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

This paper is a summary of work presented in a technical report, which may be obtained via anonymous ftp at research.microsoft.com://pub/tech-reports/spring94/tr-94-10.ps.

1 Introduction

In this paper, we offer three improvements to current work in causal reasoning. First, current approaches either take causality as a primitive notion, or provide only a fuzzy, intuitive definition of cause and effect. 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, 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.

Third, we describe a special type of an influence diagram known as Howard Canonical Form [Howard, 1990], developed for computing value of information, and show how it can be used to represent causal relationships more efficiently than existing representations.

2 Background

Fundamental to our discussion is the distinction between an *uncertain variable* and *a decision variable*. The state of a decision variable is an action chosen by a person, usually called the decision maker. In contrast, an uncertain variable is uncertain and its state may be at most indirectly affected by the decision maker's choices. This distinction is made at modeling time by the decision maker or his agent. For example, a decision maker who wants to determine whether or not to smoke would deem the variable *smoke* to be a decision variable, whereas this same person would deem the variable *lung cancer*, representing whether or not he develops lung cancer, to be an uncertain 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. We use a probability distribution $P\{X|Y\}$ to represent a decision maker's uncertainty about X, given that a set of uncertain and/or decision variables Y is known or determined.

We are interested in modeling relationships in a *domain* consisting of uncertain variables U and decision variables D. In this paper, we use the influence-diagram representation to illustrate some of our concepts. We assume that the reader is familiar with this representation.

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.

In the example that we have considered, we have implicitly assumed 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 $\{d = \text{heads}, c = \text{heads}, w = \text{win}\}$

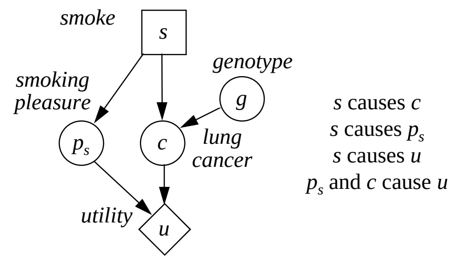

Figure 1: A decision of whether or not to smoke.

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, let us consider a simple decision problem of whether or not to smoke, illustrated by the influence diagram in Figure 1. (The arcs are suggestive of causal relationships. Nonetheless, the reader should resist this interpretation until the end of the next section.) Consider the variables *smoke*, *smoking pleasure*, and *lung cancer*, and *utility*. The variable *utility* is responsive to the decision *smoke*. Nonetheless, if we consider only the counterfactual worlds in which the variables *smoking pleasure* and *lung cancer* take on the same instance, then *utility* will be the same. We say that *utility* is unresponsive to *smoke* in counterfactual worlds where *smoking pleasure* and *lung cancer* is the same, or that *utility is unresponsive to smoke in worlds limited by* $\{$*smoking pleasure*, *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?"

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 decision problem of whether to smoke, 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.*

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. Also, 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.

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, $x \hookleftarrow D$, and C is a minimal set such that $x \not\hookleftarrow_C D$.*

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.

Our decision problem of whether to smoke helps us to illustrate the definition. As mentioned, it is reasonable to assert that *lung cancer* is responsive to $D = \{smoke\}$. Also, it is true trivially that *lung cancer* is conditionally unresponsive to D given $\{smoke\}$. 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 responsive variable x.

Also, it is reasonable to assert that *utility* is responsive to D, *utility* is conditionally unresponsive to D given $\{smoking\ pleasure,\ lung\ cancer\}$, and there is no subset C of $\{smoking\ pleasure,\ lung\ cancer\}$

such that *utility* is conditionally unresponsive to D given C. Therefore, we can conclude that $\{$*smoking pleasure*, *lung cancer*$\}$ are causes for *utility*. We may also conclude that $\{$*smoke*$\}$ is a cause for *utility*. This example illustrates an important property of our definition: causes are not unique.

Someday, it may be possible to use retroviral therapy to alter one's genetic makeup. Assuming that a decision of whether or not to undergo such therapy is available, it is reasonable to assert that *lung cancer* is responsive to $D = \{$*smoke*, *retroviral therapy*$\}$ and that $\{$*smoke*, *genotype*$\}$ is a minimal set C such that *lung cancer* is conditionally unresponsive to D given C. Thus, we can conclude that $\{$*smoke*, *genotype*$\}$ are causes for *lung cancer*. This example demonstrates that the conclusions drawn about cause and effect, given our definition, depend on what decisions are available. Thus, as in our formal definition, we say that $\{$*smoke*, *genotype*$\}$ are causes for *lung cancer* with respect to $\{$*smoke*, *retroviral therapy*$\}$.

These examples illustrates a benefit of defining cause with respect to a set of decisions. Given our definition, not all dependencies need have a causal explanation. For example, without a retroviral therapy or any other alternative for modifying genotype, 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.

5 Mapping Variables and Causal Mechanisms

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 developed for the purpose of computing value of information, it turns out that an HCF influence diagram accurately and efficiently encodes causal relationships. In the remainder of this paper, we describe HCF and show how it can be used to represent causal relationships.

An important concept concerning HCF is that of a mapping variable, which we discuss in this section. To understand what a mapping variable is, let us consider the relationship between the decision *smoke* (s) and the uncertain 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. That is, each state of $c(s)$ represents a possible set of outcomes for c, given all possible choices for s. The states of $c(s)$ are shown in Table 1.

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

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 $c(s)$.

When the argument of a mapping variable contains uncertain variables, things get more complicated. For example, in our decision problem of whether to smoke, consider a new variable *length of life* (l) and the mapping variable $l(c)$. It may be that, whether or not the decision maker smokes, he will not get lung cancer. In this case, $l(c)$ appears to have no meaning, because c does not take on the state "true." Nonetheless, we can imagine a decision where we directly set the variable *lung cancer* to each of its states, in which case, the mapping variable $l(c)$ becomes well defined.

The notion of "directly setting" a variable is a bit tricky. For example, we can imagine setting the variable *weather forecast* to be "rain" by cloud seeding, but this action is not a direct setting of the variable due to its side effects on the variable *weather*. In contrast, changing the cue cards of the weatherman on the nightly news can be considered a direct setting of *weather forecast*. Pearl and Verma (1991) discuss the notion of directly setting a variable, taking this concept to be primitive. Here, we define a *set decision*, using limited unresponsiveness to capture the meaning of "direct."

Definition 5 (Set Decision) *Given uncertain variables U and decisions D, a* set decision for $x \in U$ with respect to U and D, denoted $\hat{x}$, *is a decision variable in D such that (1) $\hat{x}$ has alternatives "set x to k" for each possible state k of x and "do nothing," and (2) for all variables $y \in U$, y is unresponsive to D in worlds limited by $\{x\} \cup D \setminus \{\hat{x}\}$.*

The idea behind the second condition is that all uncertain variables must assume the same state whether x remains not set ($\hat{x}$="do nothing") or x is directly set to the state it would have assumed had it not been set. As a matter of notation, we use $\hat{Y}$ to refer to the collection of set decisions corresponding to the variables in Y. We can now formally define a mapping variable.

Definition 6 (Mapping Variable) *Given uncertain variables X and variables $Y = Y_D \cup Y_C$, where Y_D and Y_C are sets of decision and uncertain variables, respectively, the* mapping variable $X(Y)$ *is the uncertain variable that represents all possible mappings from $Y_D \cup \hat{Y}_C$ to X.*

Note that the decisions $\hat{Y}_C$ need only be imagined. They need not be realizable in practice.

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.

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, and most important, we have the following theorem.

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

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 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) $x \notin C$, (2) x is responsive to D, and (3) C is a minimal set of variables such that $x(C)$ is unresponsive to D.*

We can think of $x(C)$—where C are causes for x—as a *causal mechanism* that relates C and x. For example, suppose uncertain 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 7 (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 Howard Canonical Form

We can now define HCF. Although Howard (1990) does not use our language, his definition is equivalent to the following:

Definition 8 (Howard Canonical Form) *An influence diagram for uncertain variables U and decisions D is said to be in* Howard Canonical Form *if (1) every uncertain node that is not a descendant of a decision node is unresponsive to D, and (2) every uncertain 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 2a is shown in Figure 2b. In this new influence diagram, we have added a node corresponding to the causal mechanism $c(s)$. This node becomes the only non-deterministic uncertain node, and is unresponsive to $D = \{smoke\}$.

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

Theorem 3 (Howard Canonical Form) *Given uncertain variables U and decisions D, an influence diagram in HCF for $U \cup D$ can be constructed as follows:*[1]

1. *Add a node to the diagram corresponding to each variable in $U \cup D$*

[1] We are not concerned with information arcs and utility nodes in this construction.

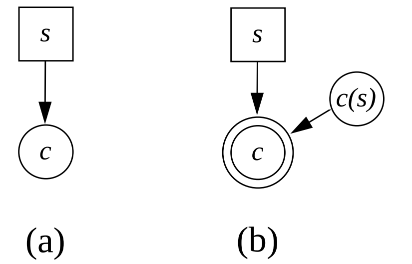

Figure 2: A transformation to Howard Canonical Form.

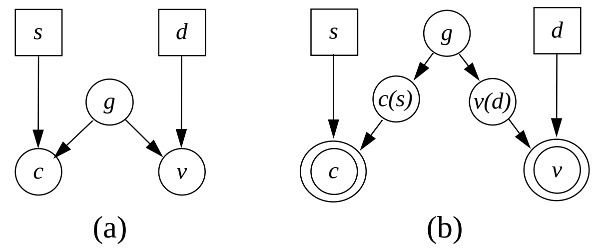

Figure 3: Another transformation to Howard Canonical Form.

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* $x_i \hookleftarrow D$,

- *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* $C_i \cup x_i(C_i)$

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 3a. 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 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 3b. This example illustrates an important point that causal mechanisms may be dependent.

From our construction, it follows that every responsive variable x_i has at least one set of causes explicitly encoded in the diagram (C_i). That is, an HCF constructed as in Theorem 3 accurately represents a set of causes for every caused variable. In addition, HCF is an efficient representation of cause. Namely, using HCF, we can simplify assessments by using our knowledge of what variables can be controlled. For example, in the decision problem corresponding to Figure 3, if there was a decision that affected g (e.g., *retroviral therapy*), then we would have to construct and assess the mapping variables $c(s, g)$ and $v(d, g)$ (each having 16 states assuming s, c, d, v, g are binary). Because there is no such decision, however, we can construct and assess the mapping variables $c(s)$ and $v(d)$ (each having only four states).

We note that Pearl's causal theory (e.g., Pearl and Verma, 1991, Pearl, 1994) is essentially equivalent to HCF when every uncertain variable has a corresponding set decision. (See our technical report for a detailed discussion of this point.) In particular, Pearl's representation does not make use of a decision maker's knowledge about which variables can be controlled, and is consequently less efficient than HCF in many circumstances.

References

[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.

[Pearl, 1994] Pearl, J. (1994). A probabilistic calculus of actions. In *Proceedings of Tenth Conference on Uncertainty in Artificial Intelligence,* Seattle, WA. Morgan Kaufmann.

[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.