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 deadends 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-Robin 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.

We are currently exploring axiomatic characterizations of such constraints which we hope to use as a logic for sentences of the type: "X affects Y when we hold Z fixed". With the help of such logic we hope to be able to derive, refute or confirm sentences 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, actions, causal effect, and identifiability

Definition 1 *A causal theory is a 4-tuple*

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

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 unobserved variables which represent disturbances, abnormalities or assumptions,*

(iii) $P(\mathbf{u})$ *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*

$$X_i = f_i(X_1, \ldots, X_n, U_1, \ldots, U_m) \quad i = 1, \ldots, n \tag{3}$$

[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.

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 $P(\mathbf{u})$ induces a unique distribution on the observables, which we denote by $P_T(\mathbf{v})$.

We will consider 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 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.*

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 $P_T(y|\hat{x})$, is a function from the domain of X to the space of probability distributions on Y, such that*

$$P_T(y|\hat{x}) = P_{T_x}(y) \tag{4}$$

for each realization x of X. In other words for each $x \in dom(X)$, the causal effect $P_T(y|\hat{x})$ gives the distribution of Y induced by the action $do(X = x)$.

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) *The causal effect of X on Y is said to be* identifiable *if the quantity $P(y|\hat{x})$ can be computed uniquely from any positive distribution of the observed variables, that is, if for every pair of theories T_1 an T_2 such that $P_{T_1}(\mathbf{v}) = P_{T_2}(\mathbf{v}) > 0$, we have $P_{T_1}(y|\hat{x}) = P_{T_2}(y|\hat{x})$*

Identifiability means that $P(y|\hat{x})$ can be estimated consistently from an arbitrarily large sample randomly drawn from the distribution of the observed variables.

Definition 5 (back-door path) *A path from X to Y in a graph G is said to be a* back-door path *if it contains an arrow into X.*

Figure 1 illustrates a simple causal theory in the form of a diagram. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2), whether the sprinkler is on (X_3), whether the pavement would get wet (X_4), and whether the pavement would be slippery (X_5). All variables in this figure are binary, taking a value of either true or false, except the root variable X_1 which can take one of four values: Spring, Summer, Fall, or Winter. Here, the absence of a direct link between X_1 and X_5, for example, captures our understanding that the influence of seasonal variations on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement).

The theory corresponding to Figure 1 consists of five functions, each representing an autonomous mechanism:

$$\begin{aligned}
X_1 &= U_1 \\
X_2 &= f_2(X_1, U_2) \\
X_3 &= f_3(X_1, U_3) \\
X_4 &= f_4(X_3, X_2, U_4) \\
X_5 &= f_5(X_4, U_5)
\end{aligned} \tag{5}$$

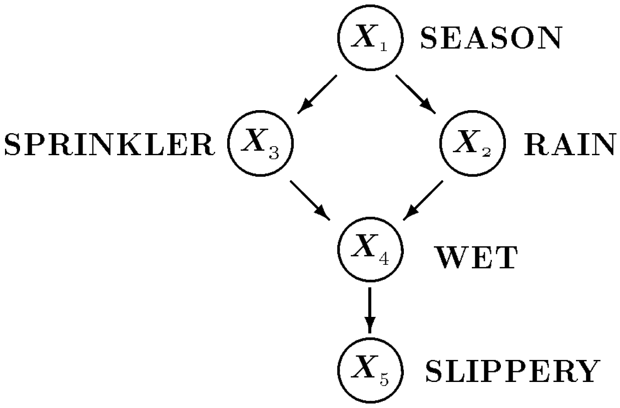

Figure 1: A diagram representing a causal theory on five variables.

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. (5), 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 actions on other variables. For example, 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 descendant of the manipulated variable X_3.

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

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

$$X_i = f_i(\mathbf{pa}_i, U_i), \qquad \mathbf{pa}_i \subseteq \{X_1, \ldots, X_{i-1}\} \tag{6}$$

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

$$U_i \perp\!\!\!\perp \{X_1, \ldots, X_{i-1}\} \tag{7}$$

These two conditions, also called Markovian, are the basis of Bayesian networks [Pearl, 1988] and they enable us to compute causal effects directly from the conditional probabilities $P(x_i|\mathbf{pa}_i)$, without specifying the functional form of the functions f_i, or the distributions $P(u_i)$ of the disturbances. This is seen immediately from the following observations:

The distribution induced by any Markovian theory T is given by the product

$$P_T(x_1, \ldots, x_n) = \prod_i P(x_i|\mathbf{pa}_i) \tag{8}$$

where $\mathbf{pa}_i$ are the direct predecessors (called *parents*) of X_i in the diagram. On the other hand the distribution induced by the subtheory $T_{x'_j}$, representing the action $do(X_j = x'_j)$ is given by product

$$P_{T_{x'_j}}(x_1, \ldots, x_n) \;=\; \begin{cases} \prod_{i \neq j} P(x_i|\mathbf{pa}_i) = \frac{P(x_1,\ldots,x_n)}{P(x_j|\mathbf{pa}_j)} & \text{if } \; x_j = x'_j \\ 0 & \text{if } \; x_j \neq x'_j \end{cases} \tag{9}$$

where the partial product reflects the surgical removal of the

$$X_j = f_j(\mathbf{pa}_j, U_j)$$

from the theory of equation (6).

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

$$P_T(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{10}$$

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

$$P_T(x_1, x_2, x_4, x_5|do(X_3 = \mathrm{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = \mathrm{ON})\ P(x_5|x_4) \tag{11}$$

Note the difference between the action $do(X_3 = \mathrm{ON})$ and the observation $X_3 = \mathrm{ON}$. The latter is encoded by ordinary Bayesian conditioning, while the former by conditioning a mutilated graph, with the link $X_1 \to X_3$ 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 amputation of $X_3 = f_3(X_1, U_3)$ from (5) ensures the suppression of any abductive inferences from any of the action's consequences.

Note also that Equations (8) through (11) are independent of T, in other words, the pre-actions and post-action distributions depend only on observed conditional probabilities but are independent of the particular functional form of $\{f_i\}$ or the distribution $P(\mathbf{u})$ which generate those probabilities. This is the essence of identifiability as given in Definition 4, which stems from the Markovian assumptions (6) and (7). Section 3.2 will demonstrate that certain causal effects, though not all, are identifiable even when the Markovian property is destroyed by introducing dependencies among the disturbance terms.

Generalization to multiple actions and conditional actions are straightforward. Multiple actions $do(X = x)$, where X is a compound variable result in a distribution similar to (9), except that all factors corresponding to the variables in X are removed from the product in (8). Stochastic conditional strategies of the form

$$do(X_j = x_j) \text{ with probability } P^*(x_j|\mathbf{pa}_j^*)$$

where $\mathbf{pa}_j^*$ is the support of the decision strategy, also result in a product decomposition similar to (9), except that each factor $P(x_j|\mathbf{pa}_j)$ is replaced with $P^*(x_j|\mathbf{pa}_j^*)$.

The surgical procedure described above is not limited to probabilistic analysis. The causal knowledge represented in Figure 1 can be captured by logical theories as well, for example,

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

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for trigerring and inhibiting abnormalities. The double arrows represent the assumption that the events on the r.h.s. of each equation are the *only* causes for the l.h.s.

It should be emphasized though that the models of a causal theory are not made up merely of truth value assignments which satisfy the equations in the theory. Since each equation represents an autonomous process, the scope of each individual equation must be specified in any model of the theory, and this can be encoded using either the graph (as in Figure 1) or the generic description of the theory, as in (5). Alternatively, we can view a model of a causal theory to consist of a mutually consistent set of submodels, with each submodel being a standard model of a single equation in the theory.

3.2 Action Calculus

The identifiability of causal effects demonstrated in Section 3.1 relies critically on the Markovian assumptions (6) and (7). 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 (6) is violated and identifiability may be destroyed. This can be seen easily from Eq. (9); if any parent of the manipulated variable X_j is unobserved, one cannot estimate the conditional probability $P(x_j|\mathbf{pa}_j)$, 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 $\mathbf{pa}_j$ are unobservable and, 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 [Pearl, 1995, Galles, 1995].

These tests and derivations are based on a symbolic calculus 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, i.e., they overrule equations that tie the manipulated variables to their pre-intervention causes.

Let X, Y, and Z be arbitrary disjoint sets of nodes in a DAG G. We say that X and Y are independent given Z in G, denoted $(X \perp\!\!\!\perp Y|Z)_G$, if the set Z d-separates X from Y in G. 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 1 Let G be the directed acyclic graph 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

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}} \tag{12}$$

Rule 2 Action/observation exchange

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \text{ if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}} \tag{13}$$

Rule 3 Insertion/deletion of actions

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}} \tag{14}$$

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. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in Spirtes et al. (1993)) which supports all three rules.

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $do(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms.

Rule 2 provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{X}\overline{Z}}$).

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 1, which reduces q into a standard (i.e., hat-free) probability expression involving observed quantities. □

Although Theorem 1 and Corollary 1 require the Markovian property, they do not require all variables to be observable and, hence, they can be applied to non Markovian, recursive theories as well. To demonstrate, 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 1 tells us that the causal effect $P(x_4|x_3)$ is identifiable, because:

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

Rule 3 permits the deletion

$$P(x_2|\hat{x}_3) = P(x_2), \text{ because } (X_2 \perp\!\!\!\perp X_3)_{G_{\overline{X}_3}},$$

while Rule 2 permits the exchange

$$P(x_4|\hat{x}_3, x_2) = P(x_4|x_3, x_2), \text{ because } (X_4 \perp\!\!\!\perp X_3|X_2)_{G_{\underline{X}_3}}.$$

This gives

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

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

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, and,

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

The ability to assess the effect of interventions from nonexperimental data has immediate applications in the medical and social sciences, since subjects who undergo certain treatments often are not representative of the population as a whole. Such assessments are also important in AI applications where an agent needs to predict the effect of the next action on the basis of past performance records, and where that action has never been enacted out of free will, but in response to environmental needs or to other agent's requests.

3.3 Historical background

An explicit translation of interventions to "wiping out" equations from linear econometric models was first proposed by Strotz & Wold (1960) and later used in Fisher (1970) and Sobel (1990) . Extensions to action representation in nonmonotonic systems were reported in [Goldszmidt & Pearl, 1992, Pearl, 1993a]. Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b). A related formulation of causal effects, based on event trees and counterfactual analysis was developed by Robins (1986, pp. 1422-25). Calculi for actions and counterfactuals based on this interpretation are developed in [Pearl, 1994b] and [Balke & Pearl, 1994a], respectively.

4 Counterfactuals

A counterfactual sentence has the form

If A were true, then C would have been true?

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations, and C, the counterfactual consequent, specifies a result that is expected to hold in the 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 shooting 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 counterfactuals antecedent A [Lewis, 1973]. Ginsberg [1986], followed a similar strategy. The drawback of the "closest world" approach is that it leaves the precise specification of the closeness measure almost unconstrained. In the domino example cited in Section 1, the manifestly closest world consistent with the antecedent "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 over to the right, tile $i+1$ would be tipped as well" as plausible and valid. Thus, we see that distances among worlds are not determined merely by surface similarities but require a delicate balance between disturbed and naturally occurring mechanisms. The local surgery paradigm expounded in Section 1 offers in fact a concrete explication of the closest world approach. A world w_1 is "closer" to w than a world w_2 is, if the the set of atomic surgeries needed for transforming w into w_1 is a subset of those needed for transforming w into w_2. In the domino example, finding tile i tipped and $i+1$ erect requires the breakdown of two mechanism (e.g., by two external actions) compared with one 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.

Counterfactual thinking is important to fault diagnosis, planning, determination of liability, and policy analysis, and it dominates most 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 much lower today". The messages conveyed by such thought experiments emphasize an understanding of generic laws in the domain, and are aimed to guide future policy making, for example, that defeated countries should not be humiliated, or that lowering taxes (contrary to Reaganomics theory) tend to increase national debts.

Strangely, there is very little formal work on counterfactual reasoning, or even policy analysis in the general literature. An examination of econometric journals and textbooks, for example, reveals a strange imbalance: while an enormous mathematical machinery is centered on problems of estimation and prediction, policy analysis (the ultimate goal of economic theories) receives almost no formal treatment. Currently, the primary methods of economic policy making are based on the so called *reduced-form* analysis [Intriligator, 1978, Maddala, 1992] which boils down to the following: 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 by a process in which the decision variables X act as exogenous variables. Unfortunately, almost every realistic policy making (e.g., taxation) involves some endogenous variables, that is, variables whose values are determined by other variable in the analysis. Taking taxation policies as an example, economic data are generated by a process in which Government is reacting to various indicators and various pressures, hence, taxation is endogenous in the estimation phase of the study. Taxation becomes exogenous when we wish to predict the impact of a specific decision to raise or lower taxes. Thus, the reduced-form method is valid only when past decisions are nonresponsive to other variables in the system, and this, unfortunately, rules out almost all interesting control variables (e.g., tax, interest-rates, quotas) from the analysis.

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

The connection between counterfactuals and policy making was formulated in [Balke & Pearl, 1994a] using the network representation of causal theories. The probability of a counterfactual sentence is computed by the following steps: First, the available observations are used to update the joint probability of the exogenous U variables (viewed as constant, but unknown boundary conditions, see Eq. (3)). Second, the counterfactual antecedent is interpreted as an external intervention that forces that antecedent to hold true. In the network representation, such an intervention is simulated by severing all causal edges that lead into the antecedent variables and setting their values to those specified in the antecedent. Finally, the probability of the counterfactual consequent is evaluated using the mutilated network as in (9). It turns out that, unlike causal effect queries, counterfactual queries are not identifiable even in Markovian theories, but require that the functional-form of $\{f_i\}$ be specified. In [Balke & Pearl, 1994b] a method is devised for computing bounds on counterfactual probabilities which, under certain circumstances may collapse to point estimates. This method has been applied to the evaluation of causal effects in studied involving noncompliance, to the determination of legal liability, and to the evaluation of economic policies in non-recursive linear systems.

References

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

[Balke & Pearl, 1994a] Balke, A. and Pearl, J., "Probabilistic evaluation of counterfactual queries," in *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, Volume I, 230-237, July 31 - August 4, 1994.

[Balke & Pearl, 1994b] 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.

[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.

[Fisher, 1970] Fisher, F.M., "A correspondence principle for simultaneous equation models," *Econometrica*, 38, 73–92, 1970.

[Galles, 1995] Galles, D. and Pearl, J., "Testing Identifiability of Causal Effects," UCLA Computer Science Department, Technical Report (R-226), March 1995. Submitted to UAI-95.

[Gardenfors, 1988] Gardenfors, P., *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, MIT Press, Cambridge, MA, 1988.

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

[Ginsberg, 1986] Ginsberg, M.L., "Counterfactuals", *Artificial Intelligence*, 30, 35–79, 1986.

[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.

[Intriligator, 1978] Intriligator, M.D., *Econometric models, techniques, and applications*, Englewood Cliffs, N.J., 1978.

[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.

[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.K., *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.

[Maddala, 1992] Maddala, G.S., *Introduction to Econometrics*, Mcmillan, NY, 1992.

[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*.

[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.

[Robins, 1986] Robins, J., "A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect," *Mathematical Modelling*, 7, 1393–1512, 1986.

[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.

[Sobel 1990] Sobel, M.E., "Effect analysis and causation in linear structural equation models," *Psychometrika*, 55(3), 495–515, 1990.

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

[Strotz & Wold, 1960] Strotz, R.H. and Wold, H.O.A., "Recursive versus nonrecursive systems: An attempt at synthesis," *Econometrica* 28, 417–427, 1960.