwise. They make such assertions on the basis of relative magnitudes of forces; for example, it takes very little energy to make an input of a gate change its output, but no force applied to the output can influence the input. When such considerations are available, causal directionality can be 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. The variable which does not affect any of the others is the dependent variable. This then constitutes the operational semantics for identifying the dependent variables X_i in nonrecursive causal theories (see Section 3.1).

2.5 Imaging vs. conditioning

If action is a transformation from one probability function to another, one may ask whether every transformation corresponds to an action, or are there some constraints that are peculiar to exactly those transformations that originate from actions. Lewis (1976) formulation of counterfactuals indeed identifies such constraints: the transformation must be an *imaging* operator (Imaging is the probabilistic version of Winslett-Katsuno-Mendelzon possible worlds representation of "update").

Whereas Bayes conditioning $P(s|e)$ transfers the entire probability mass from states excluded by e to the remaining states (in proportion to their current $P(s)$), imaging works differently; each excluded state s transfers its mass individually to a select set of states $S^*(s)$, which are considered "closest" to s. The reason why imaging is a more adequate representation of transformations associated with actions can be seen more clearly through a representation theorem due to Gardenfors [1988, Theorem 5.2 pp.113] (strangely, the connection to actions never appears in Gardenfors' analysis). Gardenfors' theorem states that a probability update operator $P(s) \rightarrow P_A(s)$ is an imaging operator iff it preserves mixtures, i.e.,

$$[\alpha P(s) + (1-\alpha)P'(s)]_A = \alpha P_A(s) + (1-\alpha)P'_A(s) \tag{1}$$

for all constants $1 > \alpha > 0$, all propositions A, and all probability functions P and P'. In other words, the update of any mixture is the mixture of the updates[3].

This property, called homomorphism, is what permits us to specify actions in terms of *transition probabilities*, as it is usually done in stochastic control and Markov decision process. Denoting by $P_A(s|s')$ the probability resulting from acting A on a known state s', homomorphism (1) dictates:

$$P_A(s) = \sum_{s'} P_A(s|s')P(s') \tag{2}$$

saying that, whenever s' is not known with certainty, $P_A(s)$ is given by a weighted sum of $P_A(s|s')$ over s', with the weight being the current probability function $P(s')$.

This characterization, however, is too permissive; while it requires any action-based transformation to be describable in terms of transition probabilities, it also accepts any transition probability specification, however whimsical as a descriptor of some action. The valuable information that actions are defined as *local* surgeries, is totally ignored in this characterization. For example, the transition probability associated with the atomic action $A_i = do(X_i = x_i)$ originates from the deletion of just one mechanism in the assembly. Hence, one would expect that the transition probabilities associated with the set of atomic actions would not be totally arbitrary but would constrain one another.

An axiomatic characterization of such constraints is formulated in [Galles & Pearl, 1996], aiming toward a logic of action-based modalities, as in: "X affects Y when we hold Z fixed". With such logic one should be able to derive, refute or confirm theorems such as "If X has no effect on Y and Z affects Y, then Z will continue to affect Y when we fix X." The reader might find some challenge proving or refuting the sentence above, that is, testing whether it holds in every causal theory, when "affecting" and "fixing" are interpreted by the local-surgery semantics described in this paper.

3 FORMAL UNDERPINNING

3.1 Causal theories and actions

Definition 1 *A causal theory is a four-tuple*

$$T = < V, U, P(u), \{f_i\} >$$

where

(i) $V = \{X_1, \ldots, X_n\}$ *is a set of observed variables,*

(ii) $U = \{U_1, \ldots, U_m\}$ *is a set of exogenous (often unmeasured) variables that represent disturbances, abnormalities, or assumptions,*

(iii) $P(u)$ *is a distribution function over* $U_1, \ldots, U_m$*, and*

(iv) $\{f_i\}$ *is a set of* n *deterministic functions, each of the form*

$$X_i = f_i(PA_i, u) \quad i = 1, \ldots, n \tag{3}$$

[3] Assumption (1) is reflected in the (U8) postulate of [Katsuno & Mendelzon, 1991]: $(K_1 \vee K_2)o\mu = (K_1 o\mu) \vee (K_2 o\mu)$, where o is an update operator.

where PA_i is a subset of variables in V not containing X_i.

We will assume that the set of equations in (iv) has a unique solution for $X_i, \ldots, X_n$, given any value of the disturbances $U_1, \ldots, U_m$. Therefore, the distribution $P(u)$ induces a unique distribution on the observables, which we denote by $P_T(v)$. The variables PA_i (connoting "parents") are considered the direct causes of X_i and they define a directed graph G which may, in general, be cyclic. However, unlike the standard definition of "parents" in Bayesian networks [Pearl, 1988], PA_i is selected from V by considering outcomes of manipulative experiments (according to Lemma 1 below), not by conditional independence considerations. Encoding this manipulative information in the equations enabling reasoning about *implicit actions*, i.e., actions that were not anticipated by the modeller.

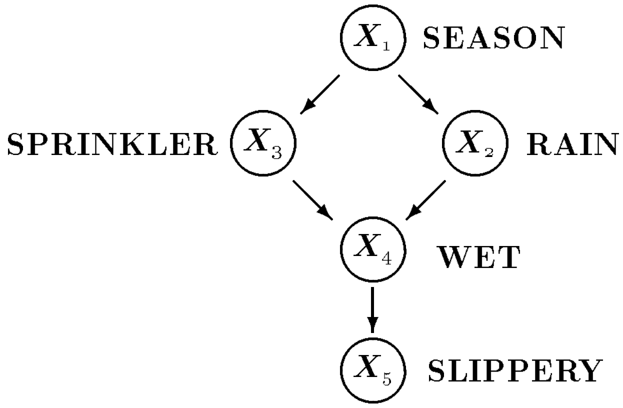

Figure 1: A Bayesian network representing causal influences among five variables.

Consider the example depicted in Figure 1. The corresponding theory consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= f_2(X_1, U_2) \\ X_3 &= f_3(X_1, U_3) \\ X_4 &= f_4(X_3, X_2, U_4) \\ X_5 &= f_5(X_4, U_5) \end{aligned} \tag{4}$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in the graph of Figure 1, but are understood to govern the uncertainties associated with the causal relationships. A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean theory below:

$$\begin{aligned} x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\ x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\ x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\ x_5 &= (x_4 \vee ab_5) \wedge \neg ab'_5 \end{aligned} \tag{5}$$

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for triggering and inhibiting abnormalities.[4] For example, ab_4 stands for (unspecified) events which might cause the

[4] Goldszmidt and Pearl (1992, 1995) describe a qualitative method of causal analysis based on attributing infinitesimal probabilities to the ab predicates.

ground to get wet (x_4) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$), while $\neg ab'_4$ stands for events which will keep the ground dry despite the rain, the sprinkler and ab_4, say covering the ground with plastic sheet.

Now consider local concurrent actions of the form $do(X = x)$, where $X \subseteq V$ is a set of variables and x is a set of values from the domain of X. In other words, $do(X = x)$ represents a combination of direct actions that forces the variables in X to attain the values x.

Definition 2 (effect of actions) *The effect of the action $do(X = x)$ on a causal theory T is given by a subtheory T_x of T, where T_x obtains by deleting from T all equations corresponding to variables in X and substituting the equations $X = x$ instead.*

For example, to represent the action "turning the sprinkler ON," $do(X_3 = \text{ON})$, we delete the equation $X_3 = f_3(X_1, U_3)$ from the theory of Eq. (4), and replace it with $X_3 = \text{ON}$. The resulting subtheory, $T_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on other variables. It is easy to see from this subtheory that the only variables affected by the action are X_4 and X_5, that is, the descendants, of the manipulated variable X_3. This is to be expected, since nondescendants of X_3 (i.e., season and rain) are presumed to be causally irrelevant to X_3. Note, however, that the operation $do(X_3 = \text{ON})$ stands in marked contrast to the operation of *finding* $X_3 = \text{ON}$ which may potentially influence (the belief in) every variable in the network. The mathematics underlying these two operations, and the conditions that enable us to predict the effects of actions without specifying $\{f_i\}$, will be discussed in the next two subsections.

Definition 2 should be taken as an integral part of Definition 1, because it assigns meaning to each individual equation in T. Specifically, it dictates what hypothetical experiments of the type $do(X = x)$ must be considered by the author of the structural equations in deciding which variables PA_i should enter into the r.h.s of each equation. By writing $X_4 = f_4(X_2, X_3, u)$, for example, the analyst defines X_2 and X_3 as the direct causes of X_4 which, according to Definition 2, means that holding X_2 and X_3 fixed determines the value of X_4 regardless of changes in the season (X_1) and regardless of any direct action we might take to make the ground slippery (X_5). In general, Definition 2 endows PA_i with the following meaning: PA_i is a set of variables that, if held fixed, would determine (for any u) the value of X_i regardless of any other action $do(Z = z)$ that one may perform, where Z is any set of variables not containing X_i or any member of PA_i. Moreover, no proper subset of PA_i possesses that quality.

Lemma 1 provides a succinct summary of this property, and can also be viewed as the structural definition of direct causes.

Lemma 1 *Let $Y(x; u)$ stand for the solution of Y under subtheory T_x, as in Definition 2. The direct causes of variable X_i are the minimal set of variables PA_i which satisfy*

$$X_i(pa_i, z; u) = X_i(pa_i; u) \tag{6}$$

for every u and for every set Z not containing X_i or any member of PA_i. (pa_i denotes a specific instantiation of PA_i.)

Clearly, if a causal theory is given explicitly, as in Definition 1, then the direct causes PA_i can be identified syntactically, as the arguments of each f_i. However, if the theory is represented implicitly in a form of a function F: Actions $\times\, U \rightarrow V$, (as is often assumed in decision theory [Savage, 1954, Heckerman & Shachter, 1995], then Lemma 1 can be used to identify, given F, the unique set of direct causes for each variable X_i.[5]

We see that the distinctive characteristic of structural equations, which sets them apart from ordinary algebraic equations, is that meaning is attached to any subset of equations from T. Mathematically, this characteristic does not show up explicitly in the equations, but rather implicitly, in the understanding that T stands for not one but 2^n sets of equations. This restricts, of course, the type of algebraic transformations admissible on T to those that preserve the solution of not one but each of the 2^n sets.

The framework provided by Definitions 1 and 2 permits the coherent formalization of many nuances and subtle concepts found in causal conversation, including causal influence, causal effect, causal relevance, average causal effect, identifiability, counterfactuals, and exogeneity. Examples are:

- X **influences** Y **in context** u if there are two values of X, x and x', such that $Y(x;u) \neq Y(x';u)$. In other words, the solution for Y under $U=u$ and $do(X=x)$ is different from the solution under $U=u$ and $do(X=x')$.

  We say, for example, that the weather (X_2) influences the wetness of the pavement (X_4) in a context u where the pavement is uncovered, and the sprinkler controller is at off position, because a change in weather from not-rain to rain is accompanied with a change in pavement condition from dry to wet. This definition interprets causal influence as the transference of change from X to Y triggered by the local intervention $do(X=x)$. Although the word "influence" is sometimes used with no intervention in mind (as in the case of the weather), the hypothetical operator $do(X=x)$ ensures that the change in Y is attributable only to changes in X, and not to spurious side effects (e.g., strange people who turn their sprinklers on whenever they see clouds in the sky.)

- X **can potentially influence** Y in context $U=u$ if there exists a subtheory T_z of T in which X influences Y.

  The difference between influence and potential influence is that the latter requires an additional intervention, $do(Z=z)$, to reveal the effect of X on Y. In our earlier example, we find it plausible to maintain that, although the weather does not influence wetness in a context (u) where the sprinkler controller is stuck at ON position, it nevertheless can potentially influence wetness, at u, as is revealed when the action $do(Sprinkler = \text{OFF})$ is implemented, say, by manual intervention. Along the same vein, we may say that seasonal variations (X_1) have potential influence on wetness, even though their influence through rain may perfectly cancel their influence through sprinkler; This potential would surface when we hold *Sprinkler* fixed (at either ON or OFF position).[6]

[5] Likewise, the local operator $do(X_i = x_i)$ can be identified from F as the unique action A for which the equality $F(A,u)_i = F(A \text{ and } B, u)_i$ holds for every action B compatible with A. In words, $do(X_i = x_i)$ is the only action which keeps the value of X_i invariant to any other action that can be implemented at the same time.

[6] The standard example in the philosophical literature [Cartwright, 1989] involves the potential positive influence of birth-control pills on thrombosis, which might be masked by its negative effect on pregnancy (another cause of thrombosis). Cartwright proposal (rejected by Eells), that the influence of the pill be as-

- **Event $X = x$ is the (singular) cause of event $Y = y$** if (i) $X = x$ and $Y = y$ are true and (ii) in every context u compatible with $X = x$ and $Y = y$, and for all $x' \neq x$, we have $Y(x'; u) \neq y$.

  This definition reflects the counterfactual explication of a singular cause: "$Y = y$ would be false if it were not for $X = x$". A separate analysis of counterfactuals will be given in Section 4.

3.2 Probabilistic causal effects and identifiability

The definitions above are deterministic. Probabilistic causality emerges when we define a probability distribution $P(u)$ for the U variables. Under the assumption that the set of equations $\{f_i\}$ and every subset thereof has a unique solution, $P(u)$ induces a unique distribution $P_{T_x}(v)$ on the endogenous variables for each combination of atomic interventions $do(X = x)$. This leads to a natural probabilistic definition of causal effects.

Definition 3 (causal effect) *Given two disjoint subsets of variables, $X \subseteq V$ and $Y \subseteq V$, the causal effect of X on Y, denoted $P_T(y|do(x))$ or $P_T(y|\hat{x})$, gives the distribution of Y induced by the action $do(X = x)$, that is,*

$$P_T(y|\hat{x}) = P_{T_x}(y) \tag{7}$$

for each realization x of X.

The probabilistic notion of causal effect is much weaker than its deterministic counterparts of causal influence and potential causal influence. For example, if U is the outcome of a fair coin, X is my bet, and Y stands for winning a dollar iff $X = U$, then the causal effect of X on Y is nil, because $P(y|do(X = Tail)) = P(y|do(X = Head) = \frac{1}{2}$. At the same time, X will qualify as having an influence on Y in every possible context, $U = Head$ and $U = Tail$. Note that causal effects are defined relative to a given causal theory T, though the subscript T is often suppressed for brevity.

Definition 4 (identifiability) *Let $Q(T)$ be any computable quantity of a theory T. Q is* identifiable *in a class M of theories if for any pairs of theories T_1 and T_2 from M, $Q(T_1) = Q(T_2)$ whenever $P_{T_1}(v) = P_{T_2}(v)$.*

Identifiability is essential for integrating statistical data (summarized by $P(v)$) with incomplete prior causal knowledge of $\{f_i\}$, as it enables the reasoner to estimate quantities Q from P alone, without specifying the details of T, so that the general characteristics of the class M suffice.[7] For the purpose of our analysis, the quantity Q of interest is the causal

sessed by considering separately the population of women that would get pregnant (or remain non-pregnant) regardless of the pill, amounts to considering a subtheory T_z in which pregnancy (Z) is held fixed.

[7] The notion of identifiability is central to much work in econometrics, where it has become synonymous to the identification of the functions $\{f_i\}$ or some of their parameters [Koopman & Reiersol, 1950], mostly under conditions of additive Gaussian noise. Definition 4, which does not assume any parametric representation of the functions $\{f_i\}$, extends the notion of identifiability to quantities Q that do not require the precision of parametric models. In particular, it permits one (see Definition 5) to dispose with the identification of functional parameters altogether, and deal directly with causal effects $P(y|\hat{x})$ – the very purpose of identifying parameters in policy-analysis applications.

effect $P_T(y|\hat{x})$ which is certainly computable from a given theory T (using Eq. (7)), but which we will now attempt to compute from incomplete specification of T, in the form of general characteristics such as the identities of the parent sets PA_i and the independencies embedded in $P(u)$. We will therefore consider a class M of theories which have the following characteristics in common:

(i) they share the same parent-child families (i.e., the same causal graph G),

(ii) they share the same set of independencies in $P(u)$, and,

(iii) they induce positive distributions on the endogenous variables,[8] i.e., $P(v) > 0$.

Relative to such classes we now define:

Definition 5 (causal-effect identifiability) *The* causal effect *of X on Y is said to be* identifiable *in M if the quantity $P(y|\hat{x})$ can be computed uniquely from the probabilities of the observed variables, that is, if for every pair of theories T_1 and T_2 in M such that $P_{T_1}(\mathrm{v}) = P_{T_2}(\mathrm{v})$, we have $P_{T_1}(y|\hat{x}) = P_{T_2}(y|\hat{x})$.*

The identifiability of $P(y|\hat{x})$ ensures that it is possible to infer the effect of action $do(X = x)$ on Y from two sources of information:

(i) passive observations, as summarized by the probability function $P(v)$,

(ii) the causal graph, G, which specifies, qualitatively, which variables make up the stable mechanisms in the domain or, alternatively, which variables participate in the determination of each variable in the domain.

Simple examples of identifiability will be discussed in the next subsection.

3.3 Inferring consequences of actions from passive observations

The probabilistic analysis of actions becomes particularly simple when two conditions are satisfied:

1. The theory is recursive, that is, there exists an ordering of the variables $V = \{X_1, \ldots, X_n\}$ such that each X_i is a function of a subset PA_i of its predecessors

$$X_i = f_i(PA_i, U_i) \qquad PA_i \subseteq \{X_1, \ldots, X_{i-1}\} \tag{8}$$

2. The disturbances $U_1, \ldots, U_n$ are mutually independent, which implies (from the exogeneity of the U_i's)

$$U_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \tag{9}$$

[8]This requirement ensures that the disturbances U are sufficiently rich to simulate a "natural experiment", that is, an experiment in which conditions change by natural phenomena rather than a human experimenter.

These two conditions, also called Markovian, are the basis of the independencies embodied in Bayesian networks [Pearl, 1988], and they enable us to compute causal effects directly from the conditional probabilities $P(x_i|pa_i)$, without specifying either the functional form of the functions f_i or the distributions $P(u_i)$ of the disturbances [Pearl, 1993a, Spirtes et al., 1993]. This is seen immediately from the following observations: On the one hand, the distribution induced by any Markovian theory T is given by the product,

$$P_T(x_1, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{10}$$

where pa_i are (values of) the parents of X_i in the diagram representing T. On the other hand, the subtheory $T_{x'_j}$, representing the action $do(X_j = x'_j)$, is also Markovian; hence, it also induces a product-like distribution

$$P_{T_{x'_j}}(x_1, \ldots, x_n) = \begin{cases} \prod_{i \neq j} P(x_i|pa_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|pa_j)} & \text{if } \; x_j = x'_j \\ 0 & \text{if } \; x_j \neq x'_j \end{cases} \tag{11}$$

where the partial product reflects the surgical removal of the equation $X_j = f_j(pa_j, U_j)$ from the theory of Eq. (8). Thus, we see that both the pre-action and the post-action distributions depend only on observed conditional probabilities but are independent of the particular functional form of $\{f_i\}$ and of the distributions $P(u)$ that generate those probabilities. This is the essence of identifiability as given in Definition 5, which stems from the Markovian assumptions (8) and (9). Section 3.4 will demonstrate that certain, though not all, causal effects are identifiable even when the Markovian property is destroyed by introducing dependencies among the disturbance terms.

In the example of Figure 1, the pre-action distribution is given by the product

$$P_T(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{12}$$

while the surgery corresponding to the action $do(X_3 = \text{ON})$ amounts to deleting the link $X_1 \rightarrow X_3$ from the graph and fixing the value of X_3 to ON, yielding the post-action distribution

$$P_T(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = \text{ON})\ P(x_5|x_4) \tag{13}$$

Note the difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$. The latter is encoded by ordinary Bayesian conditioning,

$$P_T(x_1, x_2, x_4, x_5|X_3 = \text{ON}) = \frac{P(x_1)\ P(x_2|x_1)\ P(x_3 = \text{ON}|x_1)P(x_4|x_2, X_3 = \text{ON})P(x_5|x_4)}{P(X_3 = \text{ON})}$$

The former is obtained by conditioning a mutilated graph, with the link $X_1 \rightarrow X_3$ removed. This mirrors indeed the difference between seeing and doing: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the deliberate action "turning the sprinkler ON." The excision of $X_3 = f_3(X_1, U_3)$ from (4) ensures the suppression of any abductive inferences from the action, as well as from any of its consequences.

Generalization to multiple actions and conditional actions is straightforward. Multiple actions $do(X = x)$, where X is a compound variable, result in a distribution similar to (11), except that all factors corresponding to the variables in X are removed from the product in (10). Stochastic conditional strategies [Pearl, 1994b] of the form

$$do(X_j = x_j) \text{ with probability } P^*(x_j|pa_j^*) \tag{14}$$

where PA_j^* is the support set of the decision strategy, also result in a product decomposition similar to (10), except that each factor $P(x_j|pa_j)$ is *replaced* with $P^*(x_j|pa_j^*)$.

3.4 A calculus of acting and seeing

The identifiability of causal effects demonstrated in Section 3.3 relies critically on the Markovian assumptions given in (8) and (9). If a variable that has two descendants in the graph is unobserved, the disturbances in the two equations are no longer independent, the Markovian property (8) is violated, and identifiability may be destroyed. This can be seen easily from Eq. (11); if any parent of the manipulated variable X_j is unobserved, one cannot estimate the conditional probability $P(x_j|pa_j)$, and the effect of the action $do(X_j = x_j)$ may not be predictable from the observed distribution $P(x_1, \ldots, x_n)$. Fortunately, certain causal effects are identifiable even in situations where members of pa_j are unobservable [Pearl, 1993a]. Moreover, polynomial tests are now available for deciding when $P(x_i|\hat{x}_j)$ is identifiable and for deriving closed-form expressions for $P(x_i|\hat{x}_j)$ in terms of observed quantities [Galles & Pearl, 1995].

These tests and derivations are based on a symbolic calculus [Pearl, 1994b, 1995], to be described in the sequel, in which interventions, side by side with observations, are given explicit notation and are permitted to transform probability expressions. The transformation rules of this calculus reflect the understanding that interventions perform "local surgeries" as described in Definition 2, namely, they overrule equations that tie the manipulated variables to their pre-intervention causes.

Definition 6 (d-separation) *Let a* path *in a DAG G stand for any sequence of consecutive edges (of any directionality) in G. A path p is said to be d-separated (or blocked) by a set of nodes Z iff:*

(i) p contains a chain $i \longrightarrow j \longrightarrow k$ or a fork $i \longleftarrow j \longrightarrow k$ such that the middle node j is in Z, or,

(ii) p contains an inverted fork $i \longrightarrow j \longleftarrow k$ such that neither the middle node j nor any of its descendants (in G) are in Z.

Similarly, if X, Y, and Z are three disjoint subsets of nodes in G, then Z is said to d-separate X from Y, denoted $(X \perp\!\!\!\perp Y)_G$, iff Z d-separates every path from a node in X to a node in Y.

We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$. Finally, the expression $P(y|\hat{x},z) \triangleq P(y,z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y=y$ given that $Z=z$ is observed and X is held constant at x.

Theorem 2 Let G be the DAG associated with a Markovian causal theory, and let $P(\cdot)$ stand for the probability distribution induced by that theory. For any disjoint subsets of variables X, Y, Z, and W we have:

Rule 1 Insertion/deletion of observations

$$P(y|\hat{x},z,w) = P(y|\hat{x},w) \;\text{ if }\; (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}} \tag{15}$$

Rule 2 Action/observation exchange

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},z,w) \;\text{ if }\; (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \tag{16}$$

Rule 3 Insertion/deletion of actions

$$P(y|\hat{x},\hat{z},w) = P(y|\hat{x},w) \;\text{ if }\; (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{17}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of the inference rules above follows from the basic interpretation of the $\hat{x}$ operator as a replacement of the causal mechanism that connects X to its pre-action parents by a new mechanism $X=x$ introduced by the intervening force.

Corollary 1 *A causal effect Q: $P(y_1,...,y_k|\hat{x}_1,...,\hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 2, which reduces q into a standard (i.e., hat-free) probability expression involving observed quantities.*

Although Theorem 2 and Corollary 1 require the Markovian property, they can also be applied to non-Markovian, recursive theories, because such theories become Markovian if we consider the unobserved variables as part of the analysis and represent them as nodes in the graph. To illustrate: Assume that variable X_1 in Figure 1 is unobserved, rendering the disturbances U_3 and U_2 dependent since these terms now include the common influence of X_1. Theorem 2 tells us that the causal effect $P(x_4|\hat{x}_3)$ is identifiable, because

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|\hat{x}_3,x_2)P(x_2|\hat{x}_3) \tag{18}$$

Rule 3 permits the deletion

$$P(x_2|\hat{x}_3) = P(x_2) \tag{19}$$

because $(X_2 \perp\!\!\!\perp X_3)_{G_{\overline{X}_3}}$, while Rule 2 permits the exchange

$$P(x_4|\hat{x}_3,x_2) = P(x_4|x_3,x_2) \tag{20}$$

because $(X_4 \perp\!\!\!\perp X_3|X_2)_{G_{\underline{X_3}}}$. This gives

$$P(x_4|\hat{x}_3) = \sum_{x_2} P(x_4|x_3, x_2)P(x_2) \tag{21}$$

which is a hat-free expression, involving only observed quantities.

The reader might recognize Eq. (21) as the standard formula for covariate adjustment (also called "stratification"), which is used in experimental design both for improving precision and for minimizing confounding bias. However, a formal, general criterion for deciding whether a set of covariates Z (X_2 in our example) qualifies for adjustment has long been wanting [Smith, 1957, Wainer, 1991, Shafer, 1995].[9] Theorem 2 provides such a criterion (called the "back-door criterion" in [Pearl, 1993a]) which reads:

Definition 7 *Z is an admissible set of covariates relative to the effect of X on Y if:*

(i) *no node in Z is a descendant of X, and*

(ii) *Z d-separates X from Y along any path containing an arrow into X (equivalently, $(Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}}}$).*

We see, for instance, that X_2 and X_1 (or both) qualify as admissible covariates relative the effect of X_3 on X_4, but X_5 will not qualify. The graphical definition of admissible covariates replaces statistical folklore with formal procedures, and should enable analysts to systematically select an optimal set of observations, namely, a set Z that minimizes measurement cost or sampling variability.

In general, it can be shown [Pearl, 1995] that:

1. The effect of interventions can often be identified (from nonexperimental data) without resorting to parametric models.

2. The conditions under which such nonparametric identification is possible can be determined by simple graphical criteria.

3. When the effect of interventions is not identifiable, the causal graph may suggest nontrivial experiments which, if performed, would render the effect identifiable.

While the ability to assess the effect of interventions from nonexperimental data has immediate applications in the medical and social sciences, such assessments are also important in learning theory: they explain how agents can predict the effect of the next action (e.g.,turning the sprinkler on) on the basis of past experience, where that action has never been enacted out of free will, but only in response to environmental needs (e.g., dry season) or to other agents' requests.

[9] Most of the statistical literature is satisfied with informal warnings that "Z should be quite unaffected by X" [Cox, 1958, page 48], which is necessary but not sufficient, or that X should not precede Z [Shafer, 1995, page 294], which is neither necessary nor sufficient. In some academic circles, a criterion called "ignorability" is invoked [Rosenbaum & Rubin, 1983], which merely paraphrases the problem in the language of counterfactuals. Simplified, it reads: Z is an admissible covariate relative to the effect of X on Y if, for every x, the value that Y would obtain had X been x is conditionally independent of X, given Z.

4 PROCESSING COUNTERFACTUALS

A counterfactual sentence has the form

If A were true, then C would have been true, given O

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations O, and C, the counterfactual consequent, specifies a result that is expected to hold in an alternative world where the antecedent is true. A typical example is "If Oswald were not to have shot Kennedy, then Kennedy would still be alive," which presumes the factual knowledge of Oswald's assassination of Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences have resorted to some version of Lewis' "closest world" approach: "C if it were A" is true, if C is true in worlds that are "closest" to the real world yet consistent with the counterfactual antecedent A [Lewis, 1973]. While the closest world approach leaves the precise specification of the closeness measure almost unconstrained, causal knowledge imposes very specific preferences as to which worlds should be considered closest to any given world. For example, consider the array of domino tiles discussed in Section 2 close to each other. The manifestly closest world consistent with the statement "tile i is tipped to the right" would be a world in which just tile i is tipped, while all the others remain erect. Yet, we all accept the counterfactual sentence "Had tile i been tipped to the right, tile $i+1$ would be tipped as well" as plausible and valid. Thus, distances among worlds are not determined merely by surface similarities but require a distinction between explained and unexplained dissimilarities. The local surgery paradigm expounded in Section 3.1 offers a concrete explication of the closest-world approach which respects such causal considerations. A world w_1 is "closer" to w than a world w_2 is, if the set of atomic surgeries needed for transforming w into w_1 is a proper subset of those needed for transforming w into w_2. In the domino example, finding tile i tipped and $i+1$ erect requires the alteration of two basic mechanisms (i.e., two unexplained actions or "miracles" [Lewis, 1973]) compared with one altered mechanism for the world in which all j tiles, $j > i$, are tipped. This paradigm conforms to our perception of causal influences and lends itself to economical machine representation.

4.1 Formal underpinning

The structural equations framework, coupled with the surgical operator $do(X = x)$, also offers the syntactic machinery for counterfactual analysis, while leaving the closest-world interpretation implicit. The basis for this analysis is the potential response function $Y(x; u)$ invoked in Lemma 1, which we take as the formal explication of the English phrase "the value that Y would obtain in context u, had X been x".

Definition 8 (potential response) *Given a causal theory T the potential response of Y to X in a context u, denoted $Y(x; u)$ or $Y_x(u)$, is the solution for Y under $U = u$ in the subtheory T_x.* [10]

[10]The term *unit* instead of *context* is often used in the statistical literature [Rubin, 1974], where it normally stands for the identity of a specific individual in a population, namely, the set of attributes u that characterize

Note that this definition allows for the context $U = u$ and the proposition $X = x$ to be incompatible in T. For example, if T describes a logic circuit with input U, output Y, and an intermediate variable X, it may well be reasonable to assert the counterfactual: "Given $U = u$, voltage Y would be high if current X were low," even though the input $U = u$ may preclude X from being low. It is for this reason that one must invoke some notion of intervention (alternatively, a theory change or a "miracle"[Lewis, 1973]) in the definition of counterfactuals. This is further attested by the suppression of abductive arguments in counterfactual reasoning; for example, the following sentence would be deemed unacceptable: "Had I done my homework, I would have felt miserable, because I always do my homework after my father beats me up." The reason we do not accept this argument is that it conflicts with the common understanding that the counterfactual antecedent "done my homework" should be considered an external willful act, totally free of normal inducements (e.g., beatings), as modeled by the surgical subtheory T_x.

Counterfactual sentences rarely specify a complete context u. Instead they imply a partial description of u in the form of a set o of (often implicit) facts or observations. Thus, a general counterfactual sentence would have the format $x \rightarrow y|o$, read "Given factual knowledge o, Y would obtain the value y had X been x." For example, the sentence "If Oswald were not to have shot Kennedy, then Kennedy would still be alive" would be formulated:

$$\neg Shot(Oswald, Kennedy) \rightarrow Alive(Kennedy) \mid Dead(Kennedy), Shot(Oswald, Kennedy)$$

The truth of such a sentence in a theory T can be defined in terms of the potential response $Y(x;u)$ as follows:

Definition 9 (counterfactual assertability) *The sentence $x \rightarrow y|o$ is true in T if $Y(x;u) = y$ for every u compatible with o.*

This definition parallels Lewis's closest world approach, with u playing the role of a possible world. Note the difference between the treatments of o and x; the former insists on direct compatibility between u and o, while the latter tolerates a surgical face-lift where x and u are incompatible.

If U is treated as a random variable, then the value of the counterfactual $Y(x;u)$ becomes a random variable as well, denoted $Y(x)$ or Y_x. Moreover, the distribution of this random variable is easily seen to coincide with the causal effect $P(y|\hat{x})$:

$$P((Y(x) = y) = P(y|\hat{x})$$

Thus, the probability of a counterfactual conditional $x \rightarrow y \mid o$ may be evaluated by the following procedure:

- Use the observations o to update $P(u)$, thus forming a revised causal theory $T^o = < V, U, \{f_i\}, P(u|o) >$

- Form the mutilated theory T^o_x (by deleting from T^o the equation corresponding to variables in X) and compute the probability $P_{T^o}(y|\hat{x})$ that T^o_x induces on Y.

that individual. In general, u may include the time of day, the experimental conditions under study, and so on. Practitioners of the counterfactual notation do not explicitly mention the notions of "solution" or "intervention" in the definition of $Y(x;u)$. Instead, the phrase "the value that Y would take in unit u, had X been x," viewed as basic, is posited as the definition of $Y(x;u)$.

Unlike causal-effect queries, counterfactual queries may not be identifiable in Markovian theories, but require that the functional form of $\{f_i\}$ be specified. In [Balke & Pearl, 1994], a method is devised for computing sharp bounds on counterfactual probabilities, and, under certain circumstances, those bounds may collapse to point estimates. This method has been applied to the evaluation of causal effects in studies involving noncompliance and to determination of legal liability.

4.2 Applications to Policy Analysis

Counterfactual reasoning is at the heart of every planning activity, especially real-time planning. When a planner discovers that the current state of affairs deviates from the one expected, a "plan repair" activity need be invoked to determine what went wrong and how it could be rectified. This activity amounts to an exercise of counterfactual thinking, as it calls for rolling back the natural course of events and determining, based on the factual observations at hand, whether the culprit lies in previous decisions or in some unexpected, external eventualities. Moreover, in reasoning forward to determine if things would have been different a new model of the world must be consulted, one that embodies hypothetical changes in decisions or eventualities, hence, a breakdown of the old model or theory.

The logic-based planning tools used in AI, such as STRIPS and its variants or those based on the situation calculus, do not readily lend themselves to counterfactual analysis; as they are not geared for coherent integration of abduction with prediction, and they do not readily handle theory changes. Remarkably, the formal system developed in economics and social sciences under the rubric "structural equations models" does offer such capabilities but, as will be discussed below, these capabilities are not well recognized by current practitioners of structural models. The analysis presented in this paper could serve both to illustrate to AI researchers the basic formal features needed for counterfactual and policy analysis, and to call the attention of economists and social scientists to capabilities that are dormant within structural equations models.

Counterfactual thinking dominates reasoning in political science and economics. We say, for example, "If Germany were not punished so severely at the end of World War I, Hitler would not have come to power," or "If Reagan did not lower taxes, our deficit would be lower today." Such thought experiments emphasize an understanding of generic laws in the domain and are aimed toward shaping future policy making, for example, "defeated countries should not be humiliated," or "lowering taxes (contrary to Reaganomics) tends to increase national debt."

Strangely, however, there is very little formal work on counterfactuals or even policy analysis in the behavioral science literature. An examination of a number of econometric journals and textbooks, for example, reveals a glaring imbalance: while an enormous mathematical machinery is brought to bear on problems of estimation and prediction, policy analysis (which is the ultimate goal of economic theories) receives almost no formal treatment. Not surprisingly, the methods currently used for economic policy making are grossly inadequate, and are based on so-called *reduced-form* analysis: to find the impact of a policy involving decision variables X on outcome variables Y, one examines past data and estimates the conditional expectation $E(Y|X{=}x)$, where x is the particular instantiation of X under the policy studied.

The assumption underlying this method is that the data were generated under circumstances in which the decision variables X act as exogenous variables, that is, variables whose values are determined outside the system under analysis. However, while new decisions should indeed be considered exogenous for the purpose of evaluation, past decisions are rarely enacted in an exogenous manner. Almost every realistic policy (e.g., taxation) imposes control over some endogenous variables, that is, variables whose values are determined by other variables in the analysis. Let us take taxation policies as an example. Economic data are generated in a world in which the government is reacting to various indicators and various pressures; hence, taxation is endogenous in the data-analysis phase of the study. Taxation becomes exogenous when we wish to predict the impact of a specific decision to raise or lower taxes. The reduced-form method is valid only when past decisions are non-responsive to other variables in the system, and this, unfortunately, eliminates most of the interesting control variables (e.g., tax rates, interest rates, quotas) from the analysis.

This difficulty is not unique to economic or social policy making; it appears whenever one wishes to evaluate the merit of a plan on the basis of the past performance of other agents. Even when the signals triggering the past actions of those agents are known with certainty, a systematic method must be devised for selectively ignoring the influence of those signals from the evaluation process. In fact, the very essence of *evaluation* is having the freedom to imagine and compare trajectories in various counterfactual worlds, where each world or trajectory is created by a hypothetical implementation of a policy that is free of the very pressures that compelled the implementation of such policies in the past.

Balke and Pearl (1995) demonstrate how the common economical model of linear, non-recursive equations with Gaussian noise can be used to compute counterfactual queries of the type: "Given an observation set O, find the probability that Y would have attained a value greater than y, had X been set to x". The task of inferring "causes of effects", that is, of finding the probability that $X = x$ is *the* cause for effect E, amounts to answering the counterfactual query: "Given effect E and observations O, find the probability that E would not have been realized, had X not been x". The technique developed in Balke and Pearl (1995) is based on probability propagation in dual networks, one representing the actual world, the other the counterfactual world. The method is not limited to linear functions but applies whenever we are willing to assume the functional form of the structural equations. For example, causal theories based on Boolean functions (with exceptions), such as the one described in Eq. (5) lend themselves to counterfactual analysis in the framework of Definition 8.

References

[Adams, 1975] Adams, E., *The Logic of Conditionals,* Chapter 2, D. Reidel, Dordrecht, Netherlands, 1975.

[Balke & Pearl, 1994] Balke, A. and Pearl, J., "Counterfactual probabilities: Computational methods, bounds, and applications," in R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 46-54, July 29-31, 1994.

[Balke & Pearl, 1995] Balke, A. and Pearl, J., "Counterfactuals and Policy Analysis in Structural Models," in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 11-18, 1995.

[Cartwright, 1989] Cartwright, N., *Nature's Capacities and Their Measurement*, Clarendon Press, Oxford, 1989.

[Cox, 1958] Cox, D.R., *The Planning of Experiments*, John Wiley and Sons, NY, 1958.

[Dalkey, 1994] Dalkey, N., "Webs," UCLA Cognitive Systems Laboratory, *Technical Report (R-166),* Computer Science Department, University of California, Los Angeles, March 1994.

[Darwiche & Pearl, 1994] Darwiche, A., and Pearl, J., "Symbolic causal networks for planning under uncertainty," In *Symposium Notes of the 1994 AAAI Spring Symposium on Decision-Theoretic Planning*, Stanford, CA, 41-47, March 21-23, 1994.

[Dechter & Pearl, 1991] Dechter, R. and Pearl, J., "Directed constraint networks: A relational framework for Causal Modeling," in *Proceedings, 12th International Joint Conference of Artificial Intelligence (IJCAI-91)*, Sydney, Australia, 1164-1170, August 24-30, 1991,

[Fikes & Nilsson, 1971] Fikes, R.E. and Nilsson, N.J., "STIRPS: A new approach to the application of theorem proving to problem solving," *Artificial Intelligence* 2(3/4), 189–208, 1971.

[Galles & Pearl, 1995] Galles, D. and Pearl, J., "Testing Identifiability of Causal Effects," in P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence 11*, Morgan Kaufmann, San Francisco, CA, 185–195, 1995.

[Galles & Pearl, 1996] Galles, D. and Pearl, J., "Axioms of Causal Relevance," UCLA Computer Science Department, Technical Report R-240, January 1996. Submitted to *Artificial Intelligence.*

[Geffner, 1992] Geffner, H.A., *Default Reasoning: Causal and Conditional Theories*, MIT Press, Cambridge, MA, 1992.

[Goldszmidt & Darwiche, 1994] Goldszmidt, M. and Darwiche, A., "Action networks: A framework for reasoning about actions and change under uncertainty," in R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 136–144, 1994.

[Goldszmidt & Pearl, 1992] Goldszmidt, M. and Pearl, J., "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 (Eds.), *Proceedings of the Third International Conference on Knowledge Representation and Reasoning,* Morgan Kaufmann, San Mateo, CA, 661-672, October 1992.

[Heckerman & Shachter, 1995] Heckerman, D. and Shacter, R., "A definition and graphical representation for causality," in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, 262–273, 1995.

[Katsuno & Mendelzon, 1991] Katsuno, H. and Mendelzon, A., "On the difference between updating a knowledge base and revising it," in *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Boston, MA, 387–394, 1991.

[Koopman & Reiersol, 1950] Koopman, T.C. and Reiersol, O., "The identification of structural characteristics," *Annals of Mathematical Statistics*, 21, 165–181, 1950.

[Kushmerick et al., 1993] Kushmerick, N., Hanks, S., and Weld, D., "An algorithm for probabilistic planning," *Technical Report 93-06-03*, Department of Computer Science and Engineering, University of Washington, 1993.

[Lewis, 1973] Lewis D., *Counterfactuals*, Basil Blackwell, Oxford, UK, 1973.

[Lewis, 1976] Lewis, D., "Probabilities of conditionals and conditional probabilities," *Philosophical Review*, 85, 297–315, 1976.

[Levi, 1988] Levi, I., "Iteration of conditionals and the Ramsey test," *Synthese*, 76, 49–81, 1988.

[Pearl, 1988] Pearl, J., *Probabilistic Reasoning in Intelligence Systems*, Morgan Kaufmann, San Mateo, CA, 1988.

[Pearl, 1993a] Pearl, J., "From Conditional Oughts to Qualitative Decision Theory" in D. Heckerman and A. Mamdani (Eds.), *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, Washington, D.C., Morgan Kaufmann, San Mateo, CA, 12–20, July 1993.

[Pearl, 1993b] Pearl, J., "Graphical models, causality, and intervention," *Statistical Science*, 8 (3), 266–273, 1993.

[Pearl, 1994a] Pearl, J., "From Adams' conditionals to default expressions, causal conditionals, and counterfactuals," in E. Eells and B. Skyrms (Eds.), *Probability and Conditionals*, Cambridge University Press, New York, NY, 47-74, 1994.

[Pearl, 1994b] Pearl, J., "A probabilistic calculus of actions," in R. Lopez de Mantaras and D. Poole (Eds.), *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*, Morgan Kaufmann, San Mateo, CA, 454-462, 1994.

[Pearl, 1995] Pearl, J., "Causal diagrams for experimental research," UCLA Computer Science Department, Technical Report (R-218-B), March 1995. To appear in *Biometrika*, December, 1995.

[Poole, 1985] Poole, D., "On the comparison of theories: Preferring the most specific explanations," in *Proceedings of International Conference on Artificial Intelligence (IJCAI-85)*, Los Angeles, CA, 144–147, 1985.

[Rosenbaum & Rubin, 1983] Rosenbaum, P. and Rubin, D., "The central role of propensity score in observational studies for causal effects," *Biometrica*, 70, 41–55, 1983.

[Rubin, 1974] Rubin, D.B., "Estimating causal effects of treatments in randomized and non-randomized studies," *Journal of Educational Psychology*, 66, 688–701, 1974.

[Savage, 1954] Savage, L.J., *The Foundations of Statistics*, John Wiley and Sons, Inc., New York, 1954.

[Shafer, 1995] Shafer, G., *The Art of Causal Conjecture*, MIT Press, Cambridge, MA, 1995. Forthcoming.

[Simon, 1953] Simon, H., "Causal ordering and identifiability," in W.C. Hood and T.C. Koopmans (Eds.), *Studies in Econometric Method*, New York, NY, Chapter 3, 1953.

[Smith, 1957] Smith, H.F., "Interpretation of adjusted treatment means and regressions in analysis of covariates," *Biometrics*, 13, 282–308, 1957.

[Spirtes et al., 1993] Spirtes, P., Glymour, C., and Schienes, R., *Causation, Prediction, and Search*, Springer-Verlag, New York, 1993.

[Wainer, 1991] Wainer, H., "Adjusting for differential base-rates: Lord's paradox again," *Psychological Bulletin*, 109, 147–151, 1991.