A Decision-Based View of Causality

David Heckerman
Microsoft Research, Bldg 9S
Redmond WA 98052-6399
heckerma@microsoft.com

Ross Shachter
Department of Engineering-Economic Systems
Stanford, CA 94305-4025
shachter@camis.stanford.edu

Abstract

We present a precise definition of cause and effect in terms of a more fundamental notion called unresponsiveness. Our definition departs from the traditional view of causation in that our causal assertions are made relative to a set of decisions. An important consequence of this departure is that we can reason about cause locally, not necessarily attaching a causal explanation to every dependency. Such local reasoning can be beneficial in that, given a set of real decisions to make, it may not be necessary to determine whether some dependencies are causal. Also in this paper, we examine the graphical encoding of causal relationships. We show that ordinary influence diagrams are an inadequate representation of cause, whereas influence diagrams in Howard Canonical Form can always represent cause and effect accurately. In addition, we establish a correspondence between Pearl and Verma's (1991) causal model and the influence diagram.

1 Introduction

Most traditional models of uncertainty, including Markov networks and belief networks have focused on the associational relationship among variables as captured by conditional independence and dependence. Associational knowledge, however, is not sufficient when we want to make decisions under uncertainty. For example, although we know that smoking and lung cancer are probabilistically dependent, we cannot conclude from this knowledge that we will increase our chances of getting lung cancer if we start smoking. In general, to make rational decisions, we need to be able to predict the effects of our actions.

Recent work by Artificial Intelligence researchers, statisticians, and philosophers—for example, Pearl and Verma (1991), Druzdzel and Simon (1993), and Spirtes et al. (1993)—have emphasized the importance of identifying causal relationships for purposes of modeling the effects of intervention. They argue, for example, that if we believe that smoking causes lung cancer, then we believe that our choice whether to smoke can affect whether we get lung cancer. In contrast, if we believe that smoking does not cause lung cancer, our choice whether to smoke would not affect whether we get lung cancer, and the observed correlation between our smoking and lung cancer could be explained perhaps by a common cause of both (e.g., a genetic predisposition toward cancer and the desire to smoke), which we are unable to control.

This recent work has led to significant breakthroughs in causal reasoning. For example, Pearl and Verma (1991) have shown how causal knowledge represented graphically in a *causal model* can be used to predict the effects of interventions, Spirtes et al. (1993) have shown how observational data can be used to suggest causal relationships, and Pearl (1994) has shown how, given a qualitative causal structure, the quantitative effects of intervention may be estimated from observational data alone.

In this paper, we offer two improvements to the current work in causal reasoning. First, the current approaches either take causality as a primitive notion, or provide only a fuzzy, intuitive definition of cause and effect. For example, in the introduction of their book on causation, Spirtes et al. (1993, p. 42) write:

> We understand causation to be a relation between particular events: something happens and causes something else to happen. Each cause is a particular event and each effect is a particular event. An event A can have more than one cause, none of which alone suffice to produce A. An event A can also be overdetermined: it can have more than one set of

causes that suffice for A to occur. We assume that causation is transitive, irreflexive, and antisymmetric.

In this paper, we offer a definition of causation in terms of a more fundamental relation that we call unresponsiveness. Our definition is precise, and can be used as an assessment aid when someone is having trouble determining whether or not a relationship is causal. Also, our definition can help people accurately communicate their beliefs about causal relationships.

Second, the current approaches require all relationships to be causal. That is, for any two probabilistically dependent events or variables x and y in a given domain, these methods require a user to assert either that x causes y, y causes x, x and y share a common cause, or x and y are common causes of an observed variable. For example, Verma and Pearl's (1991) causal model is a directed acyclic graph (DAG), wherein every node corresponds to a variable and every arc from nodes x to y corresponds to the assertion that x is a direct cause of y. When using a causal model to represent a domain, one of these four causal explanations must hold for every dependency in the domain.

Our definition of causation is local in that it does not require all relationships to be causal. This property can be advantageous when making decisions. Namely, given a particular problem domain consisting of a set of decisions and observable variables, there may be no need to assign a causal explanation to all dependencies in the domain in order to determine a rational course of action. Consequently, our definition may enable a decision maker to reason more efficiently.

Another advantage of our approach is that it is consistent with both current methods for reasoning about causality as well as the philosophy of decision analysis, and thereby provides a means by which the two disciplines may communicate. For over a decade, decision analysts have used the influence diagram to represent the effects of interventions on a set of uncertain variables [Howard and Matheson, 1981]. Nonetheless, they have carefully avoided using notions of causality in their work, in large part due to a lack of a precise definition of causality. Our paper offers a means by which the decision analysts may begin to understand the ongoing efforts in causal modeling and contribute to this research endeavor. We begin the dialogue by showing how causal relationships (by our definition) may be encoded in special forms of the influence diagram, and by showing a relationship between the influence diagram and existing graphical models of cause.

2 Background

Fundamental to our discussion is the distinction between a *chance variable* and *a decision variable.* In general, a variable has a (possibly infinite) set of mutually exclusive and collectively exhaustive possible *states.* The state of a decision variable corresponds to an action chosen by a person, usually called the decision maker. In contrast, a chance variable is uncertain and its state may be at most indirectly affected by the decision maker's choices. For example, whether or not to smoke is a decision variable, whereas whether or not a person develops lung cancer is a chance variable. We shall use lowercase letters to denote single variables, and uppercase letters to denote sets of variables. We call an assignment of state to every variable in set X an *instance* of X. Typically, we refer to the possible states of a decision variable as *alternatives.* We use a probability distribution $P\{X|Y\}$ to represent a decision maker's uncertainty about X, given that a set of chance and/or decision variables Y is known or determined.

In this paper, we are interested in modeling relationships in a *domain* consisting of chance variables U and decision variables D. A well-known graphical language for modeling such relationships is the influence diagram. In this paper, we use this representation to illustrate many of our concepts. In addition, we critique the influence diagram as a representation of causal relationships. In the remainder of this section, we review the representation.

An *influence diagram* is (1) a directed acyclic graph (DAG) containing decision and chance nodes corresponding to decision and chance variables, and information and relevance arcs, representing what is known at the time of a decision and probabilistic dependence, respectively, (2) a set of probability distributions associated with each chance node, and optionally (3) a utility node and a corresponding set of utilities. A *belief network* is an influence diagram containing only chance nodes and relevance arcs.

An *information arc* is one that points to a decision node. An *information arc* from chance or decision node a to decision node d encodes the assertion that variable a will be known when decision d is made. (We shall use the same notation to refer to a variable and its corresponding node in the diagram.) A *relevance arc* is one that points to a chance node. The *absence* of a relevance arc represents conditional independence. To identify relevance arcs, we start with an ordering of the variables in $U = (x_1, \ldots, x_n)$. Then, for each variable x_i in order, we ask the decision maker to identify a set $Pa(x_i) \subseteq \{x_1, \ldots, x_{i-1}, D\}$ that renders x_i and $\{x_1, \ldots, x_{i-1}, D\} \setminus Pa(x_i)$ conditionally independent.

That is,

$$P\{x_i|x_1,\ldots,x_{i-1},D\} = P\{x_i|Pa(x_i)\} \qquad (1)$$

For every variable a in $Pa(x_i)$, we place a relevance arc from a to x_i in the diagram. That is, the nodes $Pa(x_i)$ are the *parents* of x_i.

Associated with each chance node x_i in an influence diagram are the probability distributions $P\{x_i|Pa(x_i)\}$. From the chain rule of probability and Equation 1, we obtain

$$P\{x_1,\ldots,x_n|D\} = \prod_{i=1}^{n} P\{x_i|Pa(x_i)\} \qquad (2)$$

That is, every influence diagram for $U \cup D$ uniquely determines a joint probability distribution for U given D.

A *deterministic node* is a special kind of chance node that is a deterministic function of its parents. A *minimal influence diagram* is an influence diagram where Equation 1 would be violated if any arc were removed.

Finally, an influence diagram may contain a single distinguished node, called a *utility node* that encodes the decision maker's utility for each instance of the node's parents.

Figure 1 contains an influence diagram for two lifestyle decisions: whether or not to smoke and whether or not to change diet. As is illustrated in the figure, we use ovals, squares, and a diamond to represent chance, decision, and utility nodes, respectively. (Not shown in the diagram, we use double ovals to represent a deterministic nodes.) There are no information arcs in the diagram, although one can imagine that—someday—we may be able to observe our genotype prior to making these decisions. The influence diagram contains several missing relevance arcs. One assertion made by the absence of these arcs is that *lung cancer* and *cardiovascular status* are conditionally independent, given *smoke*, *diet*, and *genotype*. This assumption and others in the diagram are questionable, but they will serve for purposes of example. We note that this influence diagram is not complete, in that the decisions are not ordered. As we shall see, however, decision order is not important for our discussion.

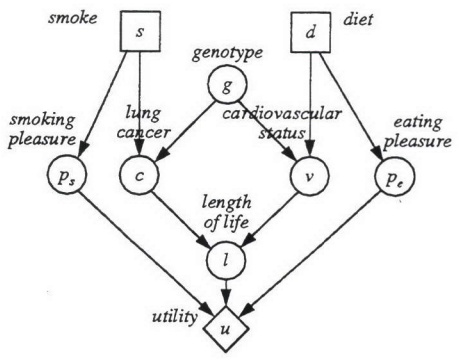

Figure 1: An influence diagram for decisions about lifestyles.

3 Unresponsiveness

In this section, we introduce the notion of responsiveness, a fundamental relation underlying causation. In the following section, we use this relation to define causal dependence.

Let us consider the simple decision d of whether or not to bet heads or tails on the outcome of a coin flip c. Let the variable w represent whether or not we win. Thus, w is a deterministic function of d and c: we win if and only if the outcome of the coin matches our bet. Let us assume that d and c are probabilistically independent and that the coin is fair—that is $P\{heads\} = 1/2$. In this case, d and w are also probabilistically independent: the probability of a win is $1/2$ whether we bet heads or tails.

In this example, we are uncertain about whether or not the coin with come up heads, but we can be certain that whatever the outcome, it would have been the same had we bet differently. We say that c *is unresponsive to* d. We cannot make the same claim about the relationship between d and w. Namely, we know that w depends on d in the sense that had we made a different bet d, the state of w would have been different. For example, we know that if we had bet heads and won, then we would have lost if we had best tails. We say that w *is responsive to* d.

In general, to determine whether or not uncertain variable x is unresponsive to decision d, we have to answer the query "Would the outcome of x have been the same had we chosen a different alternative for d?" Queries of this form are a simple type of *counterfactual query*, discussed in the philosophical literature. In our experience, we have found that people are comfortable answering such restricted counterfactual queries. One of the fundamental assumptions of our work presented here is that these queries are easily answered.

We see that probabilistic independence and unresponsiveness are not the same relation. Although both c and w are (individually) probabilistically independent of d, c is unresponsive to d whereas w is responsive to d.

Nonetheless, if an uncertain variable x is unresponsive to a decision d, then x and d must be probabilistically independent. That is, if the outcome of x is not affected by d, then the probability of x given d must

be the same for all states of d. For example, consider the decision *smoke* and the uncertain variable *genotype* in our lifestyles decision problem. Until methods improve for genetic analysis, we will be uncertain about our genotype. Nonetheless, it is reasonable to assert that whatever our genotype is, it will not be affected by whether or not we smoke. That is, it is reasonable to assert that *genotype* is unresponsive to the decision *smoke*. Given this belief, we must also believe that these two variables are probabilistically independent.

In the examples that we have considered, we have implicitly held the belief that after we have made our decision, the outcome of all uncertain variables are determined, albeit possibly unknown. We call the outcome of some or all of the uncertain variables together with our decisions that led to those outcomes a *counterfactual world.* In the coin example, we have one binary decision to make. Regardless of this decision, the coin will come up either heads or tails, although we do not know which. If the coin comes up heads, then the counterfactual worlds are $\{d = \text{heads}, c = \text{heads}, w = \text{win}\}$ and $\{d = \text{tails}, c = \text{heads}, w = \text{lose}\}$. If the coin comes up tails, then the counterfactual worlds are $\{d = \text{heads}, c = \text{tails}, w = \text{lose}\}$ and $\{d = \text{tails}, c = \text{tail}, w = \text{win}\}$. In general, the decision maker may be (and usually is) uncertain about which set of counterfactual worlds is realized.

When an uncertain variable x is responsive to a decision d, x is different in at least two counterfactual worlds of $\{x, d\}$. In some subset of those counterfactual worlds, however, x may be the same. For example, consider the variables *smoke*, *lung cancer*, and *length of life* in our lifestyles decision problem. As mentioned, *length of life* is responsive to the decision *smoke*. Nonetheless, if we consider only the counterfactual worlds in which *lung cancer* is true (or false), then *length of life* will be the same. We say that *length of life* is unresponsive to *smoke* in counterfactual worlds where *lung cancer* is the same, or that length of life *is unresponsive to* smoke *in worlds limited by* lung cancer for short. We refer to this concept as *limited unresponsiveness.*

In general, to determine whether or not an uncertain variable x is unresponsive to decision d in worlds limited by y, we have to imagine a scenario where we decide d and observe x and y and answer the counterfactual query "Would x still be the same had we decided differently, assuming that we were to find out (after deciding) that y was the same?" This counterfactual query is somewhat more complex than is the simple query associated with the unlimited form of unresponsiveness. In Section 5, we present an alternative formulation of limited unresponsiveness that some people may find easier to understand than this one.

Limited unresponsiveness has several simple properties. First, whether an uncertain variable x is unresponsive or responsive to d, it will always be unresponsive to d in worlds limited by d.

Second, if x is unresponsive to d, it follows that x is unresponsive to d in worlds limited by Y for any set of variables Y. That is, if x is unaffected by d, then it must be unaffected by d in the subsets of all counterfactual worlds where Y is the same. In our lifestyles decision problem, for example, if we believe that *genotype* would be the same whether or not we smoke, then we must believe that, *genotype* would be the same if *lung cancer* is the same, whether or not we smoke. The coin example is a bit more tricky, due to the deterministic relationship between $\{d, c\}$ and w. As we discussed, c is unresponsive to d. Consequently, c should be unresponsive to d in worlds limited by w. That is, we should answer "yes" to the query "Would c still be the same had we bet differently, assuming that we find out after betting that w is the same." Indeed, the answer is "yes" trivially, because the only way that w could be the same is if we had not changed our bet.

We now formalize these concepts.

Definition 1 (Counterfactual World) *Given uncertain variables $X \subseteq U$ and decisions D, a* counterfactual world of X and D *is any instance assumed by $X \cup D$ after the decision maker chooses a particular instance of D.*

We emphasize that the decision maker may be (and usually is) uncertain about the counterfactual world that results from deciding D.

Definition 2 ((Un)responsiveness) *Given uncertain variables X and decisions D, X* is unresponsive to D, denoted $X \not\hookleftarrow D$, *if X assumes the same instance in all counterfactual worlds of $X \cup D$. X* is responsive to D, denoted $X \hookleftarrow D$, *if X can assume different instances in different counterfactual worlds of $X \cup D$.*

Definition 3 (Limited (Un)responsiveness) *Given sets of uncertain variables X and Y and decisions D, X* is unresponsive to D in worlds limited by Y, denoted $X \not\hookleftarrow_Y D$, *if X assumes the same instance in all counterfactual worlds of $X \cup Y \cup D$ where Y assumes the same instance. X* is responsive to D in worlds limited by Y, denoted $X \hookleftarrow_Y D$, *if X can assume different instances in different counterfactual worlds of $X \cup Y \cup D$ where Y assumes the same instance.*[1]

[1] Our notions of *unresponsiveness* and *limited responsiveness* correspond with those of *fixed set* and *conditional fixed set*, respectively, in earlier work [Heckerman and Shachter, 1994].

Again, we emphasize that the identity of the counterfactual worlds may be (and usually is) uncertain. In the coin example, we do not know whether the coin will come up heads or tails, but we do know that whatever the outcome, it will be the same in the counterfactual worlds $\{d = \text{heads}, c\}$ and $\{d = \text{tails}, c\}$. Also, we emphasize that that X and Y refer to the collections of events some of which—the responsive ones—occur *after* decisions D have been made.

Finally, we note that the identification of variables that are unresponsive to D does not depend on the order in which the decisions in D are made. In the remainder of the paper, we will ignore the ordering of decisions. Consequently, we have no need to use information arcs in our influence diagrams.

4 Definition of Cause

Armed with the primitive notions of unresponsiveness and limited unresponsiveness, we can now formalize our definition of cause.

Definition 4 (Cause) *Given decisions D,* the variables C are causes for x with respect to D *if (1) $x \notin C$, (2) x is responsive to D, and (3) C is a minimal set of variables such that x is unresponsive to D in worlds limited by C—that is, $x \leftrightarrow D$, and C is a minimal set such that $x \not\leftrightarrow_C D$.*

The first condition simply says that cause is irreflexive. The second condition says that for x to be caused with respect to decisions D, it must be responsive to those decisions. The third condition says that if we can find set of variables Y such that x can be different in different counterfactual worlds only when Y is different, then Y must contain a set of causes for x.

Our definition of cause departs from traditional usage of the term in that we consider causal relationships relative to a set of decisions. At first glance, this departure may appear to be a drawback of the definition. Nonetheless, we find this departure has its advantages. First, we do not require the decisions D to be realizable in practice or at all. If we want to think about whether the moon causes the tides, we merely need to imagine a decision that affects the moon's orbit (e.g., we can imagine a decision of whether or not to destroy the moon). Therefore, our definition does not restrict the types of causal sentences that we can consider. Second, given a set of real decisions to make, it may not be necessary to determine whether some dependencies are causal. As we see in the examples that follow, our decision-based definition makes us provide causal explanations only for those relationships that matter. Using our definition, we can reason about cause locally, not necessarily having to attach a causal explanation to every dependency.

Variants of our lifestyles decision problem shown in Figure 2 help us to illustrate the definition. Some conclusions that can be drawn about each domain are shown next to the influence diagram for that domain. The arcs in the figures are suggestive of causal relationships. Nonetheless, the reader should resist this interpretation until reading Section 6.

First, let us consider the decision problem in Figure 2a. Here, we model only the decision of whether or not to smoke; and we do not bother to model the variable *length of life*. For this problem, it is reasonable to assert that *lung cancer* is responsive to $D = \{smoke\}$. Also, as we discussed, *lung cancer* is unresponsive to $\{smoke\}$ in worlds limited by $\{smoke\}$. Consequently, by our definition, we can conclude that $\{smoke\}$ is a singleton cause for *lung cancer*. Similarly, we may conclude that $\{smoke\}$ is a cause for *smoking pleasure*. In general, some subset of D will always be causes for any chance variable x.

Also in the example, it is reasonable to assert that *utility* is responsive to D, *utility* is unresponsive to D in worlds limited by $\{smoking\ pleasure,\ lung\ cancer\}$, and there is no subset C of $\{smoking\ pleasure,\ lung\ cancer\}$ such that *utility* is unresponsive to D in worlds limited by C. Therefore, we can conclude that $\{smoking\ pleasure,\ lung\ cancer\}$ are causes for *utility*. As shown in the figure, we may also conclude that $\{smoke\}$ is a cause for *utility*.

Someday, it may be possible to use retroviral therapy to alter one's genetic makeup. Figure 2b shows an influence diagram assuming that a decision of whether or not to undergo such therapy is available. Here, it is reasonable to assert that *genotype* is responsive to $D = \{smoke,\ retroviral\ therapy\}$. In contrast, it is reasonable to assert that *genotype* is unresponsive to D in worlds limited by $\{retroviral\ therapy\}$. Therefore, $\{retroviral\ therapy\}$ is a singleton cause for *genotype*. Furthermore, it is reasonable to assert that *lung cancer* is responsive to D and that $\{smoke,\ genotype\}$ is a minimal set C such that *lung cancer* is unresponsive to D in worlds limited by C. Thus, we can conclude that $\{smoke,\ genotype\}$ are causes for *lung cancer* with respect to $\{smoke,\ retroviral\ therapy\}$.

The examples in Figures 2a and b illustrate the benefits of defining cause with respect to a set of decisions. Given our definition, not all dependencies need have a causal explanation. For example, in the first example, *genotype* and *lung cancer* are dependent, but not causally related. Without a retroviral therapy or any other alternative for modifying genotype, how-

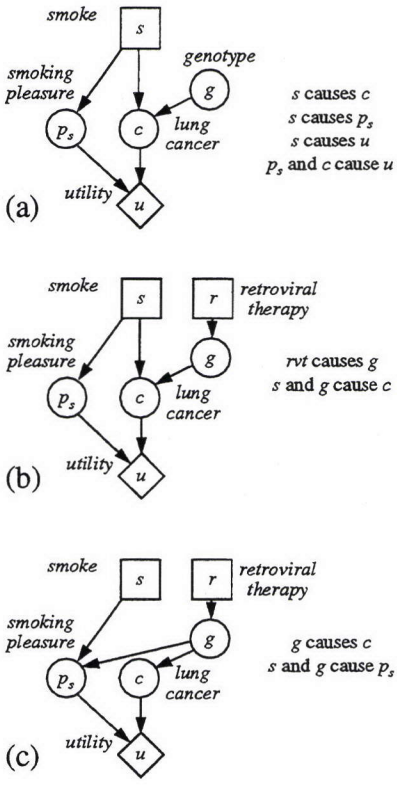

Figure 2: Variants of the decision problem in Figure 1.

ever, there is little point in knowing whether genotype causes lung cancer. Our approach allows us to ignore this question, and still make a rational decision. Of course, we may believe that someday such a therapy will be found, in which case we may want to include the decision of whether or not to wait a few years before deciding to smoke. In this case, as our formulation would show, we would want to know whether genotype is a cause of lung cancer.

The influence diagram in Figure 2c corresponds to an "alternative" view of the relationships between the decision to smoke and lung cancer. Here, smoking is not a cause for lung cancer. Rather, genotype is a cause for lung cancer, and both genotype and smoking are causes for smoking pleasure. As mentioned in the introduction, these two views predict differently what would happen should one start smoking. This example emphasizes that our definition of cause and effect is subjective. One person may hold beliefs corresponding to the model in Figure 2a (or b), whereas another person may hold beliefs corresponding to the model in Figure 2c. Both people are "correct" provided they make their decisions in a manner that is consistent with their beliefs.

Our definition of cause has two satisfying ramifications. First, it implies that x and y cannot cause each other, except for the special case where x and y are related deterministically. Namely, if $\{x\}$ is a cause for y with respect to D and $\{y\}$ is a cause for x with respect to D, then one can show that x must be a deterministic function of y and D (and y must be a deterministic function of x and D).

Second, if C are causes for x with respect to decisions D, then every variable in C must be responsive to D. Otherwise, C would not be a minimal set such that $x \not\rightharpoonup_C D$.

Table 1: The four states of the mapping variable $c(s)$, which relates smoking and lung cancer.

	state 1		state 2		state 3		state 4	
smoke	no	yes	no	yes	no	yes	no	yes
lung cancer	no	yes	yes	no	no	no	yes	yes

5 Mapping Variables and Causal Mechanisms

In this section, we show how the concept of limited unresponsiveness can be formulated in terms of the simpler concept of unresponsiveness. We thereby provide an alternative definition of cause.

Our alternative formulation of limited unresponsiveness is based on the concept of a mapping variable. To understand what a mapping variable is, let us consider the relationship between the decision *smoke* (s) and the chance variable *lung cancer* (c). In this situation, the mapping variable for c as a function of s, denoted $c(s)$, represents all possible deterministic mappings from s to c. The possible states of this variable are shown in Table 1. Each state contains a particular assignment to c for every possible state of s.

When we introduce the mapping variable $c(s)$ to a domain containing variables c and s, *lung cancer* becomes a deterministic function of *smoke* and $c(s)$. For example, if *smoke* is *yes* and $c(s)$ is in state 1, then *lung cancer* will be *yes*. The uncertainty in the relationship between *smoke* and *lung cancer*, formerly associated with the variable *lung cancer*, now is associated with the variable $c(s)$. In effect, we have *extracted* the uncertainty in the relationship between these two variables, and moved this uncertainty to the node $c(s)$.

In general, we have the following definition.

Definition 5 (Mapping Variable) *Given uncertain variables X and variables Y, the* mapping variable $X(Y)$ *is the uncertain variable that represents all possible mappings from Y to X.*

There are several important points to be made about

mapping variables. First, as in our example, X is always a deterministic function of $X(Y)$ and Y.

Second, additional assessments typically are required when introducing a mapping variable. For example, two independent assessments are needed to quantify the relationship between *smoke* and *lung cancer*, whereas three independent assessments are required for the node $c(s)$. In general, many additional assessments are required. If X has r instances and Y has q instances, then $X(Y)$ will have r^q states. In real-world domains, however, reasonable assertions of independence decrease the number of required assessments. In some cases, no additional assessments are necessary (see, e.g., Heckerman et al. 1994).

Third, although we may not be able to observe a mapping variable directly, we may be able to learn something about it. For example, we can imagine a test that measures the susceptibility of someone's lung tissue to lung cancer in the presence of tobacco smoke. The probabilities on the outcomes of this test would depend on $c(s)$.

Fourth, an most important, we have the following theorem.

Theorem 1 (Mapping Variable) *Given decisions D, uncertain variables X, and a set of variables Y, $X \not\rightsquigarrow_Y D$ if and only if $X(Y) \not\rightsquigarrow D$.*

Roughly speaking, Theorem 1 says that X is unresponsive to decisions D in worlds limited by Y if and only if the way Y depends on X does not depend on D. This equivalence provides us with an formulation of limited unresponsiveness in terms of unresponsiveness. In addition, we have an alternative set of conditions for cause.

Corollary 2 (Cause) *Given decisions D, the variables C are causes for x with respect to D if (1) x is responsive to D, and (2) C is a minimal set of variables such that $x(C)$ is unresponsive to D.*

In our experience, we have found that neither formulation is universally less complex. In different situations, we have found that one or the other or both of our formulations has been useful for the assessment of causal relationships.

We can think of $x(C)$—where C are causes for x—as a *causal mechanism* that relates C and x. For example, suppose chance variables i and o represent the voltage input and output, respectively, of an inverter in a logic circuit. Given a decision d to which i responds, we can assert that $\{i\}$ is a cause for o. In this example, the mapping variable $o(i)$, represents the mapping from the inverter's inputs to it's outputs. That is, this mapping variable represents the state of the inverter itself.

Definition 6 (Causal Mechanism) *Given decisions D and an uncertain variable x that is responsive to D, a* causal mechanism for x with respect to D *is a mapping variable $x(C)$ where C are causes for x.*

From Corollary 2, it follows that any causal mechanism for x with respect to D is unresponsive to D.

6 The Influence Diagram as a Representation of Cause

Given the known benefits of the belief network for representing conditional independence, we would like a graphical representation of cause and effect. In this section, we show that an ordinary influence diagram is not an adequate representation for causal dependence, and describe conditions under which some inadequacies may be removed. In the following section, we examine a special type of influence diagram that is always an accurate representation of cause.

In Section 4, we saw that our notion of cause and effect is intimately related to the notion of unresponsiveness. Thus, we desire a graphical representation in which we can encode the existence or lack thereof of this relation.

At first glance, the influence diagram appears to be such a representation. In particular, consider the following graphical condition.

Definition 7 (Block) *Given an influence diagram with decision nodes D and chance nodes U, $C \subseteq U$ is said to* block D *from $x \in U$ if every directed path from a node in D to x contains at least one node in C.*

If we reexamine our examples in Figure 2, we see that whenever x is not blocked from D by the empty set—that is, whenever there is a path from a decision node to x—then x is responsive to D. In addition, whenever x is blocked from D by C, then x is unresponsive to D in worlds limited by C. Thus, in these examples, we may read cause and effect directly from the influence diagram.

In other examples, however, this correspondence between the graphical condition of blocking and unresponsiveness breaks down, making the ordinary influence diagram an inadequate representation of cause and effect. To formalize these inadequacies, we introduce the following definitions, which closely parallel Pearl's concepts of $\mathcal{I}$-map and $\mathcal{D}$-map. In these definitions, $S(D, U)$ denotes the complete set of unre-

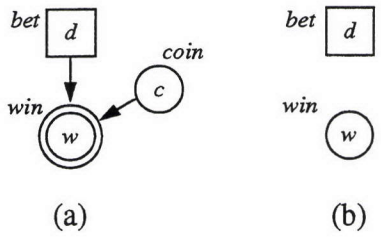

Figure 3: Influence diagrams for betting on a coin flip.

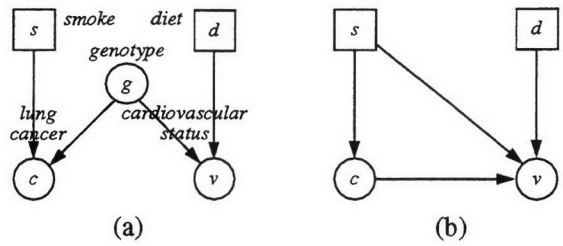

Figure 4: A subset of the lifestyles decision problem.

sponsiveness and responsiveness assertions made by a decision maker about domain $D \cup U$.

Definition 8 ($\mathcal{U}$-map)
An influence diagram $I(D,U)$ for domain $U \cup D$ is said to be a $\mathcal{U}$-map *with respect to $S(D,U)$ if*

$$C \text{ blocks } D \text{ from } x \text{ in } I(D,U) \implies x \not\leftarrow_C D \in S(D,U)$$

Definition 9 ($\mathcal{R}$-map) *An influence diagram $I(D,U)$ is said to be an* $\mathcal{R}$-map *if*

$$C \text{ does not block } D \text{ from } x \text{ in } I(D,U) \implies x \leftarrow_C D \in S(D,U)$$

Consider the influence diagram in Figure 3a for our coin example. Recall that w represents whether or not we win; and is a deterministic function of the bet d and the outcome of the coin flip c, as is indicated by the double oval. If we believe that the coin is fair, and if we do not bother to model the variable c explicitly (as shown in Figure 3b), then we need not place an arc from d to w, because the probability of winning will be $1/2$, regardless of our choice d. Nonetheless, w is responsive to d, because w will take on different states for different bets. Consequently, we have a situation where there is no path from d to w, and yet w is responsive to d. That is, our influence diagram for $U = \{d, w\}$ is not a $\mathcal{U}$-map with respect to $S(D,U)$.

Conversely, consider a subset of our lifestyles decision problem shown in Figure 4a. If we choose not to model the variable *genotype*, then we can obtain the influence diagram shown in Figure 4b.[2] In this influence diagram, we cannot remove any arc without producing invalid assertions of conditional independence. Nonetheless, *cardiovascular status* is unresponsive to D in worlds limited by $\{diet\}$. That is, the influence diagram for $U = \{s, d, c, v\}$ is not a $\mathcal{R}$-map with respect to $S(D,U)$.

In general, an ordinary influence diagrams suffers from the inadequacies that it may be neither a $\mathcal{U}$-map nor $\mathcal{R}$-map of the responsiveness assertions associated with its domain. In the next section, we describe a special type of influence diagram that is always a $\mathcal{U}$-map and $\mathcal{R}$-map. In the remainder of this section, we consider simple conditions under which the inadequacies of ordinary influence diagrams can be weakened.

A decision maker can always transform an influence diagram that is not an $\mathcal{U}$-map to one that is an $\mathcal{U}$-map, by adding arcs to it. In the coin example, knowing that w is responsive to d, a decision maker can add an arc from d to w, making the diagram an $\mathcal{U}$-map. The following definition and theorem lead to a simple procedure for identifying which arcs, if any, need to be added to an influence diagram to make it an $\mathcal{U}$-map.

Definition 10 (Unresponsive Influence Diagram)
An unresponsive influence diagram *is an influence diagram in which every node $x \in U$ is unresponsive to D in worlds limited by x's parents.*

Theorem 3 *All unresponsive influence diagrams for domain $D \cup U$ are $\mathcal{U}$-maps with respect to $S(D,U)$.*

Proof: Suppose that a set of chance nodes C block D from x, but that x is responsive to D in worlds limited by C. Because the influence diagram is causal, it follows that at least one of x's parents—say—y would be responsive to D in worlds limited by C. Applying this argument recursively, until $y \in C$, we obtain a contradiction. □

This theorem provides an implicit algorithm for transforming an ordinary influence diagram into a $\mathcal{U}$-map. In particular, for every chance node x, the decision maker determines whether or not x is unresponsive to D in worlds limited by x's parents. If not, the decision maker need only add arcs to x until this condition is satisfied. The result is an unresponsive influence diagram, and consequently a $\mathcal{U}$-map with respect to $S(D,U)$.

The following definition and theorem demonstrates a condition under which we can identify causes for some variables in an influence diagram. Following Pearl and Verma (1991), let us consider a special type of decision,

[2] We obtained this influence diagram using the ordering (s, d, c, v). Had we used the ordering (s, d, v, c), we would have obtained the influence diagram where v has the single parent d and c has parents s, d, and v.

called a set decision.

Definition 11 (Set Decision) *Given an influen diagram for uncertain variables U and decisions D a* set decision for $x \in U$ with respect to D *is any a cision node $s_x \in D$ such that (1) s_x has alternati "set x to k" for each state k of x and "do nothing and (2) x is the only child of s_x.*

Given a set decision s_x, we can literally set x to any its states, or we can do nothing. When we set x to o: if its states, none of the other ancestors of x contribu to the determination of x. When we do nothing, x determined as if it had no set-decision parent.

Theorem 4 *Given a minimal responsiveness infl ence diagram for uncertain variables U and decisio D, and a variable $x \in U$ that has nonempty parer $Pa(x)$, if D includes a set decision for each chanc node parent of x, then $Pa(x)$ are causes for x.*

Proof: Consider any node x. If x has no chance-node parents, then x is caused by its decision parents. If x has chance-node parents, then because the influence diagram is minimal and there exist set decisions for each such parent of x, x must be responsive to D. Furthermore, because the influence diagram is minimal and causal, $Pa(x)$, which includes the decisions pointing to x, must be a minimal set C such that x is unresponsive to D in worlds limited by C. Consequently, x is caused by its parents. □

7 Howard Canonical Form

Howard (1990) introduced a special type of influence diagram, which has become to be known as an influence diagram in *Howard Canonical Form* (HCF). Although not developed for this purpose, it turns out that an HCF influence diagram for $D \cup U$ is always both a $\mathcal{U}$-map and $\mathcal{R}$-map of $S(D, U)$. In this section, we describe HCF and develop a method for building influence diagrams in this form.

Although Howard (1990) does not use our language, his definition is equivalent to the following:

Definition 12 (Howard Canonical Form) *An influence diagram for uncertain variables U and decisions D is said to be in* Howard Canonical Form *if (1) every chance node that is not a descendant of a decision node is unresponsive to D, and (2) every chance node that is a descendant of a decision node is a deterministic node.*

We can transform any given influence diagram into one that is in HCF by adding causal-mechanism variables. For example, the HCF influence diagram corresponding to the ordinary influence diagram in Figure 5a is shown in Figure 5b. In this new influence diagram, we have added a node corresponding to the causal mechanism $c(s)$. This node becomes the only non-deterministic chance node, and is unresponsive to $D = \{smoke\}$.

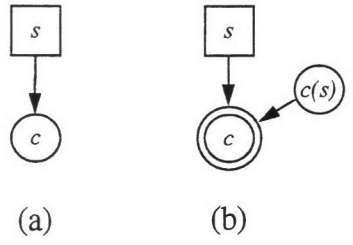

Figure 5: A transformation to Howard Canonical Form.

The following theorem describes, in general, how we can construct an influence diagram in HCF for a given domain.

Theorem 5 (Howard Canonical Form) *Given uncertain variables U and decisions D, an influence diagram in HCF for $U \cup D$ can be constructed as follows:*[3]

1. *Add a node to the diagram corresponding to each variable in $U \cup D$*
2. *Order the variables $x_1, \ldots, x_n$ in U so that the variables unresponsive to D come first*
3. *For $i := 1, \ldots, n$, if $x_i \not\leftarrow D$,*
   - *Add a causal-mechanism node $x_i(C_i)$ to the diagram, where $C_i \subseteq D \cup \{x_1, \ldots, x_{i-1}\}$*
   - *Make x_i a deterministic function of $C_i \cup x(C_i)$*
4. *Assess dependencies among the variables that are unresponsive D*

Proof: In step 3, all causal-mechanism nodes added to the diagram will be unresponsive to D and will not be descendants of decisions. Also, after step 3, all nodes in U that are responsive to D will be descendants of D and will be deterministic functions of their parents. In step 4, only the parents of nodes responsive to D will be altered. In no case will such a variable gain any variable in D as a parent. □

To illustrate this algorithm, consider the influence diagram shown in Figure 6a. We begin the construction by adding the variables $\{s, d, g, c, v\}$ to the diagram and choosing the ordering (g, c, v). Both c and

[3] As noted in the introduction, we are not concerned with information arcs and utility nodes in our construction.

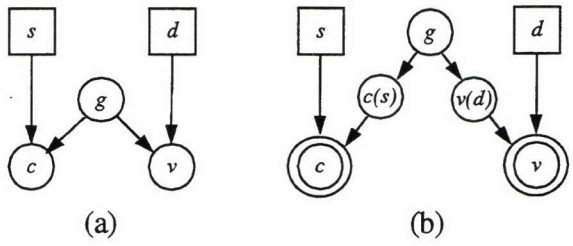

Figure 6: Another transformation to Howard Canonical Form.

v are responsive to $D = \{s, d\}$, and have causes s and d, respectively. Consequently, we add causal mechanisms $c(s)$ and $v(d)$ to the diagram, and make c a deterministic function of $\{s, c(s)\}$ and v a deterministic function of $\{d, v(d)\}$. Finally, we assess the dependencies among the unresponsive variables $\{g, c(s), v(d)\}$, adding arcs from g to $c(s)$ and $v(d)$ under the assumption that the causal mechanisms are conditionally independent given g. The resulting HCF influence diagram is shown in Figure 6b. This example illustrates an important point that causal mechanisms may be dependent. We return to this issue in Section 8.

By definition of HCF, we have the following theorem.

Theorem 6 *A minimal HCF influence diagram for $D \cup U$ is both a $\mathcal{U}$-map and $\mathcal{R}$-map with respect to $S(D, U)$.*

Thus, the HCF influence diagram is a suitable representation for causal relationships.

It is interesting to note that Howard did not develop HCF for purposes of modeling causal relationships. Rather, he developed the representation to facilitate the computation of value of information [Howard, 1990]. Before making important decisions, decision analysts investigate how useful it is to gather additional information. This investigation is typically done by computing the value of information about one or more chance nodes in the domain. To compute the *value of information of observing a chance variable x with respect to a decision d,* one computes the decision maker's expected value given that x is observed before the decision d is made, and subtracts it from the decision maker's expected value given that x is not observed before the decision is made. If the cost of learning something about x is greater than the value of information about x, then we know that it is not worthwhile to gather such information. Given an ordinary influence diagram, we cannot compute the value of information of variables responsive to D, because such variables cannot be observed before decisions D are made. In contrast, we can always compute the value of information of any non-deterministic variable in a HCF influence diagram, because all such variables are unresponsive to D by definition.

We also note that Theorem 5 contains the only algorithm known to us for constructing HCF influence diagrams. Until now, practicing decision analysts have had to use their intuitions (undoubtedly employing causal knowledge implicitly) in order to build HCF influence diagrams in accordance with the definition. Now, however, armed with our definitions of unresponsiveness and Theorem 5, we believe decision analysts will be able to construct these diagrams more systematically and effectively.

8 Global Causal Models

As we have mentioned, most previous work on the graphical representation of causality concerns the situation where all interactions in a domain are causal (see, e.g., Pearl and Verma 1991, Druzdzel and Simon 1993, and Spirtes et al. 1993, and Pearl 1994). We consider this special case, and describe correspondences between our work and the work of Pearl and Verma (1991), which is representative of this body of work.

Pearl and Verma (1991, p. 2) take causation to be a primitive notion, and define a causal model as follows:

Definition 13 (Causal Model, Pearl and Verma) *A* causal model *of a set of variables U is a DAG, in which each node corresponds to a distinct element of U.*

They qualify their definition, writing:

> The nodes of the DAG correspond to the variables under analysis, while the links denote direct causal influences among variables.

Each variable in Pearl and Verma's analysis plays a dual role of chance and decision variable. In particular, a variable may be observed, or set to a particular state (in the sense of a set decision as defined in Section 6). Therefore, we can convert a causal model to an influence diagram by copying the causal-model DAG, interpreting all nodes as chance nodes, and introducing a set-decision node for every chance node. Furthermore, as shown by the following definition and corollary of Theorem 4, every relationship in the resulting influence diagram is causal.

Definition 14 (Causal Network) *An influence diagram with uncertain variables U and decisions D is said to be a* causal network for U with respect to D *if $Pa(x)$ are causes for x with respect to D for all $x \in U$.*

Corollary 7 *A minimal causal influence diagram for uncertain variables U and decisions D such that D includes a set decision for each nonleaf uncertain variable in U is a causal network.*

For example, if we take our influence diagram for lifestyles decisions given in Figure 1 and ignore the distinction between chance and decision nodes, then we obtain a causal model for $U = \{s, d, g, p_s, c, v, p_e, l, u\}$ (by Pearl and Verma's (1991) definition). Furthermore, if we introduce set decisions for every node in U, then we obtain an influence diagram that is a causal network for U (by our definition).

We note that our definition of causal network explicitly allows some or all chance node in U to have no corresponding set decision. Examples of causal networks of this kind include the influence diagrams in Figures 2b and c. In the same spirit, Pearl and Verma do not require that all nodes in U have realizable set decisions.

Pearl and Verma (1991, p. 3) go on to define a causal theory:

Definition 15 (Causal Theory, Pearl and Verma) *A* causal theory *is a pair $T =< D, \Theta_D >$ consisting of a causal model D and a set of parameters Θ_D compatible with D. Θ_D assigns a function $x_i = f_i[Pa(x_i), \epsilon_i]$ and a probability measure g_i to each $x_i \in U$, where $Pa(x_i)$ are the parents of x_i in D and each ϵ_i is a random disturbance distributed according to g_i, independently of the other ϵ's and of any preceding variable $x_j : 0 < j < i$.*

This causal theory is closely related to HCF influence diagrams. In particular, let us suppose that we are given a causal model for U by Pearl and Verma's definition. First, create a causal network as we have just described. Next, transform the influence diagram to HCF, assuming the causal mechanisms are independent. Finally, collapse the set-decision nodes back into their corresponding chance nodes. The result will be a causal theory, where the causal mechanism for node x_i (by our definition) corresponds to the random disturbance ϵ_i (by Pearl and Verma's definition). Indeed, Pearl and Verma state later in the text that the random disturbances are not affected by set decisions. This condition corresponds to our requirement that causal mechanisms be unresponsive to all decisions in the domain.

We note that HCF is more general than is Pearl and Verma's causal theory. Namely, in HCF, causal mechanisms may be dependent. The HCF influence diagram in Figure 6b illustrates a case where there is such dependence. The generality of HCF suggests that there are some problems that can be represented by an HCF influence diagram but not by a causal theory. This conclusion, however, is not so. Rather, the generality of HCF leads to a practical advantage. For example, to represent the relationships in Figure 6b using a causal theory, we would introduce mapping variables (or random disturbances) $c(s, g)$ and $v(d, g)$. Assuming s, d, g, c, and v are binary variables, each mapping variable in the causal theory would have 16 states, whereas each mapping variable in Figure 6b has only four states. Consequently, the nodes $c(s, g)$ and $v(d, g)$ would require 30 probabilities in total, whereas the nodes $c(s)$ and $v(d)$ require only 12 probabilities in total.

9 Counterfactual Reasoning

Causal models have been used to answer counterfactual queries that are more complex than the simple counterfactual queries we use to define unresponsiveness [Balke and Pearl, 1994]. For example, methods exist for answering the query "given that I have not smoked, have maintained a good diet, have not gotten lung cancer, and my cardiovascular status has been good, what is the probability that I would have gotten lung cancer had I smoked and eaten poorly?" Methods for counterfactual reasoning can be important, for example, in legal argumentation.

Influence diagrams in HCF also can be used to answer counterfactual queries. For example, to answer our lung-cancer query, we begin with the influence diagram in HCF shown in Figure 6b. Then, we make two copies of all variables that are responsive to the decisions, as shown in Figure 7. The first copy represents the actual state of affairs; and the second copy represents the counterfactuals (in our example, *smoke* = *yes* and *diet* = *poor*). The unresponsive variables are not copied, because, by definition, they cannot be affected by decisions. Also, by definition of causal mechanism, each copy of an observable variable has the same deterministic relationship with its mechanism. To answer our query, we instantiate the decision and chance variables in the first copy of the diagram to their observed values (no smoking, good diet, no lung cancer, and good cardiovascular status, in our example). In addition, we instantiate our counterfactual decisions in the second copy of the diagram. Then, we use a standard belief-network inference method to compute the probability of the variable(s) of interest (*lung cancer* in our example).

Using this approach, we can answer arbitrary counterfactual queries, including queries where unresponsive variables have been observed. For example, we can answer the query, "given that I have not smoked, have maintained a good diet, and have a genotype predis-

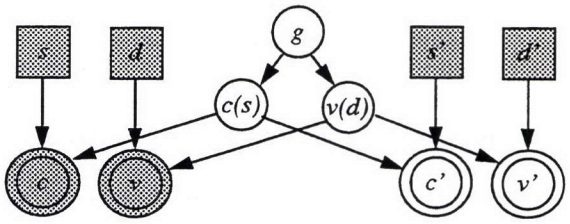

Figure 7: The use of HCF to compute a counterfactual query. The primed variables denote counterfactuals. Shaded variables are instantiated.

posing me to lung cancer, what is the probability that I would have gotten lung cancer had I smoked."

Our HCF-influence-diagram approach is identical to that of Balke and Pearl (1994), except that they use Pearl and Verma's (1991) causal theory, whereas we use the HCF influence diagram as the base representation. Therefore, our method has the computational advantage described in Section 8.

10 Conclusions and Future Work

We have presented a precise definition of cause and effect in terms of the more fundamental notion of unresponsiveness. Our definition departs from the traditional view of causation in that our causal assertions are made relative to a set of decisions. Also, our definition allows for models where only some dependencies have a causal explanation. We have argued that these properties are advantageous.

In addition, we have examined the graphical encoding of causation. We have shown how the ordinary influence diagram is sometimes inadequate as a graphical representation of cause, but that the HCF influence diagram is always an accurate language for causal dependence. Also, we have established correspondences between Pearl and Verma's (1991) causal model and theory and the influence diagram.

An important aspect of causality that we have barely touched upon in this paper is the notion of time. Although many of the results presented here are applicable to time-varying domains, where two different nodes in an influence diagram may represent the same system variable at different points in time, there are interesting aspects of such domains that we have yet to explore.

Acknowledgments

We thank Jack Breese, Tom Chavez, Eric Horvitz, Ron Howard, Judea Pearl, Mark Peot, Glenn Shafer, and Patrick Suppes for useful comments.

References

[Balke and Pearl, 1994] Balke, A. and Pearl, J. (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA. Morgan Kaufmann.

[Druzdzel and Simon, 1993] Druzdzel, M. and Simon, H. (1993). Causality in Bayesian belief networks. In *Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence,* Washington, DC, pages 3–11. Morgan Kaufmann.

[Heckerman et al., 1994] Heckerman, D., Breese, J., and Rommelse, K. (1994). Sequential troubleshooting under uncertainty. In *Proceedings of Fifth International Workshop on Principles of Diagnosis,* New Paltz, NY, pages 121–130.

[Heckerman and Shachter, 1994] Heckerman, D. and Shachter, R. (1994). A decision-based view of causality. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA, pages 302–310. Morgan Kaufmann.

[Howard, 1990] Howard, R. (1990). From influence to relevance to knowledge. In Oliver, R. and Smith, J., editors, *Influence Diagrams, Belief Nets, and Decision Analysis*, chapter 1. Wiley and Sons, New York.

[Howard and Matheson, 1981] Howard, R. and Matheson, J. (1981). Influence diagrams. In Howard, R. and Matheson, J., editors, *Readings on the Principles and Applications of Decision Analysis*, volume II, pages 721–762. Strategic Decisions Group, Menlo Park, CA.

[Pearl, 1994] Pearl, J. (1994). Causal diagrams for empirical research. Technical Report R-218-L, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles.

[Pearl and Verma, 1991] Pearl, J. and Verma, T. (1991). A theory of inferred causation. In Allen, J., Fikes, R., and Sandewall, E., editors, *Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, New York.

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