# Part III: Causality

# 17

# Instrumental Sets

Carlos Brito

## 1 Introduction

The research of Judea Pearl in the area of causality has been very much acclaimed. Here we highlight his contributions for the use of graphical languages to represent and reason about causal knowledge.[1]

The concept of causation seems to be fundamental to our understanding of the world. Philosophers like J. Carroll put it in these terms: "With regard to our total conceptual apparatus, causation is the center of the center" [Carroll 1994]. Perhaps more dramatically, David Hume states that causation together with resemblance and contiguity are "the only ties of our thoughts, ... for us the cement of the universe" [Hume 1978]. In view of these observations, the need for an adequate language to talk about causation becomes clear and evident.

The use of graphical languages was present in the early times of causal modelling. Already in 1934, Sewall Wright [Wright 1934] represented the causal relation among several variables with diagrams formed by points and arrows (i.e., a directed graph), and noted that the correlations observed between the variables could be associated with the various paths between them in the diagram. From this observation he obtained a method to estimate the strength of the causal connections known as The Method of Path Coefficients, or simply Path Analysis.

With the development of the research in the field, the graphical representation gave way to a mathematical language, in which causal relations are represented by equations of the form $Y = \alpha+\beta X+e$. This movement was probably motivated by an increasing interest in the quantitative aspects of the model, or by the rigorous and formal appearance offered by the mathematical language. However it may be, the consequence was a progressive departure from our basic causal intuitions. Today people ask whether such an equation represents a functional or a causal relation [Reiss 2005]. Sewall Wright and Judea Pearl would presumably answer: "Causal, of course!".

## 2 The Identification Problem

We explore the feasibility of inferring linear cause-effect relationships from various combinations of data and theoretical assumptions. The assumptions are represented

[1] This contribution is a simplified version of a joint paper with Judea Pearl in UAI 2002. A great deal of technicality was removed, and new discussion was added, in the hope that the reader will be able to easily follow and enjoy the argument.


Figure 1. (a) a bow-pattern; and (b) a bow-free model.

in the form of an acyclic causal diagram, which contains both arrows and bidirected arcs [Pearl 1995; Pearl 2000a]. The arrows represent the potential existence of direct causal relationships between the corresponding variables, and the bidirected arcs represent spurious correlations due to unmeasured common causes. All interactions among variables are assumed to be linear. Our task is to decide whether the assumptions represented in the diagram are sufficient for assessing the strength of causal effects from non-experimental data, and, if sufficiency is proven, to express the target causal effect in terms of estimable quantities.

This decision problem has been tackled in the past half century, primarily by econometricians and social scientists, under the rubric "The Identification Problem" [Fisher 1966] - it is still unsolved. Certain restricted classes of models are nevertheless known to be identifiable, and these are often assumed by social scientists as a matter of convenience or convention [Duncan 1975]. A hierarchy of three such classes is given in [McDonald 1997]: (1) no bidirected arcs, (2) bidirected arcs restricted to root variables, and (3) bidirected arcs restricted to variables that are not connected through directed paths.

In a further development [Brito and Pearl 2002], we have shown that the identification of the entire model is ensured if variables standing in direct causal relationship (i.e., variables connected by arrows in the diagram) do not have correlated errors; no restrictions need to be imposed on errors associated with indirect causes. This class of models was called "bow-free", since their associated causal diagrams are free of any "bow-pattern" [Pearl 2000a] (see Figure 1).

Most existing conditions for identification in general models are based on the concept of Instrumental Variables (IV) [Pearl 2000b; Bowden and Turkington 1984]. IV methods take advantage of conditional independence relations implied by the model to prove the identification of specific causal-effects. When the model is not rich in conditional independence relations, these methods are not informative. In [Brito and Pearl 2002] we proposed a new graphical criterion for identification which does not make direct use of conditional independence, and thus can be successfully applied to models in which the IV method would fail.

The result presented in this paper is a generalization of the graphical version

of the method of instrumental variables, offered by Judea Pearl [Pearl 2000a], to deal with several parameters of the model simultaneously. The traditional method of instrumental variables involves conditions on the independence of the relevant variables and on the rank of a certain matrix of correlations [McFadden ]. The first of these is captured by the notion of d-separation. As for the second, since we know from [Wright 1934] that correlations correspond to paths in the causal diagram, we can investigate which structural properties of the model give rise to the proper conditions of the IV method. The results are graphical criteria that allow us to conclude the identification of some parameters from consideration of the qualitative information represented in the causal diagram.

## 3 Linear Models and Identification

A linear model for the random variables $Y_1, \ldots, Y_n$ is defined by a set of equations of the form:

$$(1) \quad Y_j = \sum_i c_{ji} Y_i + e_j, \quad j = 1, \ldots, n$$

An equation $Y = cX + e$ encodes two distinct assumptions: (1) the possible existence of (direct) causal influence of $X$ on $Y$; and, (2) the absence of causal influence on $Y$ of any variable that does not appear on the right-hand side of the equation. The parameter $c$ quantifies the (direct) causal effect of $X$ on $Y$. That is, the equation claims that a unit increase in $X$ would result in $c$ units increase of $Y$, assuming that everything else remains the same. The variable $e$ is called an error or disturbance; it represents unobserved background factors that the modeler decides to keep unexplained; this variable is assumed to have a normal distribution with zero mean.

The specification of the equations and the pairs of error-terms $(e_i, e_j)$ with non-zero correlation defines the structure of the model. This structure can be represented by a directed graph, called causal diagram, in which the set of nodes is defined by the variables $Y_1, \ldots, Y_n$, and there is a directed edge from $Y_i$ to $Y_j$ if $Y_i$ appears on the right-hand side of the equation for $Y_j$. Additionally, if error-terms $e_i$ and $e_j$ are assumed to have non-zero correlation, we add a (dashed) bidirected edge between $Y_i$ and $Y_j$. Figure 2 shows a model with the respective causal diagram.

In this work, we consider only recursive models, which are defined by the restriction that $c_{ji} = 0$, for all $i \geq j$. This simply means that the directed edges in the causal diagram do not form cycles.

The set of parameters of the model, denoted by $\Theta$, is formed by the coefficients $c_{ij}$ and the non-zero entries of the error covariance matrix $\Psi$, $[\Psi_{ij}] = cov(e_i, e_j)$.

Fixing the structure and assigning values to the parameters $\Theta$, the model determines a unique covariance matrix $\Sigma$ over the observed variables $Y_1, \ldots, Y_n$, given by (see [Bollen 1989], page 85):

$$(2) \quad \Sigma(\Theta) = (I - C)^{-1} \Psi [(I - C)^{-1}]'$$

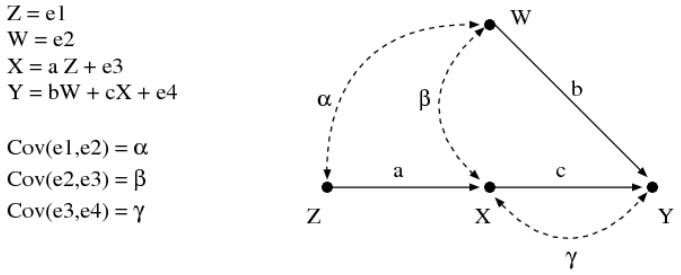



Figure 2. A simple linear model and its causal diagram.

where $C$ is the matrix of coefficients $c_{ji}$.

Conversely, in the Identification Problem, after fixing the structure of the model, one attempts to solve for $\Theta$ in terms of the observed covariance $\Sigma$. This is not always possible. In some cases, no parametrization of the model is compatible with a given $\Sigma$. In other cases, the structure of the model may permit several distinct solutions for the parameters. In these cases, the model is called *non-identified*.

However, even if the model is non-identified, some parameters may still be uniquely determined by the given assumptions and data. Whenever this is the case, the specific parameters are said to be *identified*.

Finally, since the conditions we seek involve the structure of the model alone, and do not depend on the numerical values of the parameters $\Theta$, we insist only on having identification almost everywhere, allowing few pathological exceptions. The concept of identification almost everywhere can be formalized as follows.

Let $h$ denote the total number of parameters in the model. Then, each vector $\Theta \in \Re^h$ defines a parametrization of the model. For each parametrization $\Theta$, the model $G$ generates a unique covariance matrix $\Sigma(\Theta)$. Let $\Theta(\lambda_1, \ldots, \lambda_n)$ denotes the vector of values assigned by $\Theta$ to the parameters $\lambda_1, \ldots, \lambda_n$.

Parameters $\lambda_1, \ldots, \lambda_n$ are identified almost everywhere if

$$\Sigma(\Theta) = \Sigma(\Theta') \quad \text{implies} \quad \Theta(\lambda_1, \ldots, \lambda_n) = \Theta'(\lambda_1, \ldots, \lambda_n)$$

except when $\Theta$ resides on a subset of Lebesgue measure zero of $\Re^h$.

## 4 Graph Background

DEFINITION 1.

1. A *path* in a graph is a sequence of edges such that each pair of consecutive edges share a common node, and each node appears only once along the path.

2. A *directed path* is a path composed only by directed edges, all of them oriented

in the same direction. If there is a directed path going from $X$ to $Y$ we say that $Y$ is a *descendant* of $X$.

3. A path is *closed* if it has a pair of consecutive edges pointing to their common node (e.g., $\ldots \rightarrow X \leftarrow \ldots$ or $\ldots \leftrightarrow X \leftarrow \ldots$). In this case, the common node is called a *collider*. A path is *open* if it is not closed.

DEFINITION 2. A path $p$ is blocked by a set of nodes $\mathbf{Z}$ (possibly empty) if either

1. $\mathbf{Z}$ contains some non-collider node of $p$, or

2. at least one collider of $p$ and all of its descendants are outside $\mathbf{Z}$.

The idea is simple. If the path is closed, then it is naturally blocked by its colliders. However, if a collider, or one of its descendants, belongs to $\mathbf{Z}$, then it ceases to be an obstruction. But if a non-collider of $p$ belongs to $\mathbf{Z}$, then the path is definitely blocked.

DEFINITION 3. A set of nodes $\mathbf{Z}$ d-separates $X$ and $Y$ if $\mathbf{Z}$ simultaneously blocks all the paths between $X$ and $Y$. If $\mathbf{Z}$ is empty, then we simply say that $X$ and $Y$ are d-separated.

The significance of this definition comes from a result showing that if $X$ and $Y$ are d-separated by $\mathbf{Z}$ in the causal diagram of a linear model, then the variables $X$ and $Y$ are conditionally independent given $\mathbf{Z}$ [Pearl 2000a]. It is this sort of result that makes the connection between the mathematical and graphical languages, and allows us to express our conditions for identification in graphical terms.

DEFINITION 4. Let $p_1, \ldots, p_n$ be unblocked paths connecting the variables $Z_1, \ldots, Z_n$ and the variables $X_1, \ldots, X_n$, respectively. We say that the set of paths $p_1, \ldots, p_n$ is incompatible if we cannot rearrange their edges to form a different set of unblocked paths $p'_1, \ldots, p'_n$ between the same variables.

A set of disjoint paths (i.e., paths with no common nodes) consists in a simple example of an incompatible set of paths.

## 5 Instrumental Variable Methods

### 5.1 Identification of a Single Parameter

The method of Instrumental Variables (IV) for the identification of causal effects is intended to address the situation where we cannot attribute the entire correlation between two variables, say $X$ and $Y$, to their causal connection. That is, part of the correlation between $X$ and $Y$ is due to common causes and/or correlations between disturbances. Figure 3 shows examples of this situation.

In the simplest cases, like in Figure 3a, we can find a conditioning set $\mathbf{W}$ such that the partial correlation of $X$ and $Y$ given $\mathbf{W}$ can indeed be attributed to the causal relation. In this example, if we take $\mathbf{W} = \{W\}$ we eliminate the source

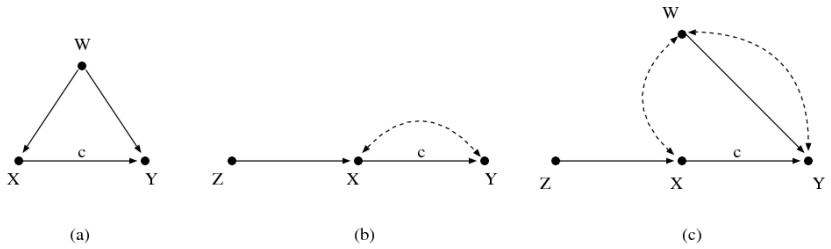



Figure 3. Models with spurious correlation between $X$ and $Y$.

of spurious correlation. The causal effect of $X$ on $Y$ is identified and given by $c = \sigma_{XY.\mathbf{W}}$.

There are cases, however, where this idea does not work, either because the spurious correlation is originated by disturbances outside the model (Figure 3b), or else because the conditioning itself introduces spurious correlations (Figure 3c). In situations like these, the IV method asks us to look for a variable $Z$ with the following properties [Bowden and Turkington 1984]:

IV-1. $Z$ is not independent of $X$.

IV-2. $Z$ is independent of all error terms that have an influence on $Y$ that is not mediated by $X$.

The first condition simply states that there is a correlation between $Z$ and $X$. The second condition says that the only source of correlation between $Z$ and $Y$ is due to a covariation bewteen $Z$ and $X$ that subsequently affects $Y$ through the causal connection $X \xrightarrow{c} Y$.

If we can find a variable $Z$ with these properties, then the causal effect of $X$ on $Y$ is identified and given by $c = \sigma_{ZY}/\sigma_{ZX}$.

Using the notion of d-separation we can express the conditions (1) and (2) of the IV method in graphical terms, thus obtaining a criterion for identification that can be applied directly to the causal diagram of the model. Let $G$ be the graph representing the causal diagram of the model, and let $G_c$ be the graph obtained after removing the edge $X \xrightarrow{c} Y$ from $G$ (see Figure 4). Then, $Z$ is an instrumental variable relative to $X \xrightarrow{c} Y$ if:

1. $Z$ is not d-separated from $X$ in $G_c$.

2. $Z$ is d-separated from $Y$ in $G_c$.

Using this criterion, it is easy to verify that $Z$ is an instrumental variable relative to $X \xrightarrow{c} Y$ in the models of Figure 3b and c.

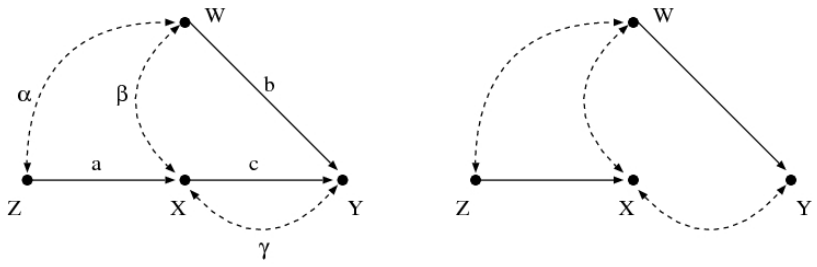



Figure 4. The causal diagram $G$ of a linear model and the graph $G_c$.

### 5.2 Conditional Instrumental Variables

A generalization of the method of instrumental variables is offered through the use of conditioning. A conditional instrumental variable is a variable $Z$ that may not have the properties (IV-1) and (IV-2) above, but after conditioning on a subset $\mathbf{W}$ these properties do hold. When such pair $(Z, \mathbf{W})$ is found, the causal effect of $X$ on $Y$ is identified and given by $c = \sigma_{ZY.\mathbf{W}}/\sigma_{ZX.\mathbf{W}}$.

Again, we obtain a graphical criterion for a conditional IV using the notion of d-separation. Variable $Z$ is a conditional instrumental variable relative to $X \stackrel{c}{\rightarrow} Y$ given $\mathbf{W}$, if

1. $\mathbf{W}$ contains only non-descendants of $Y$.
2. $\mathbf{W}$ does not d-separate $Z$ from $X$ in $G_c$.
3. $\mathbf{W}$ d-separates $Z$ from $Y$ in $G_c$.

### 5.3 Identification of Multiple Parameters

So far we have been concerned with the identification of a single parameter of the model, but in its full version the method of instrumental variables allows to prove simultaneously the identification of several parameters in the same equation (i.e., the causal effects of several variables $X_1, \ldots, X_k$ on the same variable $Y$).

Following [McFadden ], assume that we have the equation

$$Y = c_1 X_1 + \ldots + c_k X_k + e$$

in our linear model. The variables $Z_1, \ldots, Z_j$, with $j \geq k$, are called instruments if

1. The matrix of correlations between the variables $X_1, \ldots, X_k$ and the variables $Z_1, \ldots, Z_j$ is of maximum possible rank (i.e., rank $k$).
2. The variables $Z_1, \ldots, Z_j$ are uncorrelated with the error term $e$.

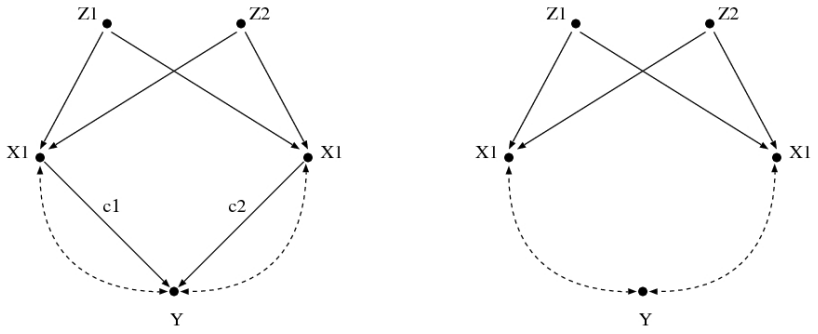



Figure 5. The causal diagram $G$ of a linear model and the graph $\bar{G}$.

Next, we develop our graphical intuition and obtain a graphical criterion for identification that corresponds to the full version of the IV method.

Consider the model in Figure 5a. Here, the variables $Z_1$ and $Z_2$ do not qualify as instrumental variables (or even conditional IVs) with respect to either $X_1 \stackrel{c_1}{\rightarrow} Y$ or $X_2 \stackrel{c_2}{\rightarrow} Y$. But, following ideas similar to the ones developed in the previous sections, in Figure 5b we show the graph obtained by removing edges $X_1 \rightarrow Y$ and $X_2 \rightarrow Y$ from the causal diagram. Observe that now both d-separation conditions for an instrumental variable hold for $Z_1$ and $Z_2$. This leads to the idea that $Z_1$ and $Z_2$ could be used together as instruments to prove the identification of parameters $c_1$ and $c_2$. Indeed, next we give a graphical criterion that is sufficient to guarantee the identification of a subset of parameters of the model.

Fix a variable $Y$, and consider the edges $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ in the causal diagram $G$ of the model. Let $\bar{G}$ be the graph obtained after removing the edges $X_1 \rightarrow Y, \ldots, X_k \rightarrow Y$ from $G$. The variables $Z_1, \ldots, Z_k$ are instruments relative to $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ if

1. There exists an incompatible set of unblocked paths $p_1, \ldots, p_k$ connecting the variables $Z_1, \ldots, Z_k$ to the variables $X_1, \ldots, X_k$.

2. The variables $Z_i$ are d-separated from $Y$ in $\bar{G}$.

3. Each variable $Z_i$ is not d-separated from the corresponding variable $X_i$ in $\bar{G}$. [2]

THEOREM 5. *If we can find variables $Z_1, \ldots, Z_k$ satisfying the conditions above, then the parameters $c_1, \ldots, c_k$ are identified almost everywhere, and can be computed by solving a system of linear equations.*

[2] Notice that this condition is redundant, since it follows from the first condition.

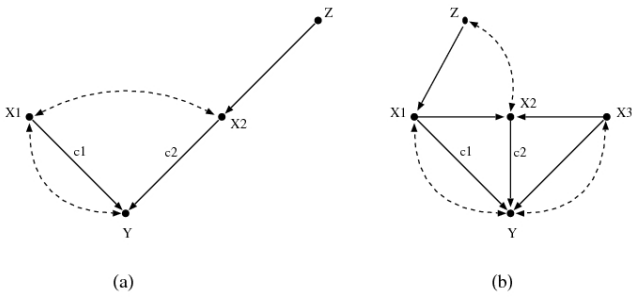



Figure 6. More examples of the new criterion.

Figure 6 shows more examples of application of the new graphical criterion. Model (a) illustrates an interesting case, in which variable $X_2$ is used as the instrumental variable for $X_1 \rightarrow Y$, while $Z$ is the instrumental variable for $X_2 \rightarrow Y$. Finally, in model (b) we have an example in which the parameter of edge $X_3 \rightarrow Y$ is non-identified, and still the graphical criterion allows to show the identfication of $c_1$ and $c_2$.

## 6 Wright's Method of Path Coefficients

Here, we describe an important result introduced by Sewall Wright [Wright 1934], which is extensively explored in our proofs.

Given variables $X$ and $Y$ in a recursive linear model, the correlation coefficient of $X$ and $Y$, denoted by $\rho_{XY}$, can be expressed as a polynomial on the parameters of the model. More precisely,

$$\sigma_{XY} = \sum_{p} T(p) \quad (3)$$

where the summation ranges over all unblocked paths $p$ between $X$ and $Y$, and each term $T(p)$ represents the contribution of the path $p$ to the total correlation between $X$ and $Y$. The term $T(p)$ is given by the product of the parameters of the edges along the path $p$. We refer to Equation 3 as Wright's equation for $X$ and $Y$.

Wright's method of path coefficients for identification consists in forming Wright's equations for each pair of variables in the model, and then solving for the parameters in terms of the observed correlations. Whenever there is a unique solution for a parameter $c$, this parameter is identified.

## 7 Proof of Theorem 1

### 7.1 Notation

Fix a variable $Y$ in the model. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be the set of all non-descendants of $Y$ which are connected to $Y$ by an edge. Define the following set of edges incoming $Y$:

$$(4) \quad Inc(Y) = \{(X_i, Y) : X_i \in \mathbf{X}\}$$

Note that for some $X_i \in \mathbf{X}$ there may be more than one edge between $X_i$ and $Y$ (one directed and one bidirected). Thus, $|Inc(Y)| \geq |\mathbf{X}|$. Let $\lambda_1, \ldots, \lambda_m$, $m \geq k$, denote the parameters of the edges in $Inc(Y)$.

It follows that edges $X_1 \stackrel{c_1}{\rightarrow} Y, \ldots, X_k \stackrel{c_k}{\rightarrow} Y$ all belong to $Inc(Y)$, because $X_1, \ldots, X_k$ are clearly non-descendants of $Y$. We assume that $\lambda_i = c_i$, for $i = 1, \ldots, k$, while $\lambda_{k+1}, \ldots, \lambda_m$ are the parameters of the remaining edges of $Inc(Y)$.

Let $Z$ be any non-descendant of $Y$. Wright's equation for the pair $(Z, Y)$ is given by:

$$(5) \quad \sigma_{ZY} = \sum_p T(p)$$

where each term $T(p)$ corresponds to an unblocked path $p$ between $Z$ and $Y$. The next lemma proves a property of such paths.

LEMMA 6. *Any unblocked path between $Y$ and one of its non-descendants $Z$ must include exactly one edge from $Inc(Y)$.*

Lemma 6 allows us to write equation 4 as:

$$(6) \quad \sigma_{ZY} = \sum_{j=1}^{m} a_j \cdot \lambda_j$$

Thus, the correlation between $Z$ and $Y$ can be expressed as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$, with no constant term. In addition, we can say something about the coefficients $a_j$. Each term in Equation 5 corresponds to an unblocked path that reaches $Y$ through some egge, say $X_j \stackrel{\lambda_j}{\rightarrow} Y$. When we group the terms together according to the parameter $\lambda_j$ and factor it out, we are, in a sense, removing the edge $X_j \rightarrow Y$ from those paths. Thus, each coefficient $a_j$ in Equation 6 is a sum of terms associated with unblocked paths between $Z$ and $X_j$.

### 7.2 Basic Linear Equations

We have just seen that the correlations between the instrumental variables $Z_i$ and $Y$ can be written as a linear function of the parameters $\lambda_1, \ldots, \lambda_m$:

$$(7) \quad \rho_{Z_iY} = \sum_{j=1}^{m} a_{ij} \cdot \lambda_j$$

Next, we prove an important result

LEMMA 7. *The coefficients $a_{i,k+1}, \ldots, a_{im}$ in Equation 7 are all identically zero.*

**Proof.** The fact that $Z_i$ is d-separated from $Y$ in $\bar{G}$ implies that $\rho_{Z_iY} = 0$ in any probability distribution compatible with $\bar{G}$. Hence, the expression for $\rho_{Z_iY}$ must vanish when evaluated in the causal diagram $\bar{G}$. But this implies that each

coefficient $a_{ij}$ in Equation 7 is identically zero, when the expression is evaluated in $\bar{G}$.

Next, we show that the only difference between the expression for $\rho_{Z_iY}$ on the causal diagrams $G$ and $\bar{G}$ are the coefficients of the parameters $\lambda_1, \ldots, \lambda_k$.

Recall from the previous section that each coefficient $a_{ij}$ is a sum of terms associated with paths which can be extended by the edge $\stackrel{\lambda_j}{\rightarrow} Y$ to form an unblocked path between $Z$ and $Y$.

Fixing $j > k$, we observe that the insertion of edges $x_1 \rightarrow Y, \ldots, X_k \rightarrow Y$ in $\bar{G}$ does not create any new such path (and clearly does not eliminate any existing one). Hence, for $j > k$, the coefficients $a_{ij}$ in the expression for $\rho_{Z_iY}$ in the causal diagrams $G$ and $\bar{G}$ are exactly the same, namely, identically zero. □

The conclusion from Lemma 7 is that the expression for $\rho_{Z_iY}$ is a linear function only of parameters $\lambda_1, \ldots, \lambda_k$:

$$\rho_{Z_iY} = \sum_{j=1}^{k} a_{ij} \cdot \lambda_j \tag{8}$$

### 7.3 System of Equations Φ

Writing Equation 8 for each instrumental variable $Z_i$, we obtain the following system of linear equations on the parameters $\lambda_1, \ldots, \lambda_k$:

$$\Phi = \begin{cases} \rho_{Z_1Y} = a_{11}\lambda_1 + \ldots, a_{1k}\lambda_k \\ \ldots \\ \rho_{Z_kY} = a_{k1}\lambda_1 + \ldots, a_{kk}\lambda_k \end{cases} \tag{9}$$

Our goal now is to show that Φ can be solved uniquely for the parameters $\lambda_i$, and so prove the identification of $\lambda_1, \ldots, \lambda_k$. Next lemma proves an important result in this direction.

Let $A$ denote the matrix of coefficients of Φ.

LEMMA 8. *Det(A) is a non-trivial polynomial on the parameters of the model.*

**Proof.** The determinant of $A$ is defined as the weighted sum, for all permutations $\pi$ of $\langle 1, \ldots, k \rangle$, of the product of the entries selected by $\pi$. Entry $a_{ij}$ is selected by a permutation $\pi$ if the $i^{th}$ element of $\pi$ is $j$. The weights are either 1 or -1, depending on the parity of the permutation.

Now, observe that each diagonal entry $a_{ii}$ is a sum of terms associated with unblocked paths between $Z_i$ and $X_i$. Since $p_i$ is one such path, we can write $a_{ii} = T(p_i) + \hat{a}_{ii}$. From this, it is easy to see that the term

$$T^* = \prod_{j=1}^{k} T(p_j) \tag{10}$$

appears in the product of permutation $\pi = \langle 1, \ldots, n \rangle$, which selects all the diagonal entries of $A$.

We prove that $det(A)$ does not vanish by showing that $T^*$ is not cancelled out by any other term in the expression for det(A).

Let $\tau$ be any other term appearing in the summation that defines the determinant of $A$. This term appears in the product of some permutation $\pi$, and has as factors exactly one term from each entry $a_{ij}$ selected by $\pi$. Thus, associated with such factor there is an unblocked path between $Z_i$ and $X_j$. Let $p'_1, \ldots, p'_k$ be the unblocked paths associated with the factors of $\tau$.

We conclude the proof observing that, since $p_1, \ldots, p_k$ is an incompatible set, its edges cannot be rearranged to form a different set of unblocked paths between the same variables, and so $\tau \neq T^*$. Hence, the term $T^*$ is not cancelled out in the summation, and the expression for $det(A)$ does not vanish. □

### 7.4 Identification of $\lambda_1, \ldots, \lambda_k$

Lemma 8 gives that $det(Q)$ is a non-trivial polynomial on the parameters of the model. Thus, $det(Q)$ only vanishes on the roots of this polynomial. However, [Okamoto 1973] has shown that the set of roots of a polynomial has Lebesgue measure zero. Thus, the system $\Phi$ has unique solution almost everywhere.

It just remains to show that we can estimate the entries of the matrix of coefficients $A$ from the data. But this is implied by the following observation.

Once again, coefficient $a_{ij}$ is given by a sum of terms associated with unblocked paths between $Z_i$ and $X_j$. But, in principle, not every unblocked path between $Z_i$ and $X_j$ would contribute with a term to the sum; just those which can be extended by the edge $X_j \rightarrow Y$ to form an unblocked path between $Z_i$ and $Y$. However, since the edge $X_j \rightarrow Y$ does not point to $X_j$, every unblocked path between $Z_i$ and $X_j$ can be extended by the edge $X_j \rightarrow Y$ without creating a collider. Hence, the terms of all unblocked paths between $Z_i$ and $X_j$ appear in the expression for $a_{ij}$, and by the method of path coefficients, we have $a_{ij} = \rho_{Z_i X_j}$.

We conclude that each entry of matrix $A$ can be estimated from data, and we can solve the system of linear equations $\Phi$ to obtain the parameters $\lambda_1, \ldots, \lambda_k$.

## References

Bollen, K. (1989). *Structural Equations with Latent Variables.* John Wiley, New York.

Bowden, R. and D. Turkington (1984). *Instrumental Variables.* Cambridge Univ. Press.

Brito, C. and J. Pearl (2002). A graphical criterion for the identification of causal effects in linear models. *In Proc. of the AAAI Conference, Edmonton, Canada.*.

Carroll, J. (1994). *Laws of Nature.* Cambridge University Press.

Duncan, O. (1975). *Introduction to Structural Equation Models.* Academic Press.

Fisher, F. (1966). *The Identification Problem in Econometrics.* McGraw-Hill.

Hume, D. (1978). *A Treatise of Human Nature.* Oxford University Press.

McDonald, R. (1997). Haldane's lungs: A case study in path analysis. *Mult. Beh. Res.*, 1–38.

McFadden, D. *Lecture Notes for Econ 240b.* Dept of Economics, UC Berkeley.

Okamoto, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Annals of Statistics*, 763–765.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 669–710.

Pearl, J. (2000a). *Causality: Models, Reasoning and Inference.* Cambridge Press.

Pearl, J. (2000b). Parameter identification: A new perspective. *Technical Report R-276, UCLA.*

Reiss, J. (2005). Causal instrumental variables and interventions. *Philosophy of Science. 72*, 964–976.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statistics.*, 161–215.

# 18

# Seeing and Doing: The Pearlian Synthesis

PHILIP DAWID

## 1 Introduction

It is relatively recently that much attention has focused on what, for want of a better term, we might call "statistical causality", and the subject has developed in a somewhat haphazard way, without a very clear logical basis. There is in fact a variety of current conceptions and approaches [Campaner and Galavotti 2007; Hitchcock 2007; Galavotti 2008]—here we shall distinguish in particular *agency*, *graphical*, *probabilistic* and *modular* conceptions of causality—that tend to be mixed together in an informal and half-baked way, based on "definitions" that often do not withstand detailed scrutiny. In this article I try to unpick this tangle and expose the various different strands that contribute to it. Related points, with a somewhat different emphasis, are made in a companion paper [Dawid 2009].

The approach of Judea Pearl [2009] cuts through this Gordian knot like the sword of Alexander. Whereas other conceptions of causality may be philosophically questionable, definitionally unclear, pragmatically unhelpful, theoretically skimpy, or simply confused, Pearl's theory is none of these. It provides a valuable framework, founded on a rich and fruitful formal theory, by means of which causal assumptions about the world can be meaningfully represented, and their implications developed. Here we will examine both the relationships of Pearl's theory with the other conceptions considered, and its differences from them. We extract the essence of Pearl's approach as an assumption of "modularity", the transferability of certain probabilistic properties between observational and interventional regimes: so, in particular, forging a synthesis between the very different activities of "seeing" and "doing". And we describe a generalisation of this framework that releases it from any necessary connexion to graphical models.

The plan of the paper is as follows. In § 2, I describe the agency, graphical and probabilistic conceptions of causality, and their connexions and distinctions. Section 3 introduces Pearl's approach, showing its connexions with, and differences from, the other theories. Finally, in § 4, I present the generalisation of that approach, emphasising the modularity assumptions that underlie it, and the usefulness of the theory of "extended conditional independence" for describing and manipulating these.

**Disclaimer** I have argued elsewhere [Dawid 2000, 2007a, 2010] that it is important to distinguish arguments about "Effects of Causes" (EoC, otherwise termed "type", or "generic" causality"), from those about "Causes of Effects" (CoE, also termed "token", or "individual" causality); and that these demand different formal frameworks and analyses. My concern here will be entirely focused on problems of generic causality, EoC. A number of

the current frameworks for statistical causality, such as Rubin's "potential response models" [Rubin 1974, 1978], or Pearl's "probabilistic causal models" [Pearl 2009, Chapter 7], are more especially suited for handling CoE type problems, and will not be discussed further here. There are also numerous other conceptions of causality, such as *mechanistic causality* [Salmon 1984; Dowe 2000], that I shall not be considering here.

## 2 Some conceptions of causality

There is no generally agreed understanding of what "causality" is or how it should behave. There are two conceptions in particular that are especially relevant for "statistical causality": *Agency Causality* and *Probabilistic Causality*. The latter in turn is closely related to what we might term *Graphical Causality*.

### 2.1 Agency causality

The "agency" or "manipulability" interpretation of causality [Price 1991; Hausman 1998; Woodward 2003] depends on an assumed notion of external "manipulation" (or "intervention"), that might itself be taken as a primitive—at any rate we shall not try and explicate it further here. The basic idea is that causality is all about how an external manipulation that sets the value of some variable (or set of variables) $X$ will affect some other (unmanipulated) "response variable" (or set of variables) $Y$. The emphasis is usually on comparison of the responses ensuing from different settings $x$ for $X$: a version of the "contrastive" or "difference-making" understanding of causality. Much of Statistical Science—for example, the whole subfield of Experimental Design—aims to address exactly these kinds of questions about the comparative effects of interventions on a system, which are indeed a major object of all scientific enquiry.

We can define certain causal terms quite naturally within the agency theory [Woodward 2003]. Thus we could interpret the statement

"$X$ *has no effect on* $Y$"[1]

as holding whenever, considering regimes that manipulate only $X$, the resulting value of $Y$ (or some suitable codification of uncertainty about $Y$, such as its probability distribution) does not depend on the value $x$ assigned to $X$. When this fails, $X$ *has an effect on* $Y$; we might then go on to quantify this dependence in various ways.

We could likewise interpret

"$X$ *has no (direct) effect on* $Y$*, after controlling for* $W$"

as the property that, considering regimes where we manipulate both $W$ and $X$, when we manipulate $W$ to some value $w$ and $X$ to some value $x$, the ensuing value (or uncertainty) for $Y$ will depend only on $w$, and not further on $x$.

Now suppose that, explicitly or implicitly, we restrict consideration to some collection $\mathcal{V}$ of manipulable variables. Then we might interpret the statement

[1] Just as "zero" is fundamental to arithmetic and "independence" is fundamental to probability, so the concept of "no effect" is fundamental to causality.

"$X$ *is a direct cause of* $Y$ (relative to $\mathcal{V}$)"

(where $\mathcal{V}$ might be left unmentioned, but must be clearly understood) as the negation of "$X$ has no direct effect on $Y$, after controlling for $\mathcal{V} \setminus \{X, Y\}$".[2]

It is important to bear in mind that all these assertions relate to properties of the real world under the various regimes considered: in particular, they can not be given purely mathematical definitions. And in real world problems there are typically various ways of manipulating variables, so we must be very clear as to exactly what is intended.

EXAMPLE 1. **Ideal gas law**

Consider the "ideal gas law":

(1) $\quad PV = kNT$

where $P$ is the absolute pressure of the gas, $V$ is its volume, $N$ is the number of molecules of gas present, $k$ is Boltzmann's constant, and $T$ is the absolute temperature. For our current purposes this will be supposed to be universally valid , no matter how the values of the variables in (1) may have come to arise.

Taking a fixed quantity $N$ of gas in an impermeable container, we might consider interventions on any of $P$, $V$ and $T$. (Note however that, because of the constraint (1), we can not simultaneously and arbitrarily manipulate all three variables.)

An intervention that sets $V$ to $v$ and $T$ to $t$ will lead to the unique value $p = kNt/v$ for $P$. Because this depends on both $v$ and $t$, we can say that there is a *direct effect* of each of $V$ and $T$ on $P$ (relative to $\mathcal{V} = \{V, P, T\}$). Similarly, $P$ has a direct effect on each of $V$ and $T$.

What if we wish to quantify, say, "the causal effect of $V$ on $P$"? Any attempt to do this must take account of the fact that the problem requires additional specification to be well-defined. Suppose the volume of the container can be altered by applying a force to a piston. Initially the gas has $V = v_0$, $P = p_0$, $T = t_0$. We wish to manipulate $V$ to a new value $v_1$. If we do this *isothermally*, *i.e.* by sufficiently slow movement of the piston that, through flow of heat through the walls of the container, the temperature of the gas always remains the same as that of the surrounding heat bath, we will end up with $V = v_1$, $P = p_1 = v_0 p_0/v_1$, $T = t_1 = t_0$. But if we move the piston *adiabatically*, *i.e.* so fast that no heat can pass through the walls of the container, the relevant law is $PV^\gamma =$ constant, where $\gamma = 5/3$ for a monatomic gas. Then we get $V = v_1$, $P = p_1^* = p_0(v_0/v_1)^\gamma$, $T = t_1^* = p_1^* v_1/kN$.

### 2.2 Graphical causality

By *graphical causality* we shall refer to an interpretation of causality in terms of an underlying *directed acyclic graph* (DAG) (noting in passing that other graphical representations are also possible). As a basis for this, we suppose that there is a suitable "causal ambit"[3] $\mathcal{A}$ of variables (not all necessarily observable) that we regard as relevant, and a "causal DAG"

[2] Neapolitan [2003, p. 56] has a different and more complex interpretation of "direct cause".

[3] The importance of the causal ambit will become apparent later.

$\mathcal{D}$ over a collection $\mathcal{V} \subseteq \mathcal{A}$. These ingredients are "known to Nature", though not necessarily to us: $\mathcal{D}$ is "Nature's DAG". Given such a causal DAG $\mathcal{D}$, for $X, Y \in \mathcal{V}$ we interpret "$X$ is a *direct cause* of $Y$" as synonymous with "$X$ is a parent of $Y$ in $\mathcal{D}$", and similarly equate "*cause*" with "ancestor in $\mathcal{D}$". One can also use the causal DAG to introduce further graphically defined causal terms, such as "causal chain", "intermediate variable", ...

The concepts of causal ambit and causal DAG might be regarded as primitive notions, or attempts might be made to define them in terms of pre-existing understandings of causal concepts. In either case, it would be good to have criteria to distinguish a putative causal ambit from a non-causal ambit, and a causal DAG from a non-causal DAG.

For example, we typically read [Hernán and Robins 2006]:

> "A *causal DAG* $\mathcal{D}$ is a DAG in which:
>
> (i). the lack of an arrow from $V_j$ to $V_m$ can be interpreted as the absence of a direct causal effect of $V_j$ on $V_m$ (relative to the other variables on the graph)
>
> (ii). all common causes, even if unmeasured, of any pair of variables on the graph are themselves on the graph.[4]

If we start with a DAG $\mathcal{D}$ over $\mathcal{V}$ that we accept as being a causal DAG, and interpret "direct cause" *etc.* in terms of that, then conditions (i) and (ii) will be satisfied by definition. However, this begs the question of how we are to tell a causal from a non-causal DAG.

More constructively, suppose we start with a prior understanding of the term "direct cause" (relative to $\mathcal{V}$)—for example, though by no means necessarily,[5] based on the agency interpretation described in § 2.1 above. It appears that we could then use the above definition to check whether a proposed DAG $\mathcal{D}$ is indeed "causal". But while this is essentially straightforward so far as condition (i) is concerned (except that there is no obvious reason to require a DAG representation), interpretation and implementation of condition (ii) is more problematic. First, what is a "common cause"? Spirtes et al. [2000, p. 44] say that a variable $X$ is a common cause of variables $Y$ and $Z$ if and only if $X$ is both a direct cause of $Y$ and a direct cause of $Z$ — but in each case relative to the set $\{X, Y, Z\}$, so that this definition is not dependent on the causal ambit $\mathcal{V}$. Neapolitan [2003, p. 57] has a different interpretation, which apparently is relative to an essentially arbitrary set $\mathcal{V}$ — but then states that that problems can arise when at least one common cause is not in $\mathcal{V}$, a possibility that seems to be precluded by his definition.

As another attempt at clarification, Spirtes and Scheines [2004] require "that the set of variables in the causal graph be *causally sufficient*, *i.e.* if $\mathcal{V}$ is the set of variables in the causal graph, that there is no variable $L$ not in $\mathcal{V}$ that is a direct cause (relative to $\mathcal{V} \cup \{L\}$) of two variables in $\mathcal{V}$". If "$L \notin \mathcal{V}$ is not a direct cause of $V \in \mathcal{V}$" is interpreted in agency terms, it would mean that $V$ would not respond to manipulations of $L$, when holding fixed all the other variables in $\mathcal{V}$. But whatever the interpretation of direct cause, such a "definition" of causal sufficiency is ineffective when the range of possible choices

[4] The motivation for this requirement is not immediately obvious, but is related to the defensibility of the *causal Markov* property described in § 2.3 below.

[5] See § 2.2 below.

for the additional variable $L$ is entirely unrestricted—for then how could we ever be sure that it holds, without conducting an infinite search over all unmentioned variables $L$? That is why we posit an appropriate clearly-defined "causal ambit" $\mathcal{A}$: we can then restrict the search to $L \in \mathcal{A}$.

It seems to me that we should, realistically, allow that "causality" can operate, in parallel, at several different levels of granularity. Thus while it may or may not be possible to describe the medical effects of aspirin treatment in terms of quantum theory, even if we could, it would be a category error to try and do so in the context of a clinical trial. So there may be various different causal descriptions of the world, all operating at different levels, each with its associated causal ambit $\mathcal{A}$ of variables and various causal DAGs $\mathcal{D}$ over sets $\mathcal{V} \subseteq \mathcal{A}$. The meaning of any causal terms used should then be understood in relation to the appropriate level of description.

The obvious questions to ask about graphical causality, which are however not at all easy to answer, are: "When can a collection $\mathcal{A}$ of variables be regarded as a causal ambit?", and "When can a DAG be regarded as a causal DAG?".

In summary, so long as we *start* with a DAG $\mathcal{D}$ over $\mathcal{V}$ that we are willing to accept as a *causal* DAG (taken as a primitive concept), we can take $\mathcal{V}$ itself as our causal ambit, and use the structure of $\mathcal{D}$ to *define* causal terms. Without having a prior primitive notion of what constitutes a "causal DAG", however, conditions such as (i) and (ii) are unsatisfactory as a definition. At the very least, they require that we have specified (but how?) an appropriate causal ambit $\mathcal{A}$, relevant to our desired level of description, and have a clear pre-existing understanding (*i.e.* not based on the structure of $\mathcal{D}$, since that would be logically circular) of the terms "direct causal effect", "common cause" (perhaps relative to a set $\mathcal{V}$).

**Agency causality and graphical causality**

It is tempting to use the agency theory as a basis for such prior causal understanding. However, graphical causality does not really sit well with agency causality. For, as seen clearly in Example 1, in the agency intepretation it is perfectly possible for two variables each to have a direct effect on the other—which could not hold under any DAG representation. Similarly [Halpern and Pearl 2005; Hall 2000] there is no obvious reason to expect agency causality to be a transitive relation, which would again be a requirement under the graphical conception. For better or worse, the agency theory does not currently seem to be endowed with a sufficiently rich axiomatic structure to guide manipulations of its causal properties; and however such a general axiomatic structure might look, it would seem unduly restrictive to relate it closely to DAG models.

### 2.3 Probabilistic causality

*Probabilistic Causality* [Reichenbach 1956; Suppes 1970; Spohn 2001] depends on the existence and properties of a probability distribution $P$ over quantities of interest. At its (over-)simplest, it equates *causality* with *probability raising*: "$A$ is a *cause* of $B$" (where $A$ and $B$ are events) if $P(B \mid A) > P(B)$. This is more usefully re-expressed in its null form, and referred to random variables $X$ and $Y$: $X$ is not a cause of $Y$ if the distribution of $Y$ given $X$ is the same as the marginal distribution of $Y$; and this is equivalent to

*probabilistic independence* of $Y$ from $X$: $Y \perp\!\!\!\perp X$. But this is clearly unsatisfactory as it stands, since we could have dependence between $X$ and $Y$, $Y \not\perp\!\!\!\perp X$, with, at the same time, conditional independence given some other variable (or set of variables) $Z$: $Y \perp\!\!\!\perp X \mid Z$. If $Z$ can be regarded as delimiting the context in which we are considering the relationship between $X$ and $Y$, we might still regard $X$ and $Y$ as "causally unrelated". Thus probabilistic causality is based on *conditional* (in)dependence properties of probability distributions. However there remain obvious problems in simply equating the non-symmetrical relation of cause-and-effect with the symmetrical relation of probabilistic (in)dependence, and with clarifying what counts as an appropriate conditioning "context" variable $Z$, so that additional structure and assumptions (*e.g.* related to an assumed "causal order", possibly but not necessarily temporal) are required to complete the theory.

Most modern accounts locate probabilistic causality firmly within the graphical conception — so inheriting all the features and difficulties of that approach. It is *assumed* that there is a DAG $\mathcal{D}$, over a suitable collection $\mathcal{V}$ of variables, such that

(i). $\mathcal{D}$ can be interpreted as a *causal* DAG; and, in addition,

(ii). the joint probability distribution $P$ of the variables in $\mathcal{V}$ is *Markov* over $\mathcal{D}$, *i.e.* its probabilistic conditional independence (CI) properties are represented by the same DAG $\mathcal{D}$, according to the "$d$-separation" semantics described by Pearl [1986], Verma and Pearl [1990], Lauritzen et al. [1990].

In particular, from (ii), for any $V \in \mathcal{V}$, $V$ is independent of its non-descendants, $\mathrm{nd}(V)$, in $\mathcal{D}$, given its parents, $\mathrm{pa}(V)$, in $\mathcal{D}$. Given the further interpretation (i) of $\mathcal{D}$ as a causal DAG, this can be expressed as "$V$ is independent of its non-effects, given its direct causes in $\mathcal{V}$"— the so-called *causal Markov* assumption. Also, (ii) implies that, for any sets of variables $X$ and $Y$ in $\mathcal{D}$, $X \perp\!\!\!\perp Y \mid \mathrm{an}(X) \cap \mathrm{an}(Y)$ (where $\mathrm{an}(X)$ denotes the set of ancestors of $X$ in $\mathcal{D}$, including $X$ itself): again with $\mathcal{D}$ interpreted as causal, this can be read as saying "$X$ and $Y$ are conditionally independent, given their common causes in $\mathcal{V}$". In particular, marginal independence (where $X \perp\!\!\!\perp Y$ is represented in $\mathcal{D}$) holds if and only if $\mathrm{an}(X) \cap \mathrm{an}(Y) = \emptyset$, *i.e.* (using (i)) "$X$ and $Y$ have no common cause" (including each other) in $\mathcal{V}$; in the "if" direction, this has been termed the *weak causal Markov* assumption [Scheines and Spirtes 2008]. Many workers regard the causal and weak causal Markov assumptions as compelling—but this must depend on making the "right" choice for $\mathcal{V}$ (essentially, through appropriate delineation of the causal ambit.)

Note that this conception of causality involves, simultaneously, two very different ways of interpreting the DAG $\mathcal{D}$ (see Dawid [2009] for more on this). The $d$-separation semantics by means of which we relate $\mathcal{D}$ to conditional independence properties of the joint distribution $P$, while clearly defined, are somewhat subtle: in particular, the arrows in $\mathcal{D}$ are somewhat incidental "construction lines", that only play a small rôle in the semantics. But as soon as we also give $\mathcal{D}$ an interpretation as a "causal DAG" we are into a completely different way of interpreting it, where the arrows themselves are regarded as directly carrying causal meaning. Probabilistic causality can thus be thought of as the progeny of a shotgun wedding between two ill-matched parties.

### Causal discovery

The enterprise of *Causal Discovery* [Spirtes et al. 2000; Glymour and Cooper 1999; Neapolitan 2003] is grounded in this probabilistic-cum-graphical conception of causality. There are many variations, but all share the same basic philosophy. Essentially, one analyses observational data in an attempt to identify conditional independencies (possibly involving unobserved variables) in the distribution from which they arise. Some of these might be discarded as "accidental" (perhaps because they are inconsistent with an *a priori* causal order); those that remain might be represented by a DAG. The hope is that this discovered conditional independence DAG can also be interpreted as a causal DAG. When, as is often the case, there are several Markov equivalent DAG representations of the discovered CI relationships, which, moreover, cannot be causally distinguished on *a priori* grounds (*e.g.* in terms of an assumed causal order), this hope can not be fully realised; but if we can assume that one of these, at least, is a causal DAG, then at least an arrow common to all of them can be interpreted causally.

## 2.4 A spot of bother

Spirtes et al. [2000] and Pearl [2009], among others, have stressed the fundamental importance of distinguishing between the activities of *Seeing* and *Doing*. *Seeing* involves passive observation of a system in its natural state. *Doing*, on the other hand, relates to the behaviour of the system in a disturbed state brought about by external intervention. As a simple point of pure logic, there is no reason for there to be any relationship between these two types of behaviour of a system.

The probabilistic interpretation of causality relates solely to the *seeing* regime, whereas the agency account focuses entirely on what happens in *doing* regimes. As such these two interpretations inhabit totally unrelated universes. There are non-trivial foundational difficulties with the probabilistic (or other graphical) interpretations of causality (what exactly is a causal DAG? how will we know when we have got one?); on the other hand agency causality, while less obviously problematic and perhaps more naturally appealing, does not currently appear to offer a rich enough theory to be very useful. Even at a purely technical level, agency and probabilistic causality have very little in common. Probabilistic causality, through its close ties with conditional independence, has at its disposal the well-developed theoretical machinery of that concept, while the associated graphical structure allows for ready interpretation of concepts such as "causal pathway". Such considerations are however of marginal relevance to agency causality, which need not involve any probabilistic or graphical connexions.

From the point of view of a statistician, this almost total disconnect between the causal theories relating to the regimes of seeing and doing is particularly worrying. For one of the major purposes of "causal inference" is to draw conclusions, from purely observational "seeing" data on a system, about "doing": how would the system behave were we to intervene in it in certain ways? But not only is there no necessary logical connexion between the behaviours in the different regimes, the very concepts and representations by which we try to understand causality in the different regimes are worlds apart.

## 3 The Pearlian Synthesis

Building on ideas introduced by Spirtes et al. [2000], Pearl's approach to causality, as laid out for example in his book [Pearl 2009],[6] attempts to square this circle: it combines the two apparently incommensurable approaches of agency causality and probabilistic causality[7] in a way that tries to bring together the best features of both, while avoiding many of their individual problems and pitfalls.

Pearl considers a type of stochastic model, described by a DAG $\mathcal{D}$ over a collection $\mathcal{V}$ of variables, that can be simultaneously interpreted in terms of both agency and probabilistic causality. We could, if we wished, think of $\mathcal{V}$ as a "causal ambit", and $\mathcal{D}$ as a "causal DAG", but little is gained (or lost) by doing so, since the interpretations of any causal terms we may employ are provided internally by the model, rather than built on any pre-existing causal conceptions.

In its probabilistic interpretation, such a DAG $\mathcal{D}$ represents the conditional independence properties of the undisturbed system, which is supposed Markov with respect to $\mathcal{D}$. In its agency interpretation, the same DAG $\mathcal{D}$ is used to describe precisely how the system responds, probabilistically, to external interventions that set the values of (an arbitrary collection of) its variables. Specifically, such a disturbed probability distribution is supposed still Markov with respect to $\mathcal{D}$, and the conditional distribution of any variable $V$ in $\mathcal{V}$, given its parents in $\mathcal{D}$, is supposed the same in all regimes, seeing or doing (except of course those that directly set the value of $V$ itself, say at $v$, for which that distribution is replaced by the 1-point distribution at $v$). The "parent-child" conditional distributions thus constitute invariant "modular components" that (with the noted exception) can be transferred unchanged from one regime to another.

We term such a causal DAG model "Pearlian". Whether or not a certain DAG $\mathcal{D}$ indeed supplies a Pearlian DAG model for a given system can never be a purely syntactical question about its graphical structure, but is, rather, a semantic question about its relationship with the real world: do the various regimes actually have the probabilistic properties and relationships asserted? This may be true or false, but at least it is a meaningful question, and it is clear in principle how it can be addressed in purely empirical fashion: by observing and comparing the behaviours of the system under the various regimes.[8] A Pearlian DAG

[6] We in fact shall deal only with Pearl's earlier, fully stochastic, theory. More recently (see the second-half of Pearl [2009], starting with Chapter 7), he has moved to an interpretation of DAG models based on deterministic functional relationships, with stochasticity deriving solely from unobserved exogenous variables. That interpretation does however imply all the properties of the stochastic theory, and can be regarded as a specialisation of it. We shall not here be considering any features (such as the possibility of counterfactual analysis) dependent on the additional structure of Pearl's deterministic approach, since these only become relevant when analysing "causes of effects"—see Dawid [2000, 2002] for more on this.

[7] We have already remarked that probabilistic causality is itself the issue of an uneasy alliance between two quite different ways of interpreting graphs. Further miscegenation with the agency conception of causality looks like a eugenically risky endeavour!

[8] For this to be effective, the variables in $\mathcal{V}$ should have clearly-defined meanings and be observable in the real-world. Some Pearlian models incorporate unobservable latent variables without clearly identified external referents, in which case only the implications of such a model for the behaviour of observables can be put to empirical test.

model thus has the great virtue, all too rare in treatments of causality, of being totally clear and explicit about what is being said—allowing one to accord it, in a principled way, acceptance or rejection, as deemed appropriate, in any given application. And when a system can indeed be described by a Pearlian DAG, it is straightforward to learn (not merely qualitatively, but quantitatively too), from purely observational data, about the (probabilistic) effects of any interventions on variables in the system.

### 3.1 Justification

The falsifiability of the property of being a Pearlian DAG (unlike, for example, the somewhat ill-defined property of being a "causal DAG") is at once a great strength of the theory (especially for those with a penchant for Karl Popper's "falsificationist" Philosophy of Science), and something of an Achilles' heel. For all too often it will be impossible, for a variety of pragamatic, ethical or financial reasons, to conduct the experiments that would be needed to falsify the Pearlian assumptions. A lazy reaction might then simply be to assume that a DAG found, perhaps by "causal discovery", to represent observational conditional independencies, but without any interventions having been applied, is indeed Pearlian—and so also describes what would happen under interventions. While this may well be an interesting working hypothesis to guide further experimental investigations, it would be an illogical and dangerous point at which to conclude our studies. In particular, further experimental investigations could well result in rejection of our assumed Pearlian model.

Nevertheless, if forced to make a tentative judgment on the Pearlian nature, or otherwise, of a putative DAG model[9] of a system, there are a number of more or less reasonable, more or less intuitive, arguments that can be brought to bear. As a very simple example, we would immediately reject any putative "Pearlian DAG" in which an arrow goes backwards in time,[10] or otherwise conflicts with an accepted causal order. As another, if an "observational" regime itself involves an imposed physical randomisation to generate the value of some variable $X$, in a way that might possibly take account of variables $Z$ temporally prior to $X$, we might reasonably regard the conditional distribution of some later variable $Y$, given $X$ and $Z$, as a modular component, that would be the same in a regime that intervenes to set the value of $X$ as it is in the (observational) randomisation regime.[11] Such arguments can be further extended to "natural experiments", where it is Nature that imposed the external randomisation. This is the case for "Mendelian randomisation" [Didelez and Sheehan 2007], which capitalises on the random assortment of genes under Mendelian genetics. Other natural experiments rely on other causal assumptions about Nature: thus the "discontinuity design" [Trochim 1984] assumes that Nature supplies continuous dose-response cause-effect relationships. But all such justifications are, and must be, based on (what we think are) properties of the real world, and not solely on the internal structure of

[9] Assumed, for the sake of non-triviality, already to be a Markov model of its observational probabilistic properties.

[10] Assuming, as most would accept, that an intervention in a variable at some time can not affect any variable whose value is determined at an earlier time.

[11] See Dawid [2009] for an attempted argument for this, as well as caveats as to its general applicability.

the putative Pearlian DAG. In particular, they are founded on pre-existing ideas we have about causal and non-causal processes in the world, even though these ideas may remain unformalised and woolly: the important point is that we have enough, perhaps tacit, shared understanding of such processes to convince both ourselves and others that they can serve as external justification for a suggested Pearlian model. Unless we have sufficient justification of this kind, all the beautiful analysis (*e.g.* in Pearl [2009]) that develops the implications of a Pearlian model will be simply irrelevant. To echo Cartwright [1994, Chapter 2], "No causes in, no causes out".

## 4 Modularity, extended conditional independence and decision-theoretic causality

Although Pearlian causality as described above appears to be closely tied to graphical representation, this is really an irrelevance. We can strip it of its graphical clothing, laying bare its core ingredient: the property that certain conditional distributions[12] are the same across several different regimes. This *modular* conception provides us with yet another interpretation of causality. When, as here, the regimes considered encompass both observation (seeing) and intervention (doing), it has the great advantage over other theories of linking those disparate universes, thus supporting *causal inference*.

The modularity assumption can be conveniently expressed formally in the algebraic language of conditional independence, suitably interpreted [Dawid 1979, 2002, 2009], making no reference to graphs. Thus let $F$ be a "regime indicator", a non-stochastic parameter variable, whose value indicates the regime whose probabilistic properties are under consideration. If $X$ and $Y$ are stochastic variables, the "extended conditional independence" (ECI) property

(2) $\quad Y \perp\!\!\!\perp F \mid X$

can be interpreted as asserting that the conditional distribution of $Y$, for specified regime $F = f$ and given observed value $X = x$, depends only on $x$ and not further on the regime $f$ that is operating: in terms of densities we could write $p(y \mid f, x) = p(y \mid x)$. If $F$ had been a stochastic variable this would be entirely equivalent to stochastic conditional independence of $Y$ and $F$ given $X$; but it remains meaningful, with the above interpretation, even when $F$ is a non-stochastic regime indicator: Indeed, it asserts exactly the modular nature of the conditional distribution $p(y \mid x)$, as being the same across all the regimes indicated by values of $F$. Such modularity properties, when expressed in terms of ECI, can be formally manipulated—and, in those special cases where this is possible and appropriate, represented and manipulated graphically—in essentially the same fashion as for regular probabilistic conditional independence.

For applications of ECI to causal inference, we would typically want one or more of the regimes indicated by $F$ to represent the behaviour of the system when subjected to an intervention of a specified kind—thus linking up nicely with the agency interpretation; and one

[12]More generally, we could usefully identify features of the different regimes other than conditional distributions—for example, conditional expectations, or odds ratios—as modular components.

regime to describe the undisturbed system on which observations are made—thus allowing the possibility of "causal inference" and making links with probabilistic causality, but in a non-graphical setting. Modularity/ECI assumptions can now be introduced, as considered appropriate, and their implications extracted by algebraic or graphical manipulations, using the established theory of conditional independence. We emphasise that, although the notation and technical machinery of conditional independence is being used here, this is applied in a way that is very different from the approach of probabilistic causality: no assumptions need be made connecting causal relationships with ordinary probabilistic conditional independence.

Because it concerns the probabilistic behaviour of a system under interventions—a particular interpretation of agency causality—this general approach can be termed "decision-theoretic" causality. With the emphasis now on modularity, intuitive or graphically motivated causal terms such as "direct effect" or "causal pathway" are best dispensed with (and with them such assumptions as the causal Markov property). The decision-theoretic approach should not be regarded as providing a philosophical foundation for "causality", or even as a way of interpreting causal terms, but rather as very useful machinery for expressing and manipulating whatever modularity assertions one might regard as appropriate in a given problem.

### 4.1 Intervention DAGs

The assumptions that are implicit in a Pearlian model can be displayed very explicitly in the decision-theoretic framework, by associating a non-stochastic "intervention variable" $F_X$ with each "domain variable" $X \in \mathcal{V}$. The assumed ECI properties are conveniently displayed by means of a DAG, $\mathcal{D}^*$, which extends the Pearlian DAG $\mathcal{D}$ by adding extra nodes for these regime indicators, and extra arrows, from $F_X$ to $X$ for each $X \in \mathcal{V}$ [Spohn 1976; Spirtes et al. 2000; Pearl 2009; Dawid 2002; Dawid 2009]. If $\mathcal{X}$ is the set of values for $X$, then that for $F_X$ is $\mathcal{X} \cup \{\emptyset\}$: the intended interpretation is that $F_X = \emptyset$ (the "idle" regime) corresponds to the purely observational regime, while $F_X = x \in \mathcal{X}$ corresponds to "setting" $X$ at $x$.

To be precise, we specify the distribution of $X \in \mathcal{V}$ given its parents $(\mathrm{pa}(X), F_X)$ in $\mathcal{D}^*$ (where $\mathrm{pa}(X)$ denotes the "domain" parents of $X$, in $\mathcal{D}$) as follows. When $F_X = \emptyset$, this is the same as the observational conditional distribution of $X$, given $\mathrm{pa}(X)$; and when $F_X = x$ it is just a 1-point distribution on $x$, irrespective of the values of $\mathrm{pa}(X)$. The extended DAG $\mathcal{D}^*$, supplied with these parent-child specifications, is the *intervention DAG* representation of the problem.

With this construction, for any settings of all the regime indicators, some to idle and some to fixed values, the implied joint distribution of all the domain variables in that regime is exactly as required for the Pearlian DAG interpretation. But a valuable added bonus of the intervention DAG representation is that the Pearlian assumptions are explicitly represented. For example, the standard $d$-separation semantics applied to $\mathcal{D}^*$ allows us to read off the ECI property $X \perp\!\!\!\perp \{F_Y : Y \neq X\} \mid (\mathrm{pa}(X), F_X)$, which asserts the modular property of the conditional distribution of $X$ given $\mathrm{pa}(X)$: when $F_X = \emptyset$ (the only non-trivial case) the

conditional distribution of $X$ given $\mathrm{pa}(X)$ is the same, no matter how the other variables are set (or left idle).

### 4.2 More general causal models

It is implicit in the Pearlian conception that every variable in $\mathcal{V}$ should be manipulable (the causal Markov property then follows). But there is no real reason to require this. We can instead introduce intervention variables for just those variables that we genuinely wish to consider as "settable". The advantage of this is that fewer assumptions need be made and justified, but useful conclusions can often still be drawn.

EXAMPLE 2. **(Instrumental variable)**

Suppose we are interested in the "causal effect" of a binary exposure variable $X$ on some response $Y$. However we can not directly manipulate $X$. Moreover the observational relationship between $X$ and $Y$ may be distorted because of an unobserved "confounder" variable, $U$, associated with both $X$ and $Y$. In an attempt to evade this difficulty, we also measure an "instrumental variable" $Z$.

To express our interest in the *causal* effect of $X$ on $Y$, we introduce an intervention variable $F_X$ associated with $X$, defined and interpreted exactly as in § 4.1 above. The aim of our causal inference is to make some kind of comparison between the distributions of the response $Y$ in the interventional regimes, $F_X = 0$ and $F_X = 1$, corresponding to manipulating the value of $X$. The available data, however, are values of $(X, Y, Z)$ generated under the observational regime, $F_X = \emptyset$. We must make some assumptions if we are to be able to use features of that observational joint distribution to address our causal question, and clearly these must involve some kind of transference of information across regimes.

A useful (when valid!) set of assumptions about the relationships between all the variables in the problem is embodied in the following set of ECI properties (the "core conditions"[13] for basing causal inferences on an instrumental variable):

$$(U, Z) \quad \perp\!\!\!\perp \quad F_X \tag{3}$$

$$U \quad \perp\!\!\!\perp \quad Z \mid F_X \tag{4}$$

$$Y \quad \perp\!\!\!\perp \quad F_X \mid (X, U) \tag{5}$$

$$Y \quad \perp\!\!\!\perp \quad Z \mid (X, U; F_X) \tag{6}$$

$$X \quad \not\perp\!\!\!\perp \quad Z \mid F_X = \emptyset \tag{7}$$

Property (3) is to be interpreted as saying that the joint distribution of $(U, Z)$ is independent of the regime $F_X$: *i.e.*, it is the same in all three regimes. That is to say, it is entirely unaffected by whether, and if so how, we intervene to set the value of $X$. The identity of this joint distribution across the two interventional regimes, $F_X = 0$ and $F_X = 1$, can be interpreted as expressing a causal property: manipulating $X$ has no (probabilistic) effect

[13] In addition to these core conditions, precise identification of a causal effect by means of an instrumental variable requires further modelling assumptions, such as linear regressions [Didelez and Sheehan 2007].

on the pair of variables $(U, Z)$. Moreover, since this common joint distribution is also supposed the same in the idle regime, $F_X = \emptyset$, we could in principle use observational data to estimate it—thus opening up the possibility of causal inference.

Property (4) asserts that, in their (common) joint distribution in any regime, $U$ and $Z$ are independent (this however is a purely probabilistic, not a causal, property).

Property (5) says that the conditional distribution of $Y$ given $(X, U)$ is the same in both interventional regimes, as well as in the observational regime, and can thus be considered as a modular component, fully transferable between the three regimes—again, I regard this as expressing a causal property.

Property (6) asserts that this common conditional distribution is unaffected by further conditioning on $Z$ (not in itself a causal property).

Finally, property (7) requires that $Z$ be genuinely associated with $X$ in the observational regime.

Of course, these ECI properties should not simply be assumed without some attempt at justification: for example, Mendelian randomisation attempts this in the case that $Z$ is an inherited gene. But because we have no need to consider interventions at any node other than $X$, less by way of justification is required than if we were to do so.

Once expressed in terms of ECI, these core conditions can be manipulated algebraically using the general theory of conditional independence [Dawid 1979]. Depending on what further modelling assumptions are made, it may then be possible to identify, or to bound, the desired causal effect in terms of properties of the observational joint distribution of $(X, Y, Z)$ [Dawid 2007b, Chapter 11].

In this particular case, although the required ECI conditions are expressed without reference to any graphical representation, it is possible (though not obligatory!) to give them one. This is shown in Figure 1. Properties (3)–(6) can be read off this DAG directly using the standard $d$-separation semantics. (Property (7) is only represented under a further assumption that the graphical representation is faithful.) We term such a DAG an *augmented DAG*: it differs from a Pearlian DAG in that some, but not necessarily all, variables have associated intervention indicators.

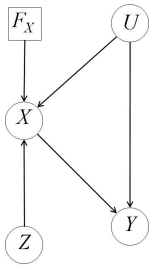



Figure 1. Instrumental variable: Augmented DAG representation

Just as for regular CI, it is possible for a collection of ECI properties, constituting a

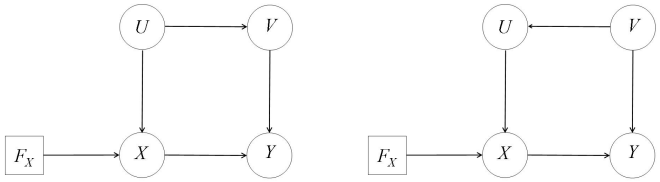



Figure 2. Two Markov-equivalent augmented DAGs

decision-theoretic causal model, to have no (augmented) DAG representation, or more than one. This latter is the case for Figure 2, where the direction of the arrow between $U$ and $V$ is not determined. This emphasises that, even when we do have an augmented DAG representation, we can not necessarily interpret the direction of an arrow in it as directly related to the direction of causality. Even in Figure 1 (and in spite of the natural connotation of the term "instrument"), the arrow pointing from $Z$ to $X$ is not be interpreted as necessarily causal, since the dependence between $Z$ and $X$ could be due to a "common cause" $U^*$ without affecting the ECI properties (3)–(6) [Dawid 2009], and Figure 1 is merely a graphical representation of these properties, based on $d$-separation semantics. In particular, one should be cautious of using an augmented DAG, which is nothing but a way of representing certain ECI statements, to introduce graphically motivated concepts such as "causal pathway". The general decision-theoretic description of causality *via* modularity, expressed in terms of ECI properties, where there is no requirement that the assumptions be representable by means of an augmented DAG at all, allows us to evade some of the restrictions of graphical causality, while still retaining a useful "agency-cum-probabilistic" causal theory.

The concept of an "interventional regime" can be made much more general, and in particular we need not require that it have the properties assumed above for an intervention variable associated with a domain variable. We could, for example, incorporate "fat hand" interventions that do not totally succeed in their aim of setting a variable to a fixed value, or interventions (such as kicking the system) that simultaneously affect several domain variables [Duvenaud et al. 2009]. So long as we understand what such regimes refer to in the real world, and can make and justify assumptions of modularity of appropriate conditional distributions as we move across regimes, we can apply the decision-theoretic ECI machinery. And at this very general level we can even apply a variant of "causal discovery" algorithms—so long as we can make observations under all the regimes considered.[14] For example, if we can observe $(X, Y)$ under the different regimes described by $F$, we can readily investigate the validity of the ECI property $X \perp\!\!\!\perp F \mid Y$ using standard tests (*e.g.*

[14]Or we might make parametric modelling assumptions about the relationships across regimes, to fill in for regimes we are not able to observe. This would be required for example when want to consider the effect of setting the value of a continuous "dose" variable. At this very general level we can even dispense entirely with the assumption of modular conditional distributions [Duvenaud et al. 2009].

the $\chi^2$-test) for conditional independence. Such discovered ECI properties (whether or not they can be expressed graphically) can then be used to model the "causal structure" of the problem.

## 5 Conclusion

Over many years, Judea Pearl's original and insightful approach to understanding uncertainty and causality have had an enormous influence on these fields. They have certainly had a major influence on my own research directions: I have often—as evidenced by this paper—found myself following in his footsteps, picking up a few crumbs here and there for further digestion.

Pearl's ideas do not however exist in a vacuum, and I believe it is valuable both to relate them to their precursors and to assess the ways in which they may develop. In attempting this task I fully acknowledge the leadership of a peerless researcher, whom I feel honoured to count as a friend.

## References

Campaner, R. and M. C. Galavotti (2007). Plurality in causality. In P. K. Machamer and G. Wolters (Eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics*, pp. 178–199. Pittsburgh: University of Pittsburgh Press.

Cartwright, N. (1994). *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B 41*, 1–31.

Dawid, A. P. (2000). Causal inference without counterfactuals (with Discussion). *Journal of the American Statistical Association 95*, 407–448.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 70*, 161–189. Corrigenda, *ibid*., 437.

Dawid, A. P. (2007a). Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In F. Russo and J. Williamson (Eds.), *Causality and Probability in the Sciences*, Volume 5 of *Texts in Philosophy*, pp. 503–32. London: College Publications.

Dawid, A. P. (2007b). Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London. <http://www.ucl.ac.uk/Stats/research/reports/psfiles/rr279.pdf>

Dawid, A. P. (2010). Beware of the DAG! *Journal of Machine Learning Research*. To appear.

Dawid, A. P. (2010). The rôle of scientific and statistical evidence in assessing causality. In R. Goldberg, J. Paterson, and G. Gordon (Eds.), *Perspectives on Causation*, Oxford. Hart Publishing. To appear.

Didelez, V. and N. A. Sheehan (2007). Mendelian randomisation as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research 16*, 309–330.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Duvenaud, D., D. Eaton, K. Murphy, and M. Schmidt (2010). Causal learning without DAGs. *Journal of Machine Learning Research*. To appear.

Galavotti, M. C. (2008). Causal pluralism and context. In M. C. Galavotti, R. Scazzieri, and P. Suppes (Eds.), *Reasoning, Rationality and Probability*, Chapter 11, pp. 233–252. Chicago: The University of Chicago Press.

Glymour, C. and G. F. Cooper (Eds.) (1999). *Computation, Causation and Discovery*. Menlo Park, CA: AAAI Press.

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy XCVII*, 198–222.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science 56*, 843–887.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

Hernán, M. A. and J. M. Robins (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology 17*, 360–372.

Hitchcock, C. (2007). How to be a causal pluralist. In P. K. Machamer and G. Wolters (Eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics*, pp. 200–221. Pittsburgh: University of Pittsburgh Press.

Lauritzen, S. L., A. P. Dawid, B. N. Larsen, and H.-G. Leimer (1990). Independence properties of directed Markov fields. *Networks 20*, 491–505.

Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Upper Saddle River, New Jersey: Prentice Hall.

Pearl, J. (1986). A constraint–propagation approach to probabilistic reasoning. In L. N. Kanal and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence*, Amsterdam, pp. 357–370. North-Holland.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (Second ed.). Cambridge: Cambridge University Press.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science 42*, 157–176.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of Los Angeles Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics 6*, 34–68.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Scheines, R. and P. Spirtes (2008). Causal structure search: Philosophical foundations and future problems. Paper presented at NIPS 2008 Workshop "Causality: Objectives and Assessment", Whistler, Canada.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction and Search* (Second ed.). New York: Springer-Verlag.

Spirtes, P. and R. Scheines (2004). Causal inference of ambiguous manipulations. *Philosophy of Science 71*, 833–845.

Spohn, W. (1976). *Grundlagen der Entscheidungstheorie*. Ph.D. thesis, University of Munich. (Published: Kronberg/Ts.: Scriptor, 1978).

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In M. C. Galavotti, P. Suppes, and D. Costantini (Eds.), *Stochastic Dependence and Causality*, Chapter 9, pp. 157–172. Chicago: University of Chicago Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. SAGE Publications.

Verma, T. and J. Pearl (1990). Causal networks: Semantics and expressiveness. In R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 4*, Amsterdam, pp. 69–76. North-Holland.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

19

# Effect Heterogeneity and Bias in Main-Effects-Only Regression Models

FELIX ELWERT AND CHRISTOPHER WINSHIP

## 1 Introduction

The overwhelming majority of OLS regression models estimated in the social sciences, and in sociology in particular, enter all independent variables as main effects. Few regression models contain many, if any, interaction terms. Most social scientists would probably agree that the assumption of constant effects that is embedded in main-effects-only regression models is theoretically implausible. Instead, they would maintain that regression effects are historically and contextually contingent; that effects vary across individuals, between groups, over time, and across space. In other words, social scientists doubt constant effects and believe in effect heterogeneity.

But why, if social scientists believe in effect heterogeneity, are they willing to substantively interpret main-effects-only regression models? The answer—not that it's been discussed explicitly—lies in the implicit assumption that the main-effects coefficients in linear regression represent straightforward averages of heterogeneous individual-level causal effects.

The belief in the averaging property of linear regression has previously been challenged. Angrist [1998] investigated OLS regression models that were correctly specified in all conventional respects except that effect heterogeneity in the main treatment of interest remained unmodeled. Angrist showed that the regression coefficient for this treatment variable gives a rather peculiar type of average—a conditional variance weighted average of the heterogeneous individual-level treatment effects in the sample. If the weights differ greatly across sample members, the coefficient on the treatment variable in an otherwise well-specified model may differ considerably from the arithmetic mean of the individual-level effects among sample members.

In this paper, we raise a new concern about main-effects-only regression models. Instead of considering models in which heterogeneity remains unmodeled in only one effect, we consider standard linear path models in which unmodeled heterogeneity is potentially pervasive.

Using simple examples, we show that unmodeled effect heterogeneity in more than one structural parameter may mask confounding and selection bias, and thus lead to biased estimates. In our simulations, this heterogeneity is indexed by latent (unobserved) group membership. We believe that this setup represents a fairly realistic scenario—one in which the analyst has no choice but to resort to a main-effects-only regression model because she cannot include the desired interaction terms since group-membership is un-

observed. Drawing on Judea Pearl's theory of directed acyclic graphs (DAG) [1995, 2009] and VanderWeele and Robins [2007], we then show that the specific biases we report can be predicted from an analysis of the appropriate DAG. This paper is intended as a serious warning to applied regression modelers to beware of unmodeled effect heterogeneity, as it may lead to gross misinterpretation of conventional path models.

We start with a brief discussion of conventional attitudes toward effect heterogeneity in the social sciences and in sociology in particular, formalize the notion of effect heterogeneity, and briefly review results of related work. In the core sections of the paper, we use simulations to demonstrate the failure of main-effects-only regression models to recover average causal effects in certain very basic three-variable path models where unmodeled effect heterogeneity is present in more than one structural parameter. Using DAGs, we explain which constellations of unmodeled effect heterogeneity will bias conventional regression estimates. We conclude with a summary of findings.

## 2 A Presumed Averaging Property of Main-Effects-Only Regression

### 2.1 Social Science Practice

The great majority of empirical work in the social sciences relies on the assumption of constant coefficients to estimate OLS regression models that contain nothing but main effect terms for all variables considered.[1] Of course, most researchers do not believe that real-life social processes follow the constant-coefficient ideal of conventional regression. For example, they aver that the effect of marital conflict on children's self-esteem is larger for boys than for girls [Amato and Booth 1997]; or that the death of a spouse increases mortality more for white widows than for African American widows [Elwert and Christakis 2006]. When pressed, social scientists would probably agree that the causal effect of almost any treatment on almost any outcome likely varies from group to group, and from person to person.

But if researchers are such firm believers in effect heterogeneity, why is the constant-coefficients regression model so firmly entrenched in empirical practice? The answer lies in the widespread belief that the coefficients of linear regression models estimate averages of heterogeneous parameters—average causal effects—representing the average of the individual-level causal effects across sample members. This (presumed) averaging property of standard regression models is important for empirical practice for at least three reasons. First, sample sizes in the social sciences are often too small to investigate effect heterogeneity by including interaction terms between the treatment and more than a few common effect modifiers (such as sex, race, education, income, or place of residence); second, the variables needed to explicitly model heterogeneity may well not have been measured; third, and most importantly, the complete list of effect modifiers along which the causal effect of treatment on the outcome varies is typically unknown (indeed, unknowable) to the analyst in any specific application. Analysts thus rely on faith that

[1] Whether a model requires an interaction depends on the functional form of the dependent and/or independent variables. For example, a model with no interactions in which the independent variables are entered in log form, would require a whole series of interactions in order to approximate this function if the independent variables where entered in nonlog form.

their failure to anticipate and incorporate all dimensions of effect heterogeneity into regression analysis simply shifts the interpretation of regression coefficients from individual-level causal effects to average causal effects, without imperiling the causal nature of the estimate.

### 2.2 Defining Effect Heterogeneity

We start by developing our analysis of the consequences of causal heterogeneity within the counterfactual (potential outcomes) model. For a continuous treatment $T \in (-\infty,\infty)$, let $T = t$ denote some specific treatment value and $T = 0$ the control condition. $Y(t)_i$ is the potential outcome of individual i for treatment $T = t$, and $Y(0)_i$ is the potential outcome of individual i for the control condition. For a particular individual, generally only one value of $Y(t)_i$ will be observed. The *individual-level causal effect* (ICE) of treatment level $T = t$ compared to $T = 0$ is then defined as: $\delta_{i,t} = Y(t)_i - Y(0)_i$ (or $\delta_i$, for short, if T is binary).

Since $\delta_{i,t}$ is generally not directly estimable, researchers typically attempt estimating the *average causal effect* (ACE) for some sample or population:

$$\bar{\delta}_t = \sum_{i=1}^{N} \delta_{i,t} / N$$

We say that the effect of treatment T is *heterogeneous* if: $\delta_{i,t} \neq \bar{\delta}_t$ for at least one i.

In other words, effect heterogeneity exists if the causal effect of the treatment differs across individuals. The basic question of this paper is whether a regression estimate for the causal effect of the treatment can be interpreted as an average causal effect if effect heterogeneity is present.

### 2.3 Regression Estimates as Conditional Variance Weighted Average Causal Effects

The ability of regression to recover average causal effects under effect heterogeneity has previously been challenged by Angrist [1998].[2] Here, we briefly sketch the main result. For a binary treatment, T=0,1, Angrist assumed a model where treatment was ignorable given covariates X and the effect of treatment varied across strata defined by the values of X. He then analyzed the performance of an OLS regression model that properly controlled for confounding in X but was misspecified to include only a main effect term for T and no interactions between T and X. Angrist showed that the regression estimate for the main effect of treatment can be expressed as a weighted average of stratum-specific treatment effects, albeit one that is difficult to interpret. For each stratum defined by fixed values of X, the numerator of the OLS estimator has the form $\delta_x W_x P(X=x)$,[3] where $\delta_x$ is the stratum-specific causal effect and $P(X=x)$ is the relative size of the stratum in the sample. The weight, $W_x$, is a function of the propensity score, $P_x=P(T=1 \mid X)$, associated with the stratum, $W_x = P_x (1- P_x)$, which equals the stratum-specific variance of treatment. This variance, and hence the weight, is largest if $P_x=.5$ and smaller as $P_x$ goes to 0 or 1.

[2]This presentation follows Angrist [1998] and Angrist and Pischke [2009].

[3]The denominator of the OLS estimator is just a normalizing constant that does not aid intuition.

If the treatment effect is constant across strata, these weights make good sense. OLS gives the minimum variance linear unbiased estimator of the model parameters under homoscedasticity assuming correct specification of the model. Thus in a model without interactions between treatment and covariates X the OLS estimator gives the most weight to strata with the smallest variance for the estimated within-stratum treatment effect, which, not considering the size of the strata, are those strata with the largest treatment variance, i.e. with the $P_x$ that are closest to .5. However, if effects are heterogeneous across strata, this weighting scheme makes little substantive sense: in order to compute the average causal effect, $\bar{\delta}$, as defined above, we would want to give the same weight to every individual in the sample. As a variance-weighted estimator, however, regression estimates under conditions of unmodeled effect heterogeneity do not give the same weight to every individual in the sample and thus do not converge to the (unweighted) average treatment effect.

## 3 Path Models with Pervasive Effect Heterogeneity

Whereas Angrist analyzed a misspecified regression equation that incorrectly assumed no treatment-covariate interaction for a *single* treatment variable, we investigate the ability of a main-effects-only regression model to recover unbiased average causal effects in simple path models with unmodeled effect heterogeneity across *multiple* parameters.

*Setup:* To illustrate how misleading the belief in the averaging power of the constant-coefficient model can be in practice, we present simulations of basic linear path models, shown in summary in Figure 1 (where we have repressed the usual uncorrelated error terms).

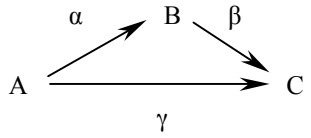



**Figure 1.** A simple linear path model

To introduce effect heterogeneity, let G = 0, 1 index membership in a latent group and permit the possibility that the three structural parameters α, β, and γ vary across (but not within) levels of G. The above path model can then be represented by two linear equations: $B = A\alpha_G + \varepsilon_B$ and $C = A\gamma_G + B\beta_G + \varepsilon_C$. In our simulations, we assume that A~N(0,1) and $\varepsilon_B$, and $\varepsilon_C$ are iid N(0,1), and hence all variables are normally distributed. From these equations, we next simulate populations of N=100,000 observations, with P(G=1) = P(G=0) = 1/2. We start with a population in which all three parameters are constant across the two subgroups defined by G, and then systematically introduce effect heterogeneity by successively permitting the structural parameters to vary by group, yielding one population for each of the $2^3 = 8$ possible combinations of constant/varying parameters. To fix ideas, we choose the group-specific parameter values shown in Table

1. For simulations in which one or more parameters do not vary by group, we set the constant parameter(s) to the average of the group specific parameters, e.g. $\alpha = (\alpha_0 + \alpha_1)/2$.

**Table 1:** Group-specific structural parameters for simulations

| | $\alpha_G$ | $\beta_G$ | $\gamma_G$ |
|---|---|---|---|
| Group: | | | |
| G=0 | 0.4 | 0.5 | 0.6 |
| G=1 | 1.2 | 2.5 | 1.4 |
| Average | 0.8 | 1.5 | 1.0 |

Finally, we estimate a conventional linear regression model for the effects of A and B on C using the conventional default specification, in which all variables enter as main effects only, $C = A\gamma + B\beta + \varepsilon$. (Note that G is latent and therefore cannot be included in the model.) The parameter, $\gamma$ refers to the direct effect of A on C holding B constant, and $\beta$ refers to the total effect of B on C.[4] In much sociological and social science research, this main-effects regression model is intended to recover average structural (causal) effects, and is commonly believed to be well suited for the purpose.

*Results:* Table 2 shows the regression estimates for the main effect parameters across the eight scenarios of effect heterogeneity. We see that the main effects regression model correctly recovers the desired (average) parameters, $\gamma$=1 and $\beta$=1.5 if none of the parameters vary across groups (column 1), or if only one of the three parameters varies (columns 2-4).

Other constellations of effect heterogeneity, however, produce biased estimates. If $\alpha_G$ and $\beta_G$ (column 5); or $\alpha_G$ and $\gamma_G$ (column 6); or $\alpha_G$, $\beta_G$, and $\gamma_G$ (column 8) vary across groups, the main-effects-only regression model fails to recover the true (average) parameter values known to underlie the simulations. For our specific parameter values, the estimated (average) effect of B on C in these troubled scenarios is always too high, and the estimated average direct effect of A on C is either too high or too low. Indeed, if we set $\gamma$=0 but let $\alpha_G$ and $\beta_G$ vary across groups, the estimate for $\gamma$ in the main-effects-only regression model would suggest the presence of a direct effect of A on C even though it is known by design that no such direct effect exists (not shown).

Failure of the regression model to recover the known path parameters is not merely a function of the number of paths that vary. Although none of the scenarios in which fewer than two parameters vary yield incorrect estimates, and the scenario in which all three parameters vary is clearly biased, results differ for the three scenarios in which exactly two parameters vary. In two of these scenarios (columns 5 and 6), regression fails to recover the desired (average) parameters, while regression does recover the correct average parameters in the third scenario (column 7).

[4]The notion of direct and indirect effects is receiving deserved scrutiny in important recent work by Robins and Greenland [1992]; Pearl [2001]; Robins [2003]; Frangakis and Rubin [2002]; Sobel [2008]; and VanderWeele [2008].

**Table 2:** OLS regression estimates for the main effects of A and B on C across eight different combinations of effect heterogeneity in α, β,and/or γ

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|---|---|---|---|
| Heterogeneity in: - | | α | β | γ | α, β | α, γ | β, γ | α, β, γ |
| Group: | G0 G1 | G0 G1 | G0 G1 | G0 G1 | G0 G1 | G0 G1 | G0 G1 | G0 G1 |
| α | 0.8 | 0.4 1.2 | 0.8 | 0.8 | 0.4 1.2 | 0.4 1.2 | 0.8 | 0.4 1.2 |
| β | 1.5 | 1.5 | 0.5 2.5 | 1.5 | 0.5 2.5 | 1.5 | 0.5 2.5 | 0.5 2.5 |
| γ | 1.0 | 1.0 | 1.0 | 0.6 1.4 | 1.0 | 0.6 1.4 | 0.6 1.4 | 0.6 1.4 |
| Pooled OLS estimate: | | | | | | | | |
| β | 1.50 | 1.50 | 1.50 | 1.50 | **1.77** | **1.64** | 1.50 | **1.91** |
| γ | 1.00 | 1.00 | 1.00 | 1.00 | **1.17** | **0.89** | 1.00 | **1.07** |

*Note:* Bold estimates are biased for the true (average) parameters. Results from independent simulations of N=100,000 for each scenario using (group-specific) parameters listed above. See text for details.

In sum, the naïve main-effects-only linear regression model recovers the correct (average) parameter values only under certain conditions of limited effect heterogeneity, and it fails to recover the true average effects in certain other scenarios, including the scenario we consider most plausible in the majority of sociological applications, i.e., where all three parameters vary across groups. If group membership is latent—because group membership is unknown to or unmeasured by the analyst— and thus unmodeled, linear regression generally will fail to recover the true average effects.

## 4 DAGs to the Rescue

These results spell trouble for empirical practice in sociology. Judea Pearl's work on causality and directed acyclic graphs (DAGs) [1995, 2009] offers an elegant and powerful approach to understanding the problem. Focusing on the appropriate DAGs conveys the critical insight for the present discussion that effect heterogeneity, rather than being a nuisance that is easily averaged away, encodes structural information that analysts ignore at their peril.

Pearl's DAGs are nonparametric path models that encode causal dependence between variables: an arrow between two variables indicates that the second variable is causally dependent on the first (for detailed formal expositions of DAGs, see Pearl [1995, 2009]; for less technical introductions see Robins [2001]; Greenland, Pearl and Robins [1999] in epidemiology, and Morgan and Winship [2007] in sociology). For example, the DAG in Figure 2 indicates that Z is a function of X and Y, $Z= f(X,Y,\varepsilon_Z)$, where $\varepsilon_Z$ is an unobserved error term independent of (X,Y).

In a non-parametric DAG—as opposed to a conventional social science path model—the term f( ) can be any function. Thus, the DAG in Figure 2 is consistent with a linear structural equation in which X only modifies (i.e. introduces heterogeneity into) the effect

of Y on Z, $Z=Y\xi + YX\psi + \varepsilon_Z$.[5] In the language of VanderWeele and Robins [2007], who provide the most extensive treatment of effect heterogeneity using DAGs to date, one may call X a "direct effect modifier" of the effect of Y on Z. The point is that a variable that modifies the effect of Y on Z is causally associated with Z, as represented by the arrow from X to Z.

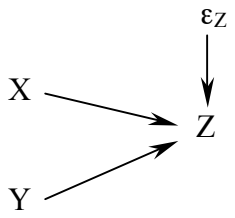



**Figure 2.** DAG illustrating direct effect modification of the effect of Y on Z in X

Returning to our simulation, one realizes that the social science path model of Figure 1, although a useful tool for informally illustrating the data generation process, does not, generally, provide a sufficiently rigorous description of the causal structure underlying the simulations. Figure 1, although truthfully representing the separate data generating mechanism for each group and each individual in the simulated population, is not the correct DAG for the pooled population containing groups G = 0 and G = 1 for all of the heterogeneity scenarios considered above. Specifically, in order to turn the informal social science path model of Figure 1 into a DAG, one would have to integrate the source of heterogeneity, G, into the picture. How this is to be done depends on the structure of heterogeneity. If only $\beta_G$ (the effect of B on C) and/or $\gamma_G$ (the direct effect of A on C holding B constant) varied with G, then one would add an arrow from G into C. If $\alpha_G$ (the effect of A on B) varied with G, then one would add an arrow from G into B. The DAG in Figure 3 thus represents those scenarios in which $\alpha_G$ as well as either $\beta_G$ or $\gamma_G$, or both, vary with G (columns 5, 6, and 8). Interpreted in terms of a linear path model, this DAG is consistent with the following two structural equations: $B = A\alpha_0 + AG\alpha_1 + \varepsilon_{\mathbf{B}}$ and $C = A\gamma_0 + AG\gamma_1 + B\beta_0 + BG\beta_1 + \varepsilon_C$ (where the iid errors, ε, have been omitted from the DAG and are assumed to be uncorrelated).[6]

In our analysis, mimicking the reality of limited observational data with weak substantive theory, we have assumed that A, B, and C are observed, but that G is not observed. It is immediately apparent that the presence of G in Figure 3 means that, first, G is a confounder for the effect of B on C; and, second, that B is a "collider" [Pearl 2009] on

[5]It is also consistent with an equation that adds a main effect of X. For the purposes of this paper it does not matter whether the main effect is present.

[6]By construction of the example, we assume that A is randomized and thus marginally independent of G. Note, however, that even though G is mean independent of B and C (no main effect of G on either B or C), G is not marginally independent of B or C because $var(B|G=1)\neq var(B|G=0)$ and $var(C|G=1)\neq var(C|G=0)$, which explains the arrows from G into B and C. Adding main effects of G on B and C would not change the arguments presented here.

the path from A to C via B and G. Together, these two facts explain the failure of the main-effects-only regression model to recover the true parameters in panels 5, 6, and 8: First, in order to recover the effect of B on C, β, one would need to condition on the confounders A and G. But G is latent so it cannot be conditioned on. Second, conditioning on the collider B in the regression opens a "backdoor path" from A to C via B and G (when G is not conditioned on), i.e. it induces a non-causal association between A and C, creating selection bias in the estimate for the direct effect of A on C, γ [Pearl 1995, 2009; Hernán et al 2004]. Hence, both coefficients in the main-effects-only regression model will be biased for the true (average) parameters.

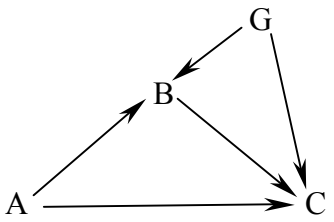



**Figure 3.** DAG consistent with effect modification of the effects of A on B, and B on C and/or A on C, in G

By contrast, if G modifies neither β nor γ, then the DAG would not contain an arrow from G into C; and if G does not modify α then the DAG would not contain an arrow from G into B. Either way, if either one (or both) of the arrows emanating from G are missing, then G is not a confounder for the effect of B on C, and conditioning on B will not induce selection bias by opening a backdoor path from A to C. Only then would the main effects regression model be unbiased and recover the true (average) parameters, as seen in panels 1-4 and 7.

In sum, Pearl's DAGs neatly display the structural information encoded in effect heterogeneity [VanderWeele and Robins 2007]. Consequently, Pearl's DAGs immediately draw attention to problems of confounding and selection bias that can occur when more than one effect in a causal system varies across sample members. Analyzing the appropriate DAG, the failure of main-effects-only regression models to recover average structural parameters in certain constellations of effect heterogeneity becomes predictable.

## 5 Conclusion

This paper considered a conventional structural model of a kind commonly used in the social sciences and explored its performance under various basic scenarios of effect heterogeneity. Simulations show that the standard social science strategy of dealing with effect heterogeneity—by ignoring it—is prone to failure. In certain situations, the main-effects-only regression model will recover the desired quantities, but in others it will not. We believe that effect heterogeneity in all arrows of a path model is plausible in many, if not most, substantive applications. Since the sources of heterogeneity are often not theorized, known, or measured, social scientists continue routinely to estimate main-effects-

only regression models in hopes of recovering average causal effects. Our examples demonstrate that the belief in the averaging powers of main-effects-only regression models may be misplaced if heterogeneity is pervasive, as estimates can be mildly or wildly off the mark. Judea Pearl's DAGs provide a straightforward explanation for these difficulties—DAGs remind analysts that effect heterogeneity may encode structural information about confounding and selection bias that requires consideration when designing statistical strategies for recovering the desired average causal effects.

**Acknowledgments:** We thank Jamie Robins for detailed comments on a draft version of this paper, and Michael Sobel, Stephen Morgan, Hyun Sik Kim, and Elizabeth Wrigley-Field for advice. Genevieve Butler provided editorial assistance.

## References

Amato, Paul R., and Alan Booth. (1997). *A Generation at Risk: Growing Up in an Era of Family Upheaval*. Cambridge, MA: Harvard University Press.

Angrist, Joshua D. (1998). "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica 66*: 249-88.

Angrist, Joshua D. and Jörn-Steffen Pischke. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Elwert, Felix, and Nicholas A. Christakis. (2006). "Widowhood and Race." *American Sociological Review 71*: 16-41.

Frangakis, Constantine E., and Donald B. Rubin. (2002). "Principal Stratification in Causal Inference." *Biometrics 58*: 21–29.

Greenland, Sander, Judea Pearl, and James M. Robbins. (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology 10*: 37-48.

Hernán, Miguel A., Sonia Hernández-Diaz, and James M. Robins. (2004). "A Structural Approach to Section Bias." *Epidemiology 155* (2): 174-184.

Morgan, Stephen L. and Christopher Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles of Social Research.* Cambridge: Cambridge University Press.

Pearl, Judea. (1995). "Causal Diagrams for Empirical Research." *Biometrika 82* (4): 669-710.

Pearl, Judea. (2001). "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 411-420.

Pearl, Judea. (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. Cambridge: Cambridge University Press.

Robins, James M. (2001). "Data, Design, and Background Knowledge in Etiologic Inference," *Epidemiology 11* (3): 313-320.

Robins, James M. (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In: *Highly Structured Stochastic Systems*, P. Green, N. Hjort and S. Richardson, Eds. Oxford: Oxford University Press.

Robins, James M, and Sander Greenland. (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology 3*:143-155.

Sobel, Michael. (2008). "Identification of Causal Parameters in Randomized Studies with Mediating Variables," *Journal of Educational and Behavioral Statistics 33* (2): 230-251.

VanderWeele, Tyler J. (2008). "Simple Relations Between Principal Stratification and Direct and Indirect Effects." *Statistics and Probability Letters 78*: 2957-2962.

VanderWeele, Tyler J. and James M. Robins. (2007). "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology 18* (5): 561-568.

# 20

# Causal and Probabilistic Reasoning in P-log

Michael Gelfond and Nelson Rushton

## 1 Introduction

In this paper we give an overview of the knowledge representation (KR) language P-log [Baral, Gelfond, and Rushton 2009] whose design was greatly influenced by work of Judea Pearl. We introduce the syntax and semantics of P-log, give a number of examples of its use for knowledge representation, and discuss the role Pearl's ideas played in the design of the language. Most of the technical material presented in the paper is not new. There are however two novel technical contributions which could be of interest. First we expand P-log semantics to allow domains with infinite Herbrand bases. This allows us to represent infinite sequences of random variables and (indirectly) continuous random variables. Second we generalize the logical base of P-log which improves the degree of elaboration tolerance of the language.

The goal of the P-log designers was to create a KR-language allowing natural and elaboration tolerant representation of commonsense knowledge involving logic and probabilities. The logical framework of P-log is Answer Set Prolog (ASP) — a language for knowledge representation and reasoning based on the answer set semantics (*aka* stable model semantics) of logic programs [Gelfond and Lifschitz 1988; Gelfond and Lifschitz 1991]. ASP has roots in declarative programing, the syntax and semantics of standard Prolog, disjunctive databases, and non-monotonic logic. The semantics of ASP captures the notion of possible beliefs of a reasoner who adheres to the *rationality principle* which says that "One shall not believe anything one is not forced to believe". The entailment relation of ASP is non-monotonic[1], which facilitates a high degree of elaboration tolerance in ASP theories. ASP allows natural representation of defaults and their exceptions, causal relations (including effects of actions), agents' intentions and obligations, and other constructs of natural language. ASP has a number of efficient reasoning systems, a well developed mathematical theory, and a well tested methodology of representing and using knowledge for computational tasks (see, for instance, [Baral 2003]). This, together with the fact that some of the designers of P-log came from the ASP community made the choice of a logical foundation for P-log comparatively easy.

[1] Roughly speaking, a language $L$ is *monotonic* if whenever $\Pi_1$ and $\Pi_2$ are collections of statements of $L$ with $\Pi_1 \subset \Pi_2$, and $W$ is a model of $\Pi_2$, then $W$ is a model of $\Pi_1$. A language which is not monotonic is said to be *nonmonotonic*.

The choice of a probabilistic framework was more problematic and that is where Judea's ideas played a major role. Our first problem was to choose from among various conceptualizations of probability: classical, frequentist, subjective, etc. Understanding the intuitive readings of basic language constructs is crucial for a software/knowledge engineer — probably more so than for a mathematician who may be primarily interested in their mathematical properties. Judea Pearl in [Pearl 1988] introduced the authors to the subjective view of probability — i.e. understanding of probabilities as degrees of belief of a rational agent — and to the use of subjective probability in AI. This matched well with the ASP-based logic side of the language. The ASP part of a P-log program can be used for describing possible beliefs, while the probabilistic part would allow knowledge engineers to quantify the degrees of these beliefs.

After deciding on an intuitive reading of probabilities, the next question was *which sorts of probabilistic statements to allow.* Fortunately, the question of concise and transparent representation of probability distributions was already addressed by Judea in [Pearl 1988], where he showed how Bayesian nets can be successfully used for this purpose. The concept was extended in [Pearl 2000] where Pearl introduced the notion of Causal Bayesian Nets (CBN's). Pearl's definition of CBN's is pioneering in three respects. First, he gives a framework where nondeterministic causal relations are the primitive relations among random variables. Second, he shows how relationships of correlation and (classical) independence *emerge* from these causal relationships in a natural way; and third he shows how this emergence is faithful to our intuitions about the difference between causality and (mere) correlation.

As we mentioned above, one of the primary desired features in the design of P-log was elaboration tolerance — defined as the ability of a representation to incorporate new knowledge with minimal revision [McCarthy 1999]. P-log inherited from ASP the ability to naturally incorporate many forms of new logical knowledge. An extension of ASP, called CR-Prolog, further improved this ability [Balduccini and Gelfond 2003]. The term "elaboration tolerance" is less well known in the field of probabilistic reasoning, but one of the primary strengths of Bayes nets as a representation is the ability to systematically and smoothly incorporate new knowledge through conditioning, using Bayes Theorem as well as algorithms given by Pearl [Pearl 1988] and others. Causal Bayesian Nets carry this a step further, by allowing us to formalize interventions in addition to (and as distinct from) observations, and smoothly incorporate either kind of new knowledge in the form of updates. Thus from the standpoint of elaboration tolerance, CBN's were a natural choice as a probabilistic foundation for P-log.

Another reason for choosing CBN's is that we simply believe Pearl's distinction between observations and interventions to be central to commonsense probabilistic reasoning. It gives a precise mathematical basis for distinguishing between the following questions: (1) what can I expect to happen given that I observe $X = x$, and (2) what can I expect to happen if I *intervene in the normal operation of*

*a probabilistic system* by fixing value of variable $X$ to $x$? These questions could in theory be answered using classical methods, but only by creating a separate probabilistic model for each question. In a CBN these two questions may be treated as conditional probabilities (one conditioned on an observation and the other on an action) of a single probabilistic model.

P-log carries things another step. There are many actions one could take to manipulate a system besides fixing the values of (otherwise random) variables — and the effects of such actions are well studied under headings associated with ASP. Moreover, besides actions, there are many sorts of information one might gain besides those which simply eliminate possible worlds: one may gain knowledge which introduces new possible worlds, alters the probabilities of possible worlds, introduces new logical rules, etc. ASP has been shown to be a good candidate for handling such updates in non-probabilistic settings, and our hypothesis was that it would serve as well when combined with a probabilistic representation. Thus some of the key advantages of Bayesian nets, which are amplified by CBN's, show plausible promise of being even further amplified by their combination with ASP. This is the methodology of P-log: to combine a well studied method for elaboration tolerant probabilistic representations (CBN's) with a well studied method for elaboration tolerant logical representations (ASP).

Finally let us say a few words about the current status of the language. It is comparatively new. The first publication on the subject appeared in [Baral, Gelfond, and Rushton 2004], and the full journal paper describing the language appeared only recently in [Baral, Gelfond, and Rushton 2009]. The use of P-log for knowledge representation was also explored in [Baral and Hunsaker 2007] and [Gelfond, Rushton, and Zhu 2006]. A prototype reasoning system based on ASP computation allowed the use of the language for a number of applications (see, for instance, [Baral, Gelfond, and Rushton 2009; Pereira and Ramli 2009]). We are currently working on the development and implementation of a more efficient system, and on expanding it to allow rules of CR-Prolog. Finding ways for effectively combining ASP-based computational methods of P-log with recent advanced algorithms for Bayesian nets is probably one of the most interesting open questions in this area.

The paper is organized as follows. Section 2 contains short introduction to ASP and CR-Prolog. Section 3 describes the syntax and informal semantics of P-log, illustrating both through a nontrivial example. Section 4 gives another example, similar in nature to Simpson's Paradox. Section 5 states a new theorem which extends the semantics of P-log from that given in [Baral, Gelfond, and Rushton 2009] to cover programs with infinitely many random variables. The basic idea of Section 5 is accessible to a general audience, but its technical details require an understanding of the material presented in [Baral, Gelfond, and Rushton 2009].

## 2 Preliminaries

This section contains a description of syntax and semantics of both ASP and CR-Prolog. In what follows we use a standard notion of a sorted signature from classical logic. Terms and atoms are defined as usual. An atom $p(\bar{t})$ and its negation $\neg p(\bar{t})$ are referred to as *literals*. Literals of the form $p(\bar{t})$ and $\neg p(\bar{t})$ are called *contrary*. ASP and CR-Prolog also contain connectives *not* and *or* which are called *default negation* and *epistemic disjunction* respectively. Literals possibly preceded by default negation are called *extended literals*.

An ASP program is a pair consisting of a signature $\sigma$ and a collection of rules of the form

$$l_0 \ or \ \ldots \ or \ l_m \leftarrow l_{m+1}, \ldots, l_k, not \ l_{k+1}, \ldots, not \ l_n \tag{1}$$

where $l$'s are literals. The right-hand side of of the rule is often referred to as the rule's *body*, the left-hand side as the rule's head.

The answer set semantics of a logic program $\Pi$ assigns to $\Pi$ a collection of *answer sets* – partial interpretations[2] corresponding to possible sets of beliefs which can be built by a rational reasoner on the basis of rules of $\Pi$. In the construction of such a set $S$, the reasoner is assumed to be guided by the following informal principles:

- $S$ must satisfy the rules of $\Pi$;
- the reasoner should adhere to the *rationality principle*, which says that *one shall not believe anything one is not forced to believe*.

To understand the former let us consider a partial interpretation $S$ viewed as a possible set of beliefs of our reasoner. A ground atom $p$ is satisfied by $S$ if $p \in S$, i.e., the reasoner believes $p$ to be true. According to the semantics of our connectives $\neg p$ means that $p$ is false. Consequently, $\neg p$ is satisfied by $S$ iff $\neg p \in S$, i.e., the reasoner believes $p$ to be false. Unlike $\neg p$, *not* $p$ has an epistemic character and is read as *there is no reason to believe that* $p$ *is true*. Accordingly, $S$ satisfies *not* $l$ if $l \notin S$. (Note that it is possible for the reasoner to believe neither $p$ nor $\neg p$). An epistemic disjunction $l_1$ *or* $l_2$ is satisfied by $S$ if $l_1 \in S$ or $l_2 \in S$, i.e., the reasoner believes at least one of the disjuncts to be true. Finally, $S$ satisfies the body (resp., head) of rule (1) if $S$ satisfies all of the extended literals occurring in its body (resp., head); and $S$ satisfies rule (1) if $S$ satisfies its head or does not satisfy its body.

What is left is to capture the intuition behind the rationality principle. This will be done in two steps.

DEFINITION 1 (Answer Sets, Part I). Let program $\Pi$ consist of rules of the form:

$$l_0 \ or \ \ldots \ or \ l_i \leftarrow l_{i+1}, \ldots, l_m.$$

An answer set of $\Pi$ is a consistent set $S$ of ground literals such that:

[2] By partial interpretation we mean a consistent set of ground literals of $\sigma(\Pi)$.

- $S$ satisfies the rules of $\Pi$.
- $S$ is minimal; i.e., no proper subset of $S$ satisfies the rules of $\Pi$.

The rationality principle here is captured by the minimality condition. For example, it is easy to see that { } is the only answer set of program consisting of the single rule $p \leftarrow p$, and hence the reasoner associated with it knows nothing about the truth or falsity of $p$. The program consisting of rules

$p(a)$.
$q(a)$ *or* $q(b) \leftarrow p(a)$.

has two answer sets: $\{p(a), q(a)\}$ and $\{p(a), q(b)\}$. Note that no rule requires the reasoner to believe in both $q(a)$ and $q(b)$. Hence he believes that the two formulas $p(a)$ and ($q(a)$ *or* $q(b)$) are true, and that $\neg p(a)$ is false. He remains undecided, however, about, say, the two formulas $p(b)$ and ($\neg q(a)$ *or* $\neg q(b)$). Now let us consider an arbitrary program:

DEFINITION 2 (Answer Sets, Part II). Let $\Pi$ be an arbitrary collection of rules (1) and $S$ a set of literals. By $\Pi^S$ we denote the program obtained from $\Pi$ by

1. removing all rules containing *not* $l$ such that $l \in S$;
2. removing all other premises containing *not* .

$S$ is an answer set of $\Pi$ iff $S$ is an answer set of $\Pi^S$.

To illustrate the definition let us consider a program

$p(a)$.
$p(b)$.
$\neg p(X) \leftarrow$ *not* $p(X)$.

where $p$ is a unary predicate whose domain is the set $\{a, b, c\}$. The last rule, which says that if $X$ is not believed to satisfy $p$ then $p(X)$ is false, is the ASP formalization of a Closed World Assumption for a relation $p$ [Reiter 1978]. It is easy to see that $\{p(a), p(b), \neg p(c)\}$ is the only answer set of this program. If we later learn that $c$ satisfies $p$, this information can be simply added to the program as $p(c)$. The default for $c$ will be defeated and the only answer set of the new program will be $\{p(a), p(b), p(c)\}$.

The next example illustrates the ASP formalization of a more general default. Consider a statement: "Normally, computer science courses are taught only by computer science professors. The logic course is an exception to this rule. It may be taught by faculty from the math department." This is a typical *default* with a *weak exception*[3] which can be represented in ASP by the rules:

[3]An exception to a default is called *weak* if it stops application of the default without defeating its conclusion. Otherwise it is called *strong*.

$$\begin{array}{lcl} \neg may\_teach(P,C) & \leftarrow & \neg member(P,cs), \\ & & course(C,cs), \\ & & not\ ab(d_1(P,C)), \\ & & not\ may\_teach(P,C). \\ ab(d_1(P,logic)) & \leftarrow & not\ \neg member(P,math). \end{array}$$

Here $d_1(P,C)$ is the name of the default rule and $ab(d_1(P,C))$ says that default $d_1(P,C)$ is not applicable to the pair $\langle P,C\rangle$. The second rule above stops the application of the default in cases where the class is *logic* and $P$ *may be* a math professor. Used in conjunction with rules:

$member(john,cs)$.
$member(mary,math)$.
$member(bob,ee)$.
$\neg member(P,D) \leftarrow not\ member(P,D)$.
$course(logic,cs)$.
$course(data\_structures,cs)$.

the program will entail that Mary does not teach data structures while she may teach logic; Bob teaches neither logic nor data structures, and John may teach both classes.

The previous examples illustrate the representation of defaults and their strong and weak exceptions. There is another type of possible exception to defaults, sometimes referred to as an **indirect exception**. Intuitively, these are rare exceptions that come into play only as a last resort, to restore the consistency of the agent's world view when all else fails. The representation of indirect exceptions seems to be beyond the power of ASP. This observation led to the development of a simple but powerful extension of ASP called **CR-Prolog** (or ASP with consistency-restoring rules). To illustrate the problem let us consider the following example.

Consider an ASP representation of the default "elements of class $c$ normally have property $p$":

$$\begin{array}{lcl} p(X) & \leftarrow & c(X), \\ & & not\ ab(d(X)), \\ & & not\ \neg p(X). \end{array}$$

together with the rule

$$q(X) \leftarrow p(X).$$

and the facts $c(a)$ and $\neg q(a)$. Let us denote this program by $E$, where $E$ stands for "exception".

It is not difficult to check that $E$ is inconsistent. No rules allow the reasoner to prove that the default is not applicable to $a$ (i.e. to prove $ab(d(a))$) or that $a$ does not have property $p$. Hence the default must conclude $p(a)$. The second rule implies $q(a)$ which contradicts one of the facts. However, there seems to exists a

commonsense argument which may allow a reasoner to avoid inconsistency, and to conclude that $a$ is an indirect exception to the default. The argument is based on the **Contingency Axiom** for default $d(X)$ which says that *any element of class c can be an exception to the default $d(X)$ above, but such a possibility is very rare, and, whenever possible, should be ignored*. One may informally argue that since the application of the default to $a$ leads to a contradiction, the possibility of $a$ being an exception to $d(a)$ cannot be ignored and hence $a$ must satisfy this rare property.

In what follows we give a brief description of CR-Prolog — an extension of ASP capable of encoding and reasoning about such rare events.

A program of CR-Prolog is a four-tuple consisting of

1. A (possibly sorted) signature.
2. A collection of regular rules of ASP.
3. A collection of rules of the form

$$l_0 \stackrel{+}{\leftarrow} l_1, \ldots, l_k, not\ l_{k+1}, \ldots, not\ l_n \tag{2}$$

   where $l$'s are literals. Rules of this type are called *consistency restoring* rules (CR-rules).

4. A partial order, $\leq$, defined on sets of CR-rules. This partial order is often referred to as a **preference relation**.

Intuitively, rule (2) says that if the reasoner associated with the program believes the body of the rule, then he "may possibly" believe its head. However, this possibility may be used only if there is no way to obtain a consistent set of beliefs by using only regular rules of the program. The partial order over sets of CR-rules will be used to select preferred possible resolutions of the conflict. Currently the inference engine of CR-Prolog [Balduccini 2007] supports two such relations, denoted $\leq_1$ and $\leq_2$. One is based on the set-theoretic inclusion ($R_1 \leq_1 R_2$ holds iff $R_1 \subseteq R_2$). The other is defined by the cardinality of the corresponding sets ($R_1 \leq_2 R_2$ holds iff $|R_1| \leq |R_2|$). To give the precise semantics we will need some terminology and notation.

The set of regular rules of a CR-Prolog program $\Pi$ will be denoted by $\Pi^r$, and the set of CR-rules of $\Pi$ will be denoted by $\Pi^{cr}$. By $\alpha(r)$ we denote a regular rule obtained from a consistency restoring rule $r$ by replacing $\stackrel{+}{\leftarrow}$ by $\leftarrow$. If $R$ is a set of CR-rules then $\alpha(R) = \{\alpha(r) \ : \ r \in R\}$. As in the case of ASP, the semantics of CR-Prolog will be given for ground programs. A rule with variables will be viewed as a shorthand for a set of ground rules.

DEFINITION 3. (Abductive Support)
A minimal (with respect to the preference relation of the program) collection $R$ of

CR-rules of $\Pi$ such that $\Pi^r \cup \alpha(R)$ is consistent (i.e. has an answer set) is called an **abductive support** of $\Pi$.

DEFINITION 4. (Answer Sets of CR-Prolog)
A set $A$ is called an *answer set* of $\Pi$ if it is an answer set of a regular program $\Pi^r \cup \alpha(R)$ for some abductive support $R$ of $\Pi$.

Now let us show how CR-Prolog can be used to represent defaults and their indirect exceptions. The CR-Prolog representation of the default $d(X)$, which we attempted to represent in ASP program $E$, may look as follows

$$\begin{array}{rcl} p(X) & \leftarrow & c(X), \\ & & not\ ab(d(X)), \\ & & not\ \neg p(X). \\ \neg p(X) & \stackrel{+}{\leftarrow} & c(X). \end{array}$$

The first rule is the standard ASP representation of the default, while the second rule expresses the Contingency Axiom for the default $d(X)$[4]. Consider now a program obtained by combining these two rules with an atom $c(a)$.

Assuming that $a$ is the only constant in the signature of this program, the program's unique answer set will be $\{c(a), p(a)\}$. Of course this is also the answer set of the regular part of our program. (Since the regular part is consistent, the Contingency Axiom is ignored.) Let us now expand this program by the rules

$q(X) \leftarrow p(X)$.
$\neg q(a)$.

The regular part of the new program is inconsistent. To save the day we need to use the Contingency Axiom for $d(a)$ to form the abductive support of the program. As a result the new program has the answer set $\{\neg q(a), c(a), \neg p(a))\}$. The new information does not produce inconsistency, as it did in ASP program $E$. Instead the program withdraws its previous conclusion and recognizes $a$ as a (strong) exception to default $d(a)$.

## 3 The Language

A P-log program consists of its *declarations*, *logical rules*, *random selection rules*, *probability atoms*, *observations*, and *actions*. We will begin this section with a brief description of the syntax and informal readings of these components of the programs, and then proceed to an illustrative example.

The declarations of a P-log program give the types of objects and functions in the program. Logical rules are "ordinary" rules of the underlying logical language

[4] In this form of Contingency Axiom, we treat $X$ as a strong exception to the default. Sometimes it may be useful to also allow weak indirect exceptions; this can be achieved by adding the rule $ab(d(X)) \stackrel{+}{\leftarrow} c(X)$.

written using light syntactic sugar. For purposes of this paper, the underlying logical language is CR-Prolog.

P-log uses *random selection rules* to declare random attributes (essentially random variables) of the form $a(\bar{t})$, where $a$ is the name of the attribute and $\bar{t}$ is a vector of zero or more parameters. In this paper we consider random selection rules of the form

$$[\ r\ ]\ random(a(\bar{t})) \leftarrow B. \tag{3}$$

where $r$ is a term used to name the random causal process associated with the rule and $B$ is a conjunction of zero or more extended literals. The name $[\ r\ ]$ is optional and can be omitted if the program contains exactly one random selection rule for $a(\bar{t})$. Statement (3) says that *if $B$ were to hold, the value of $a(\bar{t})$ would be selected at random from its range by process $r$, unless this value is fixed by a deliberate action.* More general forms of random selection rules, where the values may be selected from a range which depends on context, are discussed in [Baral, Gelfond, and Rushton 2009].

Knowledge of the numeric probabilities of possible values of random attributes is expressed through *causal probability atoms*, or *pr*-atoms. A *pr*-atom takes the form

$$pr_r(a(\bar{t}) = y|_c\ B) = v$$

where $a(\bar{t})$ is a random attribute, $B$ a conjunction of literals, $r$ is a causal process, $v \in [0, 1]$, and $y$ is a possible value of $a(\bar{t})$. The statement says that *if the value of $a(\bar{t})$ is fixed by process $r$, and $B$ holds, then the probability that $r$ causes $a(\bar{t}) = y$ is $v$*. If $r$ is uniquely determined by the program then it can be omitted. The "causal stroke" '$|_c$' and the "rule body" $B$ may also be omitted in case $B$ is empty.

Observations and actions of a P-log program are written, respectively, as

$$obs(l). \qquad\qquad do(a(\bar{t}) = y)).$$

where $l$ is a literal, $a(\bar{t})$ a random attribute, and $y$ a possible value of $a(\bar{t})$. $obs(l)$ is read *$l$ is observed to be true*. The action $do(a(\bar{t}) = y)$ is read *the value of $a(\bar{t})$, instead of being random, is set to $y$ by a deliberate action.*

This completes a general introductory description of P-log. Next we give an example to illustrate this description. The example shows how certain forms of knowledge may be represented, including deterministic causal knowledge, probabilistic causal knowledge, and strict and defeasible logical rules (a rule is *defeasible* if it states an overridable presumption; otherwise it is *strict*). We will use this example to illustrate the syntax of P-log, and, afterward, to provide an indication of the formal semantics. Complete syntax and semantics are given in [Baral, Gelfond, and Rushton 2009], and the reader is invited to refer there for more details.

EXAMPLE 5. [Circuit]
A circuit has a motor, a breaker, and a switch. The switch may be open or closed. The breaker may be tripped or not; and the motor may be turning or not. The operator may toggle the switch or reset the breaker. If the switch is closed and the system is functioning normally, the motor turns. The motor never turns when the switch is open, the breaker is tripped, or the motor is burned out. The system may break and if so the break could consist of a tripped breaker, a burned out motor, or both, with respective probabilities .9, .09, and .01. Breaking, however, is rare, and should be considered only in the absence of other explanations.

Let us show how to represent this knowledge in P-log. First we give declarations of sorts and functions relevant to the domain. As typical for representation of dynamic domains we will have sorts for actions, fluents (properties of the domain which can be changed by actions), and time steps. Fluents will be partitioned into inertial fluents and defined fluents. The former are subject to the law of inertia [Hayes and McCarthy 1969] (which says that things stay the same by default), while the latter are specified by explicit definitions in terms of already defined fluents. We will also have a sort for possible types of breaks which may occur in the system. In addition to declared sorts P-log contains a number of predefined sorts, e.g. a sort *boolean*. Here are the sorts of the domain for the circuit example:

$action = \{toggle, reset, break\}$.

$inertial\_fluent = \{closed, tripped, burned\}$.

$defined\_fluent = \{turning, faulty\}$.

$fluent = inertial\_fluent \cup defined\_fluent$.

$step = \{0, 1\}$.

$breaks = \{trip, burn, both\}$.

In addition to sorts we need to declare functions (referred in P-log as *attributes*) relevant to our domain.

$holds \ : \ fluent \times step \rightarrow boolean$.

$occurs \ : \ action \times step \rightarrow boolean$.

Here $holds(f, T)$ says that fluent $f$ is true at time step $T$ and $occurs(a, T)$ indicates that action $a$ was executed at $T$.

The last function we need to declare is a random attribute $type\_of\_break(T)$ which denotes the type of an occurrence of action *break* at step $T$.

$type\_of\_break \ : \ step \rightarrow breaks$.

The first two logical rules of the program define the direct effects of action *toggle*.

$$\begin{array}{lcl} holds(closed, T+1) & \leftarrow & occurs(toggle, T), \\ & & \neg holds(closed, T). \\ \neg holds(closed, T+1) & \leftarrow & occurs(toggle, T), \\ & & holds(closed, T). \end{array}$$

They simply say that toggling opens and closes the switch. The next rule says that resetting the breaker untrips it.

$$\neg holds(tripped, T+1) \quad \leftarrow \quad occurs(reset, T).$$

The effects of action *break* are described by the rules

$$\begin{array}{lcl} holds(tripped, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = trip. \\ holds(burned, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = burn. \\ holds(tripped, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = both. \\ holds(burned, T+1) & \leftarrow & occurs(break, T), \\ & & type\_of\_break(T) = both. \end{array}$$

The next two rules express the inertia axiom which says that *by default, things stay as they are.* They use default negation *not* — the main nonmonotonic connective of ASP —, and can be viewed as typical representations of defaults in ASP and its extensions.

$$\begin{array}{lcl} holds(F, T+1) & \leftarrow & inertial\_fluent(F), \\ & & holds(F, T), \\ & & not\ \neg holds(F, T+1). \\ \neg holds(F, T+1) & \leftarrow & inertial\_fluent(F), \\ & & \neg holds(F, T), \\ & & not\ holds(F, T+1). \end{array}$$

Next we explicitly define fluents *faulty* and *turning*.

$$\begin{array}{lcl} holds(faulty, T) & \leftarrow & holds(tripped, T). \\ holds(faulty, T) & \leftarrow & holds(burned, T). \\ \neg holds(faulty, T) & \leftarrow & not\ holds(faulty, T). \end{array}$$

The rules above say that the system is functioning abnormally if and only if the breaker is tripped or the motor is burned out. Similarly the next definition says that the motor turns if and only if the switch is closed and the system is functioning normally.

$$\begin{array}{lcl} holds(turning, T) & \leftarrow & holds(closed, T), \\ & & \neg holds(faulty, T). \\ \neg holds(turning, T) & \leftarrow & not\ holds(turning, T). \end{array}$$

The above rules are sufficient to define causal effects of actions. For instance if we assume that at Step 0 the motor is turning and the breaker is tripped, i.e.

action *break* of the type *trip* occurred at 0, then in the resulting state we will have $holds(tripped, 1)$ as the direct effect of this action; while $\neg holds(turning, 1)$ will be its indirect effect[5].

We will next have a default saying that for each action $A$ and time step $T$, in the absence of a reason to believe otherwise we assume $A$ does not occur at $T$.

$\neg occurs(A, T) \leftarrow action(A), not\ occurs(A, T).$

We next state a CR-rule representing possible exceptions to this default. The rule says that a break to the system may be considered if necessary (that is, necessary in order to reach a consistent set of beliefs).

$occurs(break, 0) \stackrel{+}{\leftarrow}$ .

The next collection of facts describes the initial situation of our story.

$\neg holds(closed, 0).\quad \neg holds(burned, 0).\quad \neg holds(tripped, 0).\quad occurs(toggle, 0).$

Next, we state a random selection rule which captures the non-determinism in the description of our circuit.

$random(type\_of\_break(T)) \leftarrow occurs(break, T).$

The rule says that if action *break* occurs at step $T$ then the type of break will be selected at random from the range of possible types of breaks, unless this type is fixed by a deliberate action. Intuitively, *break* can be viewed as a non-deterministic action, with non-determinism coming from the lack of knowledge about the precise type of *break*.

Let $\pi_0$ be the circuit program given so far. Next we will give a sketch of the formal semantics of P-log, using $\pi_0$ as an illustrative example.

The *logical part* of a P-log program $\Pi$ consists of its declarations, logical rules, random selection rules, observations, and actions; while its *probabilistic part* consists of its *pr*-atoms (though the above program does not have any). The semantics of P-log describes a translation of the logical part of $\Pi$ into an "ordinary" CR-Prolog program $\tau(\Pi)$. The semantics of $\Pi$ is then given by

[5]It is worth noticing that, though short, our formalization of the circuit is non-trivial. It is obtained using the general methodology of representing dynamic systems modeled by transition diagrams whose nodes correspond to physically possible states of the system and whose arcs are labeled by actions. A transition $\langle \sigma_0, a, \sigma_1 \rangle$ indicates that state $\sigma_1$ may be a result of execution of $a$ in $\sigma_0$. The problem of finding concise and mathematically accurate description of such diagrams has been a subject of research for over 30 years. Its solution requires a good understanding of the nature of causal effects of actions in the presence of complex interrelations between fluents. An additional level of complexity is added by the need to specify what is not changed by actions. As noticed by John McCarthy, the latter, known as the Frame Problem, can be reduced to finding a representation of the Inertia Axiom which requires the ability to represent defaults and to do non-monotonic reasoning. The representation of this axiom as well as that of the interrelations between fluents we used in this example is a simple special case of general theory of action and change based on logic programming under the answer set semantics.

1. a collection of answer sets of $\tau(\Pi)$ viewed as the set of possible worlds of a rational agent associated with $\Pi$, along with

2. a probability measure over these possible worlds, determined by the collection of the probability atoms of $\Pi$.

To obtain $\tau(\pi_0)$ we represent sorts as collections of facts. For instance, sort *step* would be represented in CR-Prolog as

$step(0). \quad step(1).$

For a non-boolean function *type_of_break* the occurrences of atoms of the form $type\_of\_break(T) = trip$ in $\pi_0$ are replaced by $type\_of\_break(T, trip)$. Similarly for *burn* and *both*. The translation also contains the axiom

$$\begin{array}{rcl} \neg type\_of\_break(T, V_1) & \leftarrow & breaks(V_1), breaks(V_2), V_1 \neq V_2, \\ & & type\_of\_break(T, V_2). \end{array}$$

to guarantee that *type_of_break* is a function. In general, the same transformation is performed for all non-boolean functions.

Logical rules of $\pi_0$ are simply inserted into $\tau(\pi_0)$. Finally, the random selection rule is transformed into

$$\begin{array}{l} type\_of\_break(T, trip) \text{ or } type\_of\_break(T, burn) \text{ or } type\_of\_break(T, both) \leftarrow \\ \qquad\qquad occurs(break, T), \\ \qquad\qquad not\ intervene(type\_of\_break(T)). \end{array}$$

It is worth pointing out here that while CBN's represent the notion of intervention in terms of transformations on graphs, P-log axiomatizes the semantics of intervention by including *not intervene*(...) in the body of the translation of each random selection rule. This amounts to a *default presumption* of randomness, overridable by intervention. We will see next how actions using *do* can defeat this presumption.

Observations and actions are translated as follows. For each literal $l$ in $\pi_0$, $\tau(\pi_0)$ contains the rule

$\leftarrow obs(l), not\ l.$

For each atom $a(\bar{t}) = y$, $\tau(\pi)$ contains the rules

$a(\bar{t}, y) \leftarrow do(a(\bar{t}, y)).$

and

$intervene(a(\bar{t})) \leftarrow do(a(\bar{t}, Y)).$

The first rule eliminates possible worlds of the program failing to satisfy $l$. The second rule makes sure that interventions affect their intervened-upon variables in the expected way. The third rule defines the relation *intervene* which, for each action, cancels the randomness of the corresponding attribute.

It is not difficult to check that under the semantics of CR-Prolog, $\tau(\pi_0)$ has a unique possible world $W$ containing $holds(closed, 1)$ and $holds(turning, 1)$, the direct and indirect effects, respectively, of the action *close*. Note that the collection of regular ASP rules of $\tau(\pi_0)$ is consistent, i.e., has an answer set. This means that CR-rule $occurs(break, 0) \stackrel{+}{\leftarrow}$ is not activated, break does not occur, and the program contains no randomness.

Now we will discuss how probabilities are computed in P-log. Let $\Pi$ be a P-log program containing the random selection rule $[r]\ random(a(\bar{t})) \leftarrow B_1$ and the *pr*-atom $pr_r(a(\bar{t}) = y \mid_c B_2) = v$. Then if $W$ is a possible world of $\Pi$ satisfying $B_1$ and $B_2$, the *assigned probability* of $a(\bar{t}) = y$ in $W$ is defined [6] to be $v$. In case $W$ satisfies $B_1$ and $a(\bar{t}) = y$, but there is no *pr*-atom $pr_r(a(\bar{t} = y \mid_c B_2) = v$ of $\Pi$ such that $W$ satisfies $B_2$, then the *default probability* of $a(\bar{t}) = y$ in $W$ is computed using the "indifference principle", which says that two possible values of a random selection are equally likely if we have no reason to prefer one to the other (see [Baral, Gelfond, and Rushton 2009] for details). The *probability* of each random atom $a(\bar{t}) = y$ occurring in each possible world $W$ of program $\Pi$, written $P_\Pi(W, a(\bar{t}) = y)$, is now defined to be the assigned probability or the default probability, as appropriate.

Let $W$ be a possible world of $\Pi$. The *unnormalized probability*, $\hat{\mu}_\Pi(W)$, of a possible world $W$ *induced by* $\Pi$ is

$$\hat{\mu}_\Pi(W) =_{def} \prod_{a(\bar{t},y) \in\ W} P_\Pi(W, a(\bar{t}) = y)$$

where the product is taken only over atoms for which $P(W, a(\bar{t}) = y)$ is defined.

Suppose $\Pi$ is a P-log program having at least one possible world with nonzero unnormalized probability, and let $\Omega$ be the set of possible worlds of $\Pi$. The *measure*, $\mu_\Pi(W)$, of a possible world $W$ *induced by* $\Pi$ is the unnormalized probability of $W$ divided by the sum of the unnormalized probabilities of all possible worlds of $\Pi$, i.e.,

$$\mu_\Pi(W) =_{def} \frac{\hat{\mu}_\Pi(W)}{\sum_{W_i \in \Omega} \hat{\mu}_\Pi(W_i)}$$

When the program $\Pi$ is clear from context we may simply write $\hat{\mu}$ and $\mu$ instead of $\hat{\mu}_\Pi$ and $\mu_\Pi$ respectively.

This completes the discussion of how probabilities of possible worlds are defined in P-log. Now let us return to the circuit example. Let program $\pi_1$ be the union of $\pi_0$ with the single observation

$obs(\neg holds(turning, 1))$

The observation contradicts our previous conclusion $holds(turning, 1)$ reached by using the effect axiom for *toggle*, the definitions of $faulty$ and $turning$, and the

[6] For the sake of well definiteness, we consider only programs in which at most one $v$ satisfies this definition.

inertia axiom for *tripped* and *burned*. The program $\tau(\pi_1)$ will resolve this contradiction by using the CR-rule $occurs(break, 0) \stackrel{+}{\leftarrow}$ to conclude that the action *break* occurred at Step 0. Now *type_of_break* randomly takes one of its possible values. Accordingly, $\tau(\pi_1)$ has three answer sets: $W_1$, $W_2$, and $W_3$. All of them contain $occurs(break, 0)$, $holds(faulty, 1)$, $\neg holds(turning, 1)$. One, say $W_1$ will contain

$type\_of\_break(1, trip)$, $holds(tripped, 1)$, $\neg holds(burned, 1)$

$W_2$ and $W_3$ will respectively contain

$type\_of\_break(1, burn)$, $\neg holds(tripped, 1)$, $holds(burned, 1)$

and

$type\_of\_break(1, both)$, $holds(tripped, 1)$, $holds(burned, 1)$

In accordance with our general definition, $\pi_1$ will have three possible worlds, $W_1$, $W_2$, and $W_3$. The probabilities of each of these three possible worlds can be computed as 1/3, using the indifference principle.

Now let us add some quantitative probabilities to our program. If $\pi_2$ is the union of $\pi_1$ with the following three *pr*-atoms

$pr(type\_of\_break(T) = trip \mid_c break(T)) = 0.9$
$pr(type\_of\_break(T) = burned \mid_c break(T)) = 0.09$
$pr(type\_of\_break(T) = both \mid_c break(T)) = 0.01$

then program $\pi_2$ has the same possible worlds as $\Pi_1$. Not surprisingly, $P_{\pi_2}(W_1) = 0.9$. Similarly $P_{\pi_2}(W_2) = 0.09$ and $P_{\pi_2}(W_3) = 0.01$. This demonstrates how a P-log program may be written in stages, with quantitative probabilities added as they are needed or become available.

Typically we are interested not just in the probabilities of individual possible worlds, but in the probabilities of certain interesting sets of possible worlds described, e.g., those described by formulae. For current purposes a rather simple definition suffices. Viz., recalling that possible worlds are sets of literals, for an arbitrary set $C$ of literals we define

$$P_\pi(C) =_{def} P_\pi(\{W : C \subseteq W\}).$$

For example, $P_{\pi_1}(holds(turning, 1)) = 0$, $P_{\pi_1}(holds(tripped, 1)) = 1/3$,
and $P_{\pi_2}(holds(tripped, 1)) = 0.91$.

Our example is in some respects rather simple. For instance, every possible world of our program contains at most one atom of the form $a(\bar{t}) = y$ where $a(\bar{t})$ is a random attribute. We hope, however, that this example gives a reader some insight in the syntax and semantics of P-log. It is worth noticing that the example shows the ability of P-log to mix logical and probabilistic reasoning, including reasoning about causal effects of actions and explanations of observations. In addition it

demonstrates the non-monotonic character of P-log, i.e. its ability to react to new knowledge by changing probabilistic models of the domain and creating new possible worlds.

The ability to introduce new possible worlds as a result of conditioning is of interest from two standpoints. First, it reflects the common sense semantics of utterances such as "the motor might be burned out." Such a sentence does not eliminate existing possible beliefs, and so there is no classical (i.e., monotonic) semantics in which the statement would be informative. If it is informative, as common sense suggests, then its content seems to introduce new possibilities into the listener's thought process.

Second, nonmonotonicity can improve performance. Possible worlds tend to proliferate exponentially with the size of a program, quickly making computations intractable. The ability to consider only those random selections which may explain our abnormal observations may make computations tractable for larger programs. Even though our current solver is in its early stages of development, it is based on well researched answer set solvers which efficiently eliminate impossible worlds from consideration based on logical reasoning. Thus even our early prototype has shown promising performance on problems where logic may be used to exclude possible worlds from consideration in the computation of probabilities [Gelfond, Rushton, and Zhu 2006].

## 4 Spider Example

In this section, we consider a variant of Simpson's paradox, to illustrate the formalization of interventions in P-log. The story we would like to formalize is as follows:

In Stan's home town there are two kinds of poisonous spider, the creeper and the spinner. Bites from the two are equally common in Stan's area — though spinner bites are more common on a worldwide basis. An experimental anti-venom has been developed to treat bites from either kind of spider, but its effectiveness is questionable.

One morning Stan wakes to find he has a bite on his ankle, and drives to the emergency room. A doctor examines the bite, and concludes it is a bite from either a creeper or a spinner. In deciding whether to administer the anti-venom, the doctor examines the data he has on bites from the two kinds of spiders: out of 416 people bitten by the creeper worldwide, 312 received the anti-venom and 104 did not. Among those who received the anti-venom, 187 survived; while 73 survived who did not receive anti-venom. The spinner is more deadly and tends to inhabit areas where the treatment is less available. Of 924 people bitten by the spinner, 168 received the anti-venom, 34 of whom survived. Of the 756 spinner bite victims who did not receive the experimental treatment, only 227 survived.

For a random individual bitten by a creeper or spinner, let $s$, $a$, and $c$ denote the

events of *survival, administering anti-venom,* and *creeper bite.* Based on the fact that the two sorts of bites are equally common in Stan's region, the doctor assigns a 0.5 probability to either kind of bite. He also computes a probability of survival, with and without treatment, from each kind of bite, based on the sampling distribution of the available data. He similarly computes the probabilities that victims of each kind of bite received the anti-venom. We may now imagine the doctor uses Bayes' Theorem to compute $P(s \mid a) = 0.522$ and $P(s \mid \neg a) = 0.394$.

Thus we see that if we choose a historical victim, in such a way that he has a 50/50 chance of either kind of bite, those who received anti-venom would have a substantially higher chance of survival. Stan is in the situation of having a 50/50 chance of either sort of bite; however, he is *not* a historical victim. Since we are intervening in the decision of whether he receives anti-venom, the computation above is not germane (as readers of [Pearl 2000] already know) — though we can easily imagine the doctor making such a mistake. A correct solution is as follows. Formalizing the relevant parts of the story in a P-log program $\Pi$ gives

$survive, antivenom \ : \ boolean.$
$spider \ : \ \{creeper, spinner\}.$

$random(spider).$
$random(survive).$
$random(antivenom).$

$pr(spider = creeper) = 0.5.$

$pr(survive \mid_c spider = creeper, antivenom) = 0.6.$
$pr(survive \mid_c spider = creeper, \neg antivenom) = 0.7.$
$pr(survive \mid_c spider = spinner, antivenom) = 0.2.$
$pr(survive \mid_c spider = spinner, \neg antivenom) = 0.3.$

and so, according to our semantics,

$P_{\Pi \cup \{do(antivenom\}}(survive) = 0.4$
$P_{\Pi \cup \{do(\neg antivenom\}}(survive) = 0.5$

Thus, the correct decision, assuming we want to intervene to maximize Stan's chance of survival, is to not administer antivenom.

In order to reach this conclusion by classical probability, we would need to consider separate probability measures $P_1$ and $P_2$, on the sets of patients who received or did not receive antivenom, respectively. If this is done correctly, we obtain $P_1(s) = 0.4$ and $P_2(s) = 0.5$, as in the P-log program.

Thus we can get a correct classical solution using separate probability measures. Note however, that we could also get an *incorrect* classical solution using separate measures, since there exist probability measures $\hat{P}_1$ and $\hat{P}_2$ on the sets of historical bite victims which capture classical conditional probabilities given $a$ and $\neg a$ respectively. We may define

$\hat{P}_1(E) =_{def} \frac{P(E \cap a)}{0.3582}$

$\hat{P}_2(E) =_{def} \frac{P(E \cap \neg a)}{0.6418}$

It is well known that each of these is a probability measure. They are seldom seen only because classical conditional probability gives us simple notations for them *in terms of a single measure capturing common background knowledge*. This allows us to refer to probabilities conditioned on observations without defining a new measure for each such observation. What we do not have, classically, is a similar mechanism for probabilities conditioned on intervention — which is sometimes of interest as the example shows. The ability to condition on interventions in this way has been a fundamental contribution of Pearl; and the inclusion in P-log of such conditioning-on-intervention is a direct result of the authors' reading of his book.

## 5 Infinite Programs

The definitions given so far for P-log apply only to programs with finite numbers of random selection rules. In this section we state a theorem which allows us to extend these semantics to programs which may contain infinitely many random selection rules. No changes are required from the syntax given in [Baral, Gelfond, and Rushton 2009], and the probability measure described here agrees with the one in [Baral, Gelfond, and Rushton 2009] whenever the former is defined.

We begin by defining the class of programs for which the new semantics are applicable. The reader is referred to [Baral, Gelfond, and Rushton 2009] for the definitions of *causally ordered*, *unitary*, and *strict probabilistic levelling*.

DEFINITION 6. [Admissible Program]
A P-log program is *admissible* if it is causally ordered and unitary, and if there exists a strict probabilistic levelling $||$ on $\Pi$ such that no ground literal occurs in the heads of rules in infinitely many $\Pi_i$ with respect to $||$.

The condition of admissibility, and the definitions it relies on, are all rather involved to state precisely, but the intuition is as follows. Basically, a program is unitary if the probabilities assigned to the possible outcomes of each selection rule are either all assigned and sum to 1, or are not all assigned and their sum does not exceed 1. The program is causally ordered if its causal dependencies are acyclic and if the only nondeterminism in it is a result of random selection rules. A strict probabilistic levelling is a well ordering of the selection rules of a program which witnesses the fact that it is causally ordered. Finally, a program which meets these conditions is admissible if every ground literal in the program logically depends on only finitely many random experiments. For example, the following program is not unitary:

$random(a) : boolean.$
$pr(a) = 1/2.$
$pr(\neg a) = 2/3.$

The following program is not causally ordered:

$random(a) : boolean.$
$random(b) : boolean.$
$pr_r(a|_c\ b) = 1/3.$
$pr_r(a|_c\ \neg b) = 2/3.$
$pr_r(b|_c\ a) = 1/5.$

and neither is the following:

$p\ \leftarrow not\ q.$
$q\ \leftarrow not\ p.$

since it has two answer sets which arise from circularity of defaults, rather than random selections. The following program is both unitary and causally ordered, but not admissible, because $atLeastOneTail$ depends on infinitely many coin tosses.

$coin\_toss : positive\_integer \rightarrow \{head, tail\}.$
$atLeastOneTail : boolean.$
$random(coin\_toss(N)).$
$atLeastOneTail \leftarrow coin\_toss(N) = tail.$

We need one more definition before stating the main theorem:

DEFINITION 7. [Cylinder algebra of Π]
Let Π be a countably infinite P-log program with random attributes $a_i(t)$, $i > 0$, and let $C$ be the collection of sets of the form $\{\omega\ : a_i(t) = y \in \omega\}$ for arbitrary $t$, $i$, and $y$. The sigma algebra generated by $C$ will be called the *cylinder algebra* of program Π.

Intuitively, the cylinder algebra of a program Π is the collection of sets which can be formed by performing countably many set operations (union, intersection, and complement) upon sets whose probabilities are defined by finite subprograms. We are now ready to state the main proposition of this section.

PROPOSITION 8. *[Admissible programs]*
*Let* Π *be an admissible P-log program with at most countably infinitely many ground rules, and let A be the cylinder algebra of* Π*. Then there exists a unique probability measure* $P_\Pi$ *defined on A such that whenever* $[r]\ random(a(\bar{t})) \leftarrow B_1$ *and* $pr_r(a(\bar{t}) = y \mid B_2) = v$ *occur in* Π*, and* $P_\Pi(B_1 \wedge B_2) > 0$*, we have* $P_\Pi(a(\bar{t}) = y \mid B_1 \wedge B_2) = v$*.*

Recall that the semantic value of a P-log program Π consists of (1) a set of possible worlds of Π and (2) a probability measure on those possible worlds. The proposition now puts us in position to give semantics for programs with infinitely many random

selection rules. The possible worlds of the program are the answer sets of the associated (infinite) CR-Prolog program, as determined by the usual definition — while the probability measure is $P_\Pi$, as defined in Proposition 8.

We next give an example which exercises the proposition, in a form of a novel paradox. Imagine a casino which offers an infinite sequence of games, of which our agent may decide to play as many or as few as he wishes. For the $n^{th}$ game, a fair coin is tossed $n$ times. If the agent chooses to play the $n^{th}$ game, then the agent wins $2^{n+1}+1$ dollars if all tosses made in the $n^{th}$ game are heads and otherwise loses one dollar.

We can formalize this game as an infinite P-log program $\Pi$. First, we declare a countable sequence of games and an integer valued variable, representing the player's net winnings after each game.

$game : positive\_integer.$
$winnings : game \rightarrow integer.$
$play : game \rightarrow boolean.$
$coin : \{\langle M, N\rangle \mid 1 \leq M \leq N\} \rightarrow \{head, tail\}.$

Note that the declaration for *coin* is not written in the current syntax of P-log; but to save space we use set-builder notation here as a shorthand for the more lengthy formal declaration. Similarly, the notation $\langle M, N\rangle$ is also a shorthand. From this point on we will write $coin(M, N)$ instead of $coin(\langle M, N\rangle)$.

$\Pi$ also contains a declaration to say that the throws are random and the coin is known to be fair:

$random(coin(M, N)).$
$pr(coin(M, N) = head) = 1/2.$

The conditions of winning the $N^{th}$ game are described as follows:

$lose(N) \leftarrow play(N),\ coin(N, M) = tail.$
$win(N) \leftarrow play(N),\ not\ lose(N).$

The amount the agent wins or loses on each game is given by

$winnings(0) = 0.$
$winnings(N+1) = winnings(N) + 1 + 2^{N+1} \leftarrow win(N).$
$winnings(N+1) = winnings(N) - 1 \leftarrow lose(N).$
$winnings(N+1) = winnings(N) \leftarrow \neg play(N).$

Finally the program contains rules which describe the agent's strategy in choosing which games to play. Note that the agent's expected winnings in the $N^{th}$ game are given by $(1/2^N)(1+2^{N+1}) - (1-1/2^N) = 1$, so each game has positive expectation for the player. Thus a reasonable strategy might be to play every game, represented as

$play(N)$.

This completes program Π. It can be shown to be admissible, and hence there is a unique probability measure $P_\Pi$ satisfying the conclusion of Proposition 1. Thus, for example, $P_\Pi(coin(3,2) = head) = 1/2$, and $P_\Pi(win(10)) = 1/2^{10}$. Each of these probabilities can be computed from finite sub-programs. As more interesting example, let $S$ be the set of possible worlds in which the agent wins infinitely many games. The probability of this event cannot be computed from any finite sub-program of Π. However, $S$ is a countable intersection of countable unions of sets whose probabilities are defined by finite subprograms. In particular,

$$S = \bigcap_{N=1}^{\infty} \bigcup_{J=N}^{\infty} \{W \mid win(J) \in W\}$$

and therefore, $S$ is in the cylinder algebra of Π and so its probability is given by the measure defined in Proposition 1.

So where is the Paradox? To see this, let us compute the probability of $S$. Since $P_\Pi$ is a probability measure, it is monotonic in the sense that no set has greater probability than any of its subsets. $P_\Pi$ must also be *countably subadditive*, meaning that the probability of a countable union of sets cannot exceed the sum of their probabilities. Thus, from the above we get for every $N$,

$$\begin{aligned} P_\Pi(S) &< P_\Pi(\bigcup_{J=N}^{\infty} \{W \mid win(J) \in W\} \\ &\leq \sum_{J=N}^{\infty} P_\Pi(\{W \mid win(J) \in W\}) \\ &= \sum_{J=N}^{\infty} 1/2^J \\ &= 1/2^N \end{aligned}$$

Now since right hand side can be made arbitrarily small by choosing a sufficiently large $N$, it follows that $P_\Pi(S) = 0$. Consequently, with probability 1, our agent will *lose* all but finitely many of the games he plays. Since he loses one dollar per play indefinitely after his final win, his winnings converge to $-\infty$ with probability 1, even though each of his wagers has positive expectation!

**Acknowledgement**
The first author was partially supported in this research by iARPA.

## References

Balduccini, M. (2007). CR-MODELS: An inference engine for CR-Prolog. In C. Baral, G. Brewka, and J. Schlipf (Eds.), *Proceedings of the 9th Inter-*

*national Conference on Logic Programming and Non-Monotonic Reasoning (LPNMR'07)*, Volume 3662 of *Lecture Notes in Artificial Intelligence*, pp. 18–30. Springer.

Balduccini, M. and M. Gelfond (2003, Mar). Logic Programs with Consistency-Restoring Rules. In P. Doherty, J. McCarthy, and M.-A. Williams (Eds.), *International Symposium on Logical Formalization of Commonsense Reasoning*, AAAI 2003 Spring Symposium Series, pp. 9–18.

Baral, C. (2003). *Knowledge representation, reasoning and declarative problem solving with answer sets*. Cambridge University Press.

Baral, C., M. Gelfond, and N. Rushton (2004, Jan). Probabilistic Reasoning with Answer Sets. In *Proceedings of LPNMR-7*.

Baral, C., M. Gelfond, and N. Rushton (2009). Probabilistic reasoning with answer sets. *Journal of Theory and Practice of Logic Programming (TPLP) 9*(1), 57–144.

Baral, C. and M. Hunsaker (2007). Using the probabilistic logic programming language p-log for causal and counterfactual reasoning and non-naive conditioning. In *Proceedings of IJCAI-2007*, pp. 243–249.

Gelfond, M. and V. Lifschitz (1988). The stable model semantics for logic programming. In *Proceedings of ICLP-88*, pp. 1070–1080.

Gelfond, M. and V. Lifschitz (1991). Classical negation in logic programs and disjunctive databases. *New Generation Computing 9*(3/4), 365–386.

Gelfond, M., N. Rushton, and W. Zhu (2006). Combining logical and probabilistic reasoning. AAAI 2006 Spring Symposium Series, pp. 50–55.

Hayes, P. J. and J. McCarthy (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press.

McCarthy, J. (1999). Elaboration tolerance. In progress.

Pearl, J. (1988). *Probabistic reasoning in intelligent systems: networks of plausable inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pereira, L. M. and C. Ramli (2009). Modelling decision making with probabilistic causation. *Intelligent Decision Technologies (IDT)*. to appear.

Reiter, R. (1978). *On Closed World Data Bases*, pp. 119–140. Logic and Data Bases. Plenum Press.

# 21

# On Computers Diagnosing Computers

Moises Goldszmidt

## 1 Introduction

I came to UCLA in the fall of 1987 and immediately enrolled in the course titled "Probabilistic Reasoning in Intelligent Systems" where we, as a class, went over the draft of Judea's book of the same title [Pearl 1988]. The class meetings were fun and intense. Everybody came prepared, having read the draft of the appropriate chapter and having struggled through the list of homework exercises that were due that day. There was a high degree of discussion and participation, and I was very impressed by Judea's attentiveness and interest in our suggestions. He was fully engaged in these discussions and was ready to incorporate our comments and change the text accordingly. The following year, I was a teaching assistant (TA) for that class. The tasks involved with being a TA gave me a chance to rethink and really digest the contents of the course. It dawned on me then what a terrific insight Judea had to focus on formalizing the notion of conditional independence: All the "juice" he got in terms of making "reasoning under uncertainty" computationally effective came from that formalization. Shortly thereafter, I had a chance to chat with Judea about these and related thoughts. I was in need of formalizing a notion of "relevance" for my own research and thought that I could adapt some ideas from the graphoid models [Pearl 1988]. In that opportunity Judea shared another of his great insights with me. After hearing me out, Judea said one word: "causality". I don't remember the exact words he used to elaborate, but the gist of what he said to me was: "we as humans perform extraordinarily complex reasoning tasks, being able to select the relevant variables, circumscribe the appropriate context, and reduce the number of factors that we should manipulate. I believe that our intuitive notions of causality enable us to do so. Causality is the holly grail [for Artificial Intelligence]".

In this short note, I would like to pay tribute to Judea's scientific work by speculating on the very realistic possibility of computers using his formalization of causality for automatically performing a nontrivial reasoning task commonly reserved for humans. Namely designing, generating, and executing experiments in order to conduct a proper diagnosis and identify the causes of performance problems on code being executed in large clusters of computers. What follows in the next two sections is not a philosophical exposition on the meaning of "causality" or on the reasoning powers of automatons. It is rather a brief description of the current state of the art

in programming large clusters of computers and then, a brief account argumenting that the conditions are ripe for embarking on this research path.

## 2 Programming large clusters of computers made easy

There has been a recent research surge in systems directed at providing programmers with the ability to write efficient parallel and distributed applications [Hadoop 2008; Dean and Ghemawat 2004; Isard et al. 2007]. Programs written in these environments are automatically parallelized and executed on large clusters of commodity machines. The tasks of enabling programmers to effectively write and deploy parallel and distributed application has of course been a long-standing problem. Yet, the relatively recent emergence of large-scale internet services, which depend on clusters of hundreds of thousands of general purpose servers, have given the area a forceful push. Indeed, this is not merely an academic exercise; code written in these environments has been deployed and is very much in everyday use at companies such as Google, Microsoft, and Yahoo (and many others). These programs process web pages in order to feed the appropriate data to the search and news summarization engines; render maps for route planning services; and update usage and other statistics from these services. Year old figures estimate that Dryad, the specific such environment created at Microsoft [Isard et al. 2007], is used to crunch on the order of a petabyte a day at Microsoft. In addition, in our lab at Microsoft Research, a cluster of 256 machines controlled by Dryad runs daily at a 100% utilization. This cluster mostly runs tests and experiments on research algorithms in machine learning, privacy, and security that process very large amounts of data.

The intended model in Dryad is for the programmer to build code as if she were programming one computer. The system then takes care of a) distributing the code to the actual cluster and b) managing the execution of the code in the cluster. All aspects of execution, including data partition, communications, and fault tolerance, are the responsibility of Dryad.

With these new capabilities comes the need for new tools for debugging code, profiling execution performance, and diagnosing system faults. By the mere fact that clusters of large numbers of computers are being employed, rare bugs will manifest themselves more often, and devices will fail in more runs (due to both software and hardware problems). In addition, as the code will be executed in a networked environment and the data will be partitioned (usually according to some hash function), communication bandwidth, data location, contention for shared disks, and data skewness will impact the performance of the programs. Most of the times the impact of these factors will be hard to reproduce in a single machine, making it an imperative that the diagnosis, profiling, and debugging be performed in the same environment and conditions as those in which the code is running.

## 3 Computers diagnosing computers

The good news is that the same infrastructure that enables the programming and control of these clusters can be used for debugging and diagnosis. Normally the computation proceeds in stages where the different nodes in the cluster perform the same computation in parallel on different portions of the data. For purposes of fault tolerance, there are mechanisms in Dryad to monitor the execution time of each node at any computation stage. It is therefore possible to gather robust statistics about the expected execution time of any particular node at a given stage and identify especially slow nodes. Currently, this information is used to restart those nodes or to migrate the computation to other nodes.

We can take this further and collect the copious amount of data that is generated by the various built-in monitors looking at things such as cpu utilization, memory utilization, garbage collection, disk utilization, and statistics on I/O.[1] The statistical analysis of these signals may provide clues pointing at the probable causes of poor performance and even of failures. Indeed we have built a system called Artemis [Creţu-Ciocârlie et al. 2008], that takes advantage of the Dryad infrastructure to collect and preprocess the data from these signals in a distributed and opportunistic fashion. Once the data is gathered, Artemis will run a set of statistical and machine learning algorithms ranging from summarizations to regression and pattern classification. In this paper we propose one more step. We can imagine a system that guided with the information from these analyses, performs active experiments on the execution of the code. The objective will be to causally diagnose problems, and properly profile dependencies between the various factors affecting the performance of the computations.

Let us ground this idea in a realistic example. Suppose that through the analysis of the execution logs of some large task we identify that, on a computationally intensive stage, a small number of machines performed significantly worse that the average/median (in terms of overall processing speed). Through further analysis, for example logistic regression with L1 regularization, we are able to identify the factors that differentiate the slower machines. Thus, we narrow down the possibilities and determine that the main difference between these machines and the machines that performed well is the speed at which the data is read by the slower machines.[2] Further factors influencing this speed are whether the data resides on a local disk and whether there are other computational nodes that share that disk (and introduce contention), and on the speed of the network. Figure 1 shows a (simplified) causal model of this scenario depicting two processing nodes. The dark nodes represent factors/variables that can be controlled or where intervention is possible. Conducting controlled experiments guided by this graph would enable the

[1] The number of counters and other signals that these monitors yield can easily reach on the order of hundreds per machine.

[2] This particular case was encountered by the author while running a benchmark based on Terasort on a cluster with hundreds of machines [Creţu-Ciocârlie et al. 2008].

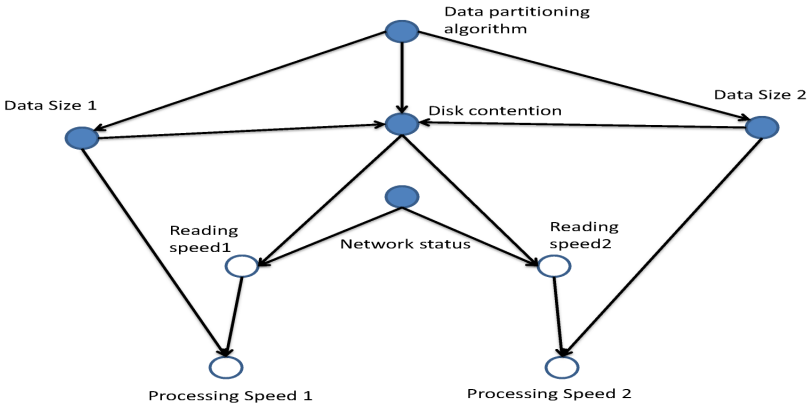



Figure 1. Simplified causal network depicting the processing speed scenario. Dark nodes represent the factor/variables than can be controlled or where intervention is possible.

precise characterization of the relationship between processing speed, data skewness, and disk contention, so that we can figure out how to partition and locate the data more efficiently and avoid having slow processing nodes. As the causal graph clearly exposes, controlling the data sizes for the computing nodes is not enough: if they reside on the same disk, contention may still cause slowdowns. This is obvious from the representation and algebra proposed by Judea in [Pearl 2000], as applied to this graph. This model also makes clear that intervening directly on the level of contention in the disk will indeed eliminate the dependency between the reading speed and the size of the data.

The idea of using graphical models for diagnosing computer systems goes back at least to [Breese and Heckerman 1996; Blake and Breese 1995]. It took close to 10 years after those papers for the first publication reporting the use of Bayesian networks for diagnosis in a nontrivial system in production to appear in a top tier systems conference [Cohen et al. 2004]. The methods in [Cohen et al. 2004] involve passive observation, and the authors make very clear that inferences concern correlation and not necessarily causation. However, hinting at root cause through correlation may not be enough in the very near future. Complexity and scale in current networked distributed systems keeps on increasing at a rapid pace. Because of service availability and reliability requirements, root cause analysis pointing at effective repair actions and accurate empirical characterization of dependencies between the different factors affecting computation are rapidly becoming a must.

Systems such as Dryad[Isard et al. 2007] enable the effective programming of large cluster of computers. In addition, they provide effective mechanisms for con-

trolling the "variables" of interest and setting up experiments in these clusters. Systems such as Artemis [Creţu-Ciocârlie et al. 2008] enable efficient collection and processing of extensive monitoring data, including the recording of the system state for recreating particular troublesome scenarios. The final ingredient for having machines automatically set up and conduct experiments is a language to describe these experiments and an algebra to reason about them in order to guarantee that the right variables are being controlled, and that we are intervening in the right spots in order to get to the correct conclusions. Through his seminal work in [Pearl 2000] and follow up papers, Judea Pearl has already given us that ingredient.

**Acknowledgments:** The author wishes to thank Mihai Budiu for numerous technical discussions on the topics of this paper, Joe Halpern for his help with the presentation, and very especially Judea Pearl for his continuous inspiration in the relentless and honest search for scientific truth.

## References

Blake, R. and J. Breese (1995). Automatic bottleneck detection. In *UAI'95: Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Breese, J. and D. Heckerman (1996). Decision theoretic troubleshooting. In *UAI'96: Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Cohen, I., M. Goldszmidt, T. Kelly, J. Symons, and J. Chase (2004). Correlating instrumentation data to systems states: A building block for automated diagnosis and control. In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*. USENIX Association.

Creţu-Ciocârlie, G. F., M. Budiu, and M. Goldszmidt (2008). Hunting for problems with Artemis. In *USENIX Workshop on the Analysis of System Logs (WASL)*.

Dean, J. and S. Ghemawat (2004). Mapreduce: simplified data processing on large clusters. In *OSDI'04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*. USENIX Association.

Hadoop (2008). The hadoop project. http://hadoop.apache.org.

Isard, M., M. Budiu, Y. Yu, A. Birrell, and D. Fetterly (2007). Dryad: distributed data-parallel programs from sequential building blocks. In *EuroSys '07: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007*. ACM.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press.

22

# Overthrowing the Tyranny of Null Hypotheses Hidden in Causal Diagrams

SANDER GREENLAND

## 1 Introduction

Graphical models have a long history before and outside of causal modeling. Mathematical graph theory extends back to the 1700s and was used for circuit analysis in the 19th century. Its application in probability and computer science dates back at least to the 1960s (Biggs et al., 1986), and by the 1980s graphical models had become fully developed tools for these fields (e.g., Pearl, 1988; Hajek et al., 1992; Lauritzen, 1996).

As *Bayesian networks*, graphical models are carriers of direct conditional independence judgments, and thus represent a collection of assumptions that confine prior support to a lower dimensional manifold of the space of prior distributions over the nodes. Such dimensionality reduction was recognized as essential in formulating explicit and computable algorithms for digital-machine inference, an essential task of artificial-intelligence (AI) research. By the 1990s, these models had been merged with causal path diagrams long used in observational health and social science (OHSS) (Wright, 1934; Duncan, 1975), resulting in a formal theory of causal diagrams (Spirtes et al., 1993; Pearl, 1995, 2000).

It should be no surprise that some of the most valuable and profound contributions to these *developments* were from Judea Pearl, a renowned AI theorist. He motivated causal diagrams as *causal* Bayesian networks (Pearl, 2000), in which the basis for the dimensionality reduction is grounded in judgments of causal independence (and especially, autonomy) rather than mere probabilistic independence. Beyond his extensive technical and philosophical contributions, Pearl fought steadfastly to roll back prejudice against causal modeling and causal graphs in statistics. Today, only a few statisticians still regard causality as a metaphysical notion to be banned from formal modeling (Lad, 1999). While a larger minority still reject some aspects of causal-diagram or potential-outcome theory (e.g., Dawid, 2000, 2008; Shafer, 2002), the spreading wake of applications display the practical value of these theories, and formal causal diagrams have advanced into applied journals and books (e.g., Greenland et al., 1999; Cole and Hernán, 2002; Hernán et al., 2002; Jewell, 2004; Morgan and Winship, 2007; Glymour and Greenland, 2008) – although their rapid acceptance in OHSS may well have been facilitated by the longstanding informal use of path diagrams to represent qualities of causal systems (e.g., Susser, 1973; Duncan, 1975).

Graphs are unsurpassed tools for illustrating certain mathematical results that hold in functional systems (whether stochastic or not, or causal or not). Nonetheless, it is essential to recognize that many if not most causal judgments in OHSS are based on

observational (purely associational) data, with little or nothing in the way of manipulative (or "surgical") experiment to test these judgments. Time order is usually known, which insures that the chosen arrow directions are correct; but rarely is there a sound basis for deleting an arrow, leaving autonomy in question. When all empirical constraints encoded by the causal network come from passive frequency observations rather than experiments, the primacy of causal independence judgments has to be questioned. In these situations (which characterize observational research), we should not neglect associational models (including graphs) that encode frequency-based judgments, for these models may be all that are identified by available data. Indeed, a deep philosophical commitment to statistically identified quantities seems to drive the arguments of certain critics of potential outcomes and causal diagrams (Dawid, 2000, 2008). Even if we reject this philosophy, however, we should retain the distinction between levels of identification provided by our data, for even experimental data will not identify everything we would like to know.

I will argue that, in some ways, the distinction of nonidentification from identification is as fundamental to modeling and statistical inference about causal effects as is the distinction of causation from association (Gustafson, 2005; Greenland, 2005a, 2009a, 2009b). Indeed, I believe that some of the controversy and confusion over causation versus association stems from the inability of statistical observations to point identify (consistently estimate) many of the causal parameters that astute scientists legitimately ask about. Furthermore, if we consider strategies that force identification from available data (such as node or arrow deletions from graphical models) we will find that identification may arise only by declaring some types of joint frequencies as justifying the corresponding conditional independence assumptions. This leads directly into the complex topic of pruning algorithms, including the choice of target or loss function.

I will outline these problems in their most basic forms, for I think that in the rush to adopt causal diagrams some realism has been lost by neglecting problems of nonidentification and pruning. My exposition will take the form of a series of vignettes that illustrate some basic points of concern. I will not address equally important concerns that many of the nodes offered as "treatments" may be ill-defined or nonmanipulable, or may correspond poorly to the treatments they ostensibly represent (Greenland, 2005b; Hernán, 2005; Cole and Frangakis, 2009; VanderWeele, 2009).

## 2 Nonidentification from Unfaithfulness in a Randomized Trial

Nonidentification can be seen and has caused controversy in the simplest causal-inference settings. Consider an experiment that randomizes a node R. Inferences on causal effects of R from subsequent associations of R with later events would then be justified, since R would be an exogenous node. R would also be an instrumental variable for certain descendants under further conditional-independence assumptions.

A key problem is how one could justify removing arrows along the line of descent from R to another node Y, even if R is exogenous. The overwhelmingly dominant approach licenses such removal if the observed R-Y association fails to meet some criterion for departure from pure randomness. This schematic for a causal-graph pruning

algorithm was employed by Spirtes et al. (1993), unfortunately with a very naïve Neyman-Pearsonian criterion (basically, allowing removal of arrows when a *P*-value exceeds an $\alpha$ level). These and related graphical algorithms (Pearl and Verma, 1991) produce what appear to be results in conflict with practical intuitions, namely causal "discovery" algorithms for single observational data sets, with no need for experimental evidence. These algorithms have been criticized philosophically on grounds related to the identification problem (Freedman and Humphreys, 1999; Robins and Wasserman, 1999ab), and there are also objections based on statistical theory (Robins et al., 2003).

One controversial assumption in these algorithms is *faithfulness* (or stability) that all connected nodes are associated. Although arguments have been put forward in its favor (e.g., Spirtes et al., 1993; Pearl, 2000, p. 63), this assumption coheres poorly with prior beliefs of some experienced researchers. Without faithfulness, two nodes may be independent even if there is an arrow linking them directly, if that arrow represents the presence of causal effects among units in a target population. A classic example of such unfaithfulness appeared in the debates between Fisher and Neyman in the 1930s, in which they disagreed on how to formulate the causal null hypothesis (Senn, 2004). The framework of their debate would be recognized today as the *potential-outcome* or counterfactual model, although in that era the model (when named) was called the randomization model. This model illustrates the benefit of randomization as a means of detecting a signal by injecting white noise into a system to drown out uncontrolled influences.

To describe the model, suppose we are to study the effect of a treatment X on an outcome $Y_{obs}$ observable on units in a specific target population. Suppose further we can fully randomize X, so X will equal the randomized node R. In the potential-outcome formulation, the outcome becomes a vector **Y** indexed by X. Specifically, X determines which component $Y_x$ of **Y** is observable conditional on X=x: $Y_{obs} = Y_x$ given X=x. To say X can causally affect a unit makes no reference to observation, however; it merely means that some components of **Y** are unequal. With a binary treatment and outcome, there are four types of units in the target population about a binary treatment X which indexes a binary potential-outcome vector **Y** (Copas, 1973):

1) Noncausal units with outcomes **Y**=(1,1) under X=1,0 ("doomed' to $Y_{obs}$=1);
2) Causal units with outcomes **Y**=(1,0) under X=1,0 (X=1 causes $Y_{obs}$=1);
3) Causal units with outcomes **Y**=(0,1) under X=1,0 (X=1 prevents $Y_{obs}$=1); and
4) Noncausal units with outcomes **Y**=(0,0) under X=1,0 ("immune" to $Y_{obs}$=1).

Suppose the proportion of type *i* in the trial population is $p_i$. There are now two null hypotheses:

$H_s$: There are no causal units: $p_2=p_3=0$ (sharp or strong null),

$H_w$: There is no net effect of treatment on the distribution of $Y_{obs}$: $p_2=p_3$ (weak null).

Under the randomization distribution we have

$E(Y_{obs}|X=1) = Pr(Y_{obs}=1|do[X=1]) = Pr(Y_1=1) = p_1+p_2$ and

$E(Y_{obs}|X=0) = Pr(Y_{obs}=1|do[X=0]) = Pr(Y_0=1) = p_1+p_3$;

hence $H_w$: $p_2=p_3$ is equivalent to the hypothesis that the expected outcome is the same for both treatment groups, and that the proportions with $Y_{obs}$=1 under the extreme population

intervention do[X=1] to every unit and do[X=0] to every unit are equal. Note however that only $H_s$ entails that the proportion with $Y_{obs}=1$ would be the same under *every* possible allocation of treatment X among the units; this property implies that the Y margin is fixed under $H_s$, and thus provides a direct causal rationale for Fisher's exact test of $H_s$ (Greenland, 1991).

$H_s$ also entails $H_w$ (or, in terms of parameter subspaces, $H_s \subset H_w$). The converse is false; but, under any of the "optimal" statistical tests that can be formulated from data on X and $Y_{obs}$ only, power is identical to the test size on all alternatives to the sharp null with $p_2=p_3$, i.e., $H_s$ is not identifiable within $H_w$, so within $H_w$ the power of any valid test of $H_s$ will not exceed its nominal alpha level. Thus, following Neyman, it is only relevant to think in terms of $H_w$, because $H_w$ could be rejected whenever $H_s$ could be rejected. Furthermore, some later authors would disallow $H_w - H_s$: $p_2 = p_3 \neq 0$ because it violates faithfulness (Spirtes et al., 2001) or because it represents an extreme treatment-by-unit interaction with no main effect (Senn, 2004).

There is also a Bayesian argument for focusing exclusively on $H_w$. $H_w$ is of Lebesgue measure zero, so under the randomization model, distinctions within $H_w$ can be ignored by inferences based on an absolutely continuous prior on $\mathbf{p} = (p_1,p_2,p_3)$ (Spirtes et al., 1993). More generally, any distinction that remains *a posteriori* can be traced to the prior. A more radical stance would dismiss both $H_s$ and the model defined by 1-4 above as "metaphysical," because it invokes constraints on the joint distribution of the components $Y_1$ and $Y_0$, and that joint distribution is not identified by randomization of X if only X and $Y_{obs}$ are observed (Dawid, 2000).

On the other hand, following Fisher one can argue that the null of key scientific and practical interest is $H_s$, and that $H_w - H_s$: $p_2 = p_3 \neq 0$ is a scientifically important and distinct hypothesis. For instance, $p_2>0$, $p_3>0$ entails the existence of units who should be treated quite differently, and provides an imperative to seek covariates that discriminate between the two causal types, even if $p_2=p_3$. Furthermore, rejection of the stronger $H_s$ is a *weaker* inference than rejection of the weaker $H_w$, and thus rejecting only $H_s$ would be a conservative interpretation of a "significant" test statistic. Thus, focusing on $H_s$ is compatible with a strictly falsificationist view of testing in which acceptance of the null is disallowed. Finally, there are real examples in which X=1 causes Y=1 in some units and causes Y=0 in others; in some of these cases there may be near-perfect balance of causation and prevention, as predicted by certain physical explanations for the observations (e.g., as in Neutra et al., 1980).

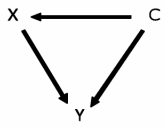
To summarize, identification problems arose in the earliest days of formal causal modeling, even when considering only the simplest of trials. Those problems pivoted not on whether one should attempt formal modeling of causation as distinct from association, but rather on what could be identified by standard experimental designs. In the face of limited (and limiting) design strategies, these problems initiated a long history of attempts to banish identification problems based on idealized inference systems and absolute philosophical assertions. But a counter-tradition of arguments, both practical and philosophical, has regarded identification problems as carriers of valuable scientific information: They are signs of study limitations which need to be recognized and can

only be dealt with effectively by innovative data collection (e.g., measuring more covariates or deploying new study designs), instead of by increasing sample sizes and defining the problems away so that "identical replications" are sufficient to narrow inferences.

## 3 Causal Diagrams Encode Numerous Uncertain Null Hypotheses

To move to the observational setting that is my main concern, consider figure 1, a typical causal diagram used to illustrate assumptions used by methods for estimating "the effect of X on Y" from observational data.

Figure 1: Naïve causal diagram




The first point to note is that this diagram is woefully incomplete relative to the epidemiologic reality, because it ignores

a) unmodeled confounders (variables not in the graph that affect more than one node in the graph),
b) selection effects (effects of factors in the graph on selection), and
c) measurement errors (which require addition of measurement nodes for each imperfectly measured node).

Put another way, typical causal DAGs like that in figure 1 are full of hidden, assumed null hypotheses, in the form of assumptions that imply problems a, b, and c are absent. For example, a causal DAG assumes that for **every** node pair (A,B) in the DAG,

1) there is **no** shared ancestor not in graph (not A↔B),
2) there is **no** unmarked conditioning event that has opened a path between A and B (not A─B),
3) if A and B are nonadjacent (neither A→B nor A←B), there is **no** mechanism that leads directly from one node to another (thus bypassing other nodes in the graph).

Not every study will seriously violate all of these assumptions. But in most studies in OHSS, none of the nulls 1-3 will have convincing support, and any purported test of a causal effect will really be a test of these 3 nulls as well as the specified causal null. This fact is just a special case of longstanding observations that statistical tests are really tests of all assumptions used in the test, not just the particular null of interest (Fisher, 1943; Box, 1980). In this regard, note that absence of arrows between nodes (3) encodes particularly strong nulls that are routinely presumed but rarely have supporting data. More often in OHSS, we observe only a conditional temporal sequence such as "A precedes B," which may be due to A→B, A↔B, A─B or some combination.

While sensitivity analysis is often recommended to examine the impact of deviations from assumptions, it becomes unintelligible if not infeasible as the number of assumptions (or corresponding parameters) increase. Then too, some causal inferences will display unlimited sensitivity to certain assumptions, requiring the introduction of priors on the corresponding parameters in order to salvage any inference (Greenland, 1998, 2005a; Gustafson, 2005). This problem arises in the model given below.

## 4 Eliminating Unsupported Nulls (graphical realism)

Let conditioning be denoted with square brackets around the conditioned event node. Then, in contrast to Figure 1, realistic causal graphs for OHSS will have

1) numerous unobserved (latent) nodes, often more of them than observed nodes,
2) few node pairs without an arc between them,
3) no **observed** set of variables sufficient for bias control, and
4) a selection node S that is bracketed and potentially affected by most other nodes.

In particular, when all variables are subject to measurement error, a realistic causal model for a single exposure-disease analysis will have at least:

X = Exposure, X*: measured X

Y = Outcome, Y*: measured Y

C = Known antecedents, C*: measured C

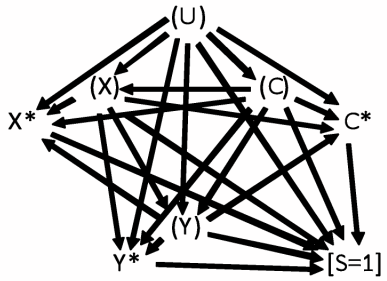
U = Other antecedents (unmeasured and possibly unknown)

S = Selection into the analysis (from selection into the study plus exclusions).

Because analysis is always conditioned on S=1, we should always show this conditioning event on the graph with a circle or brackets around it, e.g., as [S=1].

As an example, fig. 2 shows what I'd consider a **minimal** realistic causal graph for a typical case-control study of a life history and a degenerative disease outcome (e.g, nicotine intake X and Alzheimer's disease Y), which has 25 of 28 possible adjacencies.

**Figure 2: Realistic causal diagram**




What can fig. 2 provide if further assumptions cannot be justified? The only observed distribution is p(c*,x*,y*|S=1), which is not a factor in the causal Markov decomposition entailed by the graph,

$p(u,c,x,y,c^*,x^*,y^*,s) =$

$p(u)p(c|u)p(x|u,c)p(y|u,c,x)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(s|u,c,x,y,c^*,x^*,y^*)$,

which involves both S=0 events (not selected) and S=1 events (selected), i.e., the lowercase "s" is used when S can be either 0 or 1.

The marginal (total-population) potential-outcome distribution for Y after intervention on X, $p(y_x)$, equals $p(y|do[X=x])$, which under fig. 2 equals the standardized (mixing) distribution of Y given X standardized to (weighted by or mixed over) $p(c,u) = p(c|u)p(u)$:

$p(y_x) = p(y|do[x]) = \sum_{u,c} p(y|u,c,x)p(c|u)p(u)$.

This estimand involves only three factors in the decomposition, but none of them are identified if U is unobserved and no further assumptions are made. Analysis of the causal estimand $p(y_x)$ must somehow relate it to the observed distribution $p(c^*,x^*,y^*|S=1)$ using known or estimable quantities, or else remain purely speculative (i.e., a sensitivity analysis).

It is a long, hard road from $p(c^*,x^*,y^*|S=1)$ to $p(y_x)$, much longer than the current "causal inference" literature often makes it look. To appreciate the distance, rewrite the summand of the standardization formula for $p(y_x)$ as an inverse-probability-weighted (IPW) term derived from an observation $(c^*,x^*,y^*|S=1)$: From fig. 2,

$p(y|u,c,x)p(c|u)p(u) =$

$p(c^*,x^*,y^*|S=1)p(S=1)p(u,c,x,y|c^*,x^*,y^*,S=1)/$

$p(x|u,c)p(c^*|u,c,x,y)p(x^*|u,c,x,y)p(y^*|u,c,x,y)p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The latter expression includes

1) the exposure dependence on its parents, $p(x|u,c)$;
2) the measurement distributions $p(c^*|u,c,x,y)$, $p(x^*|u,c,x,y)$, $p(y^*|u,c,x,y)$; and
3) the fully conditioned selection probability $p(S=1|u,c,x,y,c^*,x^*,y^*)$.

The absence of effects corresponding to 1−3 from graphs offered as "causal" suggests that "causal inference" from observational data using formal causal models remains a theoretical and largely speculative exercise (albeit often presented without explicit acknowledgement of that fact).

When adjustments for these effects are attempted, we are usually forced to use crude empirical counterparts of terms like those in 1−3, with each substitution demanding nonidentified assumptions. Consider that, for valid inference under figure 2,

1) Propensity scoring and IPW for treatment need $p(x|u,c)$, but all we get from data is $p(x^*|c^*)$. Absence of u and c is usually glossed over by assuming "no unmeasured confounders" or "no residual confounding." These are not credible assumptions in OHSS.
2) IPW for selection and censoring needs $p(S=1|u,c,x,y,c^*,x^*,y^*)$, but usually the most we get from a cohort study or nested study is $p(S=1|c^*,x^*)$. We do not even get that much in a case-control study.
3) Measurement-error correction needs conditional distributions from $p(c^*,x^*,y^*,u,c,x,y|S=1)$, but even when a "validation" study is done, we obtain only alternative measurements $c^\dagger,x^\dagger,y^\dagger$ (which are rarely error-free) on a tiny and

biased subset. So we end up with observations from $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ where V is the validation indicator.

4) Consistency between the observed X and the intervention variable, in the sense t hat $P(Y|X=x) = P(Y|do[X=x],X=x)$. This can be hard to believe for common variables such as smoking, body-mass index, and blood pressure, even if do[X=x] is well-defined (which is not usually the case).

In the face of these realities, standard practice seems to be: Present wildly hypothetical analyses that pretend the observed distribution $p(c^*,x^*,y^*|S=1)$, perhaps along with $p(c^{\dagger},x^{\dagger},y^{\dagger},c^*,x^*,y^*|S=1,V=1)$ or $p(S=1|c^*,x^*)$, is sufficient for causal inference. The massive gaps are filled in with models or assumptions, which are priors that reduce dimensionality of the problem to something within computing bounds. For example, use of IPW with $p(S=1|c^*,x^*)$ to adjust for selection bias (as when 1−S is a censoring indicator) depends crucially on a nonidentified ignorability assumption that $S \perp\!\!\!\perp (U,C,X,Y)|(C^*,X^*)$, i.e., that selection S is independent of the latent variables U,C,X,Y given the observed variables C*,X*. We should expect this condition to be violated whenever a latent variable affects selection directly or shares unobserved causes with selection. If such effects are exist but are missing from the analysis graph, then by some definitions the graph (and hence the resulting analysis) isn't causal, no matter how much propensity scoring (PS), marginal structural modeling (MSM), inverse-probability weighting (IPW), or other causal-modeling procedures we apply to the observations $(c^*,x^*,y^*|S=1)$.

Of course, the overwhelming dimensionality of typical OHSS problems virtually guarantees that arbitrary constraints will enter at some point, and forces even the best scientists to rely on a tiny subset of all the models or explanations consistent with available facts. Personal bias in determining this subset may be unavoidable due to strong cultural influences (such as adherence to received theories, as well as moral strictures and financial incentives), which can also lead to biased censoring of observations (Greenland, 2009c). One means of coping with such bias is by being aware of it, then trying to test it against the facts one can muster (which are often few).

The remaining sections sketch some alternatives to pretending we can identify unbiased or assuredly valid estimators of causal effects in observational data, as opposed to within hypothetical models for data generation (Greenland, 1990; Robins, 2001). In these approaches, both frequentist and Bayesian analyses are viewed as hypotheticals conditioned on a data-generation model of unknown validity. Frequentist analysis provides only inferences of the form "if the data-generation process behaves like this, here is how the proposed decision rule would perform," while Bayesian analysis provides only inferences of the form "if I knew that its data-generation process behaves like this, here is how this study would alter my bets."[1] If we aren't sure how the data-generation

[1]This statement describes Bayes factors (Good, 1983) conditioned on the model. That model may include an unknown parameter that indexes a finite number of submodels scattered over some high-dimensional subspace, in which case the Bayesian analysis is called "model averaging," usually with an implicit uniform prior over the models. Model averaging may also operate over continuous parameters via priors on those parameters.

process behaves, no statistical analysis can provide more, no matter how much causal modeling is done.

## 5 Predictive Analysis

If current models for observed-data generators (whether logistic, structural, or propensity models) can't be taken seriously as "causal", what can we make of their outputs? It is hard to believe the usual excuses offered for regression outputs (e.g., that they are "descriptive") when the fitted model is asserted to be causal or "structural." Are we to consider the outputs of (say) and IPW-fitted MSM to be some sort of data summary? Or will it function as some kind of optimal predictor of outcomes in a purely predictive context? No serious case has been made for causal models in either role, and it seems that some important technical improvements are needed before causal modeling methods become credible predictive tools.

Nonetheless, graphical models remain useful (and might be less misleading) even when they are not "causal," serving instead as mere carriers of conditional independence assumptions within a time-ordered framework. In this usage, one may still employ presumed causal independencies as prior judgments for specification. In particular, for predictive purposes, some or all of the arrows in the graph may retain informal causal interpretations; but they may be causally wrong, and yet the graph can still be correct for predictive purposes.

In this regard, most of the graphical modeling literature in statistics imposes little in the way of causal burden on the graph, as when graphs are used as influence diagrams, belief and information networks, and so on without formal causal interpretation (that is, without representing a formal causal model, e.g., Pearl, 1988; Hajek et al., 1992; Cox and Wermuth, 1996; Lauritzen, 1996). DAG rules remain valid for prediction if the absence of an open path from X to Y is interpreted as entailing $X \perp\!\!\!\perp Y$, or equivalently if the absence of a directed path from X to Y (in causal terms, X is not a cause of Y; equivalently, Y is not affected by X) is interpreted as entailing $X \perp\!\!\!\perp Y|\mathbf{pa}_X$, the noncausal Markov condition (where $\mathbf{pa}_X$ is the set of parents of X). In that case, $X \rightarrow Y$ can be used in the graph even if X has no effect on Y, or vice-versa.

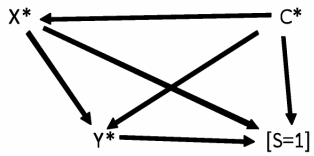
As an example, suppose X and Y are never observed without them affecting selection S, as when X is affects miscarriage S and Y is congenital malformation. If the target population is births, X predicts malformations Y among births (which have S=1). As another example, suppose X and Y are never observed without an uncontrolled, ungraphed confounder U, as when X is diet and Y is health status. If one wishes to target those at high risk for screening or actuarial purposes it does not matter if $X \rightarrow Y$ represents a causally confounded relation. Lack of a directed path from X to Y now corresponds to lack of additional predictive value for Y from X given $\mathbf{pa}_X$. Arrow directions in temporal (time-ordered) predictive graphs correspond to point priors about time order, just as they do in causal graphs.

Of course, if misinterpreted as causal, predictive inferences from graphs (or any predictive modeling) may be potentially disastrous for judging interventions on X. But, in OHSS, the causality represented by a directed path in a so-called causal diagram rarely

corresponds to more than a hypothesis, plausible perhaps but only one among a myriad of others. If most arrows shown in a graph encode no real data other than an observed conditional temporal sequencing, then labeling the graph as a "causal diagram" sets the stage for the disaster.

Figure 3 is the temporal predictive diagram for the observables in the earlier example, assuming those events occur in the order C*, X*, Y*, [S=1].

**Figure 3: Temporally predictive diagram**




Comparison to the causal diagram in figure 2 illustrates how a temporal predictive diagram for an observable frequency distribution may be derived from an underlying causal diagram for a nonidentified theory. Figure 3 is saturated in the sense that all nodes are connected by an edge, but this need not be so for a predictive diagram derived from a causal one. If there is temporal ambiguity among the observables, there may be multiple predictive diagrams compatible with the causal diagram (which will form a subset of the multiple probability graphs compatible with the causal diagram).

If we treat causal models as carriers of prior information about conditional independencies, they appear as legitimate candidates to consider as predictive models. For example, MSMs can be evaluated as devices for prediction from fixed sequences and structural nested models can be evaluated as devices for prediction from stochastic processes. I would thus offer this challenge to the current "longitudinal causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, are there predictive advantages of structural modeling (modeling potential outcomes as well as observed outcomes)? If not, what precisely is the advantage of fitting such models (compared to noncausal models) when effects are not identified?

I believe there *are* advantages of causal models, precisely as described by Pearl (2000): They provide an encoding for qualitative (structural) prior information expressed in terms of "cause" and "effect." But in current practice, fitting methods for complex causal models are quite primitive, and need to incorporate properly smoothness and other information that can be freely assumed in purely predictive-modeling approaches. This is a general problem of semi-parametric theory: It necessarily focuses sharp constraints in some dimensions and none in "most" dimensions (represented by the infinite-dimensional time component in standard Cox models). When relevant dimensions for constraint (those

where much background information is available) are not well represented by the dimensions constrained by the model, considerably efficiency can be lost for estimating parameters of interest. A striking example given by Whittemore and Keller (1986) displayed the poor small-sample performance for estimating a survival curve when using an unsmoothed nonparametric hazard estimator (Kaplan-Meier or Nelson-Altschuler estimation), relative to spline smoothing of the hazard.

## 6 Pruning the Identified Portion of the Model

Over recent decades, great strides have been made in creating predictive algorithms; the question remains however, what role should these algorithms play in causal inference? It would seem that these algorithms can be beneficially applied to fitting the marginal distribution identified by the observations. Nonetheless, the targets of causal inference in observational studies lie beyond the identified margin, and thus beyond the reach of these algorithms. At best, then, the algorithms can provide the identified foundation for building into unobserved dimensions of the phenomena under study.

Even if we focus only on the identified margin, however, there may be far more nodes and edges than seem practical to allow in the final model. A prominent feature of modern predictive algorithms is that they start with an impractically large number of terms and then aggressively prune the model, and may re-grow and re-prune repeatedly (Hastie et al., 2009). These strategies coincide with the intuition that omitting a term is justified when its contribution is too small to stand out against bias and background noise; e.g., we do not include variables like patient identification number because we know that are usually pure noise.

Nonetheless, automated algorithms often delete variables or connections that prior information instead suggests are relevant or related; thus shields from pruning are often warranted. Furthermore, a deleted node or arrow may indeed be important from a contextual perspective even if does not meet algorithmic retention criteria. Thus, model simplification strategies such as pruning may be justified by a need for dimensionality reduction, but should be recognized as part of algorithmic compression or computational prediction, not as a mode of inference about structural models.

Apart from these vague cautions, it has long been recognized that if our goal is to evaluate causal effects, different loss functions are needed from those in the pruning algorithms commonly applied by researchers. Specifically, the loss or benefit entailed by pruning needs to be evaluated in reference to the target effect under study, and not simply successful prediction of identified quantities. Operationalizing this imperative requires building out into the nonidentified (latent) realm of the target effects, which is the focus of *bias modeling*.

## 7 Modeling Latent Causal Structures (Bias Modeling)

The target effects in causal inference are functions of unobserved dimensions of the data-generating process, which consist primarily of bias sources (Greenland, 2005a). Once we recognize the nonidentification this structure entails, the major analysis task shifts away

from mathematical statistics to prior specification, because with nonidentification only proper priors on nonidentified parameters can lead to proper posteriors.

Even the simplest point-exposure case can involve complexities that transform simple and precise-looking conventional results into complex and utterly ambiguous posteriors (Greenland, 2009a, 2009b). In a model complex enough to reflect Figure 2, there are far too many elements of specification to contextually justify them all in detail. For example, one could only rarely justify fewer than two free structural parameters per arrow, and the distributional form for each parameter prior would call for at least two hyperparameters per parameter (e.g., a mean and a variance), leading to at least 50 parameters and 100 hyperparameters in a graph with 25 arrows. Allowing but one prior association parameter (e.g., a correlation) per parameter pair adds over 1,000 (50 choose 2) more hyperparameters.

As a consequence of the exponential complexity of realistic models, prior specification is difficult, ugly, ad hoc, highly subjective, and tentative in the extreme. In addition, the hard-won model will lack generalizability and elegance, making it distasteful to both the applied scientist and the theoretical statistician. Nor will it please the applied statistician concerned with "data analysis," for the analysis will instead revolve around numerous contextual judgments that enlist diverse external sources of information. In contrast to the experimental setting (in which the data-generation model may be dictated entirely by the design), the usually sharp distinction between prior and data information will be blurred by the dependence of the data-generation model on external information.

These facts raise another challenge to the current "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify predictive links among observed quantities, how can we incorporate simultaneously all error sources (systematic as well as random) known to be important into a complex longitudinal framework involving mismeasurement of entire sequences of exposures and confounders? Some progress on this front has been made, but primarily in contexts with validation data available (Cole et al., 2010), which is not the usual case.

## 8 The Descriptive Alternative

In the face of the extraordinary complexity of realistic models for OHSS, it should be an option of each study to focus on describing the study and its data thoroughly, sparing us attempts at inference about nonidentified quantities such as "causal effects." This option will likely never be popular, but should be allowed and even encouraged (Greenland et al., 2004). After all, why should I care about your causal inferences, especially if they are based on or grossly over-weighted by the one or few studies that you happened to be involved in? If I am interested in forming my own inferences, I do want to see your data and get an accurate narrative of the physical processes that produced them. In this regard, statistics may supply data summaries. Nonetheless, it must be made clear exactly how the statistics offered reflect the data as opposed to some hypothesis about the population from which they came; *P*-values do not satisfy this requirement (Greenland, 1993; Poole, 2001).

Here then is a final challenge to the "causal modeling" literature: If we know our observations are just a dim and distant projection of the causal structure and we can only identify associations among observed quantities, how can we interpret the outputs of "structural modeling" (such as confidence limits for ostensibly causal estimands which are not in fact identified) as data summaries? We should want to see answers that are sensible when the targets are effects in a context at least as complex as in fig. 2.

## 9 What is a Causal Diagram?

The above considerations call into question some epidemiological accounts of causal diagrams. Pearl (2000) describes a causal model *M* as a formal functional system giving relations among a set of variables. *M* defines a joint probability distribution p() and an intervention operator do[] on the variables. A causal diagram is then a directed graph *G* that implies the usual Markov decomposition for p() and displays additional properties relating p() and do[]. In particular, each child-parent family {X, $\mathbf{pa}_X$} in *G* satisfies

1) p(x|do[$\mathbf{pa}_X$=a]) = p(x|$\mathbf{pa}_X$=a), and
2) if Z is not in {X, $\mathbf{pa}_X$}, p(x|do[Z=z],$\mathbf{pa}_X$=a) = p(x|$\mathbf{pa}_X$=a).

(e.g., see Pearl, 2000, p. 24). These properties stake out *G* as an illustration (mapping) of structure within *M*.

Condition 1 is often described as stating that the association of each node X with its parent vector $\mathbf{pa}_X$ is unconfounded given *M*. Condition 2 says that, given *M*, the only variables in *G* that affect a node X are its parents, and is often called the causal Markov condition (CMC). Nonetheless, as seems to happen often as time passes and methods become widely adopted, details have gotten lost. In the more applied literature, causal diagrams have come to be described as "unconfounded graphs" without reference to an underlying causal model (e.g., Hernán et al., 2004; VanderWeele and Robins, 2007; Glymour and Greenland, 2008). This description not only misses the CMC (2) but, taken literally, means that all shared causes are in the graph.

Condition 1 is a property relating two mathematical objects, *G* and *M*. To claim a diagram is unconfounded is to instead make a claim about the relation of *G* the real world, thus inviting confusion between a *model* for causal processes and the actual processes. For many experts in OHSS, the claim of unconfoundedness has zero probability of being correct because of its highly restrictive empirical content (e.g., see Robins and Wasserman, 1999ab). At best, we can only hope that the diagram provides a useful computing aid for predicting the outcomes of intervention strategies.

As with regression models, causal models in OHSS are always false. Because we can never know we have a correct model (and in fact in OHSS we can't even know if we are very close), to say *G* is causal if unconfounded is a scientifically vacuous definition: It is saying the graph is causal if the causal model it represents is correct. This is akin to saying a monotone increasing function from the range of X to [0,1] is not a probability distribution if it is not in fact how X is distributed; thus a normal($\mu,\sigma^2$) cumulative function wouldn't be a probability distribution unless it is *the* actual probability distribution for X (whether that distribution is an objective event generator or a subjective betting schedule).

So, to repeat: To describe a causal diagram as an "unconfounded graph" blurs the distinction between models and reality. Model-based deductions are logical conditionals of the form "model *M* deductively yields these conclusions," and have complete certainty *given* the model *M*. But the model, and hence reality, is never known with certainty, and in OHSS cannot be claimed as known except in the most crude fashion. The point is brought home above by appreciating just how unrealistic all causal models and diagrams in OHSS must be. Thus I would encourage the description of causal diagrams as graphical causal models (or more precisely, graphical representations of certain equivalence classes of causal models), rather than as "unconfounded graphs" (or similar phrases). This usage might even be acceptable to some critics of the current causal-modeling literature (Dawid, 2008).

## 10 Summary and Conclusions

I would be among the last to deny the utility of causal diagrams; but I argue that their practical utility in OHSS is limited to (i) compact and visually immediate representation of assumptions, and (ii) illustration of sources of nonidentification and bias given realistic assumptions. Converse claims about their utility for identification seem only the latest in a long line of promises to "solve" the problem of causal inference. These promises are akin to claims of preventing and curing all cancers; while progress is possible, the enormous complexity of real systems should leave us skeptical about claims of "solutions" to the real problem.

Many authors have recognized that the problem of effect identification is unsolvable in principle. Although this logical impossibility led some to deny the scientific merit of causal thinking, it has not prevented development of useful tools that have causal-modeling components. Nonetheless, the most precision we can realistically hope for estimating effects in OHSS is about one-digit accuracy, and in many problems even that seems too optimistic. Thus some practical sense is needed to determine what is and isn't important to include as model components. Yet, despite the crudeness of OHSS, good sense seems to lead almost inevitably to including more components than can be identified by available data.

My main point is that effect identification (in the frequentist sense of identification by the observed data) should be abandoned as a primary goal in causal modeling in OHSS. My reasons are practical: Identification will often demand dropping too much of importance from the model, thus imposing null hypotheses that have no justification in either past frequency observations or in priors about mechanisms generating the observations, thus leading to overconfident and biased inferences. In particular, defining a graph as "causal" if it is unconfounded assumes a possibly large set of causal null hypotheses (at least two for every pair of nodes in the graph: no shared causes or conditioned descendants not in the graph). In OHSS, the only graphs that satisfy such a definition will need many latent nodes to be "causal" in this sense, and as a consequence will reveal the nonidentified nature of target effects. Inference may then proceed by imposing contextually defensible priors or penalties (Greenland, 2005a, 2009a, 2009b, 2010).

Despite my view and similar ones (e.g., Gustafson, 2005), I suspect the bulk of causal-inference statistics will trundle on relying exclusively on artificially identified models. It will thus be particularly important to remember that just because a method is labeled a "causal modeling" method does not mean it gives us estimates and tests of actual causal effects. For those who find identification too hard to abandon in formal analysis, the only honest recourse is to separate identified and nonidentified components of the model, focus technique on the identified portion, and leave the latent residual as a topic for sensitivity analysis, speculative modeling, and further study. In this task, graphs can be used without the burden of causality if we allow them a role as pure prediction tools, and they can also be used as causal diagrams of the largely latent structure that generates the data.

**Acknowledgments:** I am most grateful to Tyler VanderWeele, Jay Kaufman, and Onyebuchi Arah for their extensive and useful comments on this chapter.

## References

Biggs, N., Lloyd, E. and Wilson, R. (1986). *Graph Theory, 1736-1936*. Oxford University Press.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society*, *Series A* **143**, 383–430.

Cole S.R. and M.A. Hernán (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31**, 163–165.

Cole, S.R. and C.E. Frangakis (2009). The consistency assumption in causal inference: a definition or an assumption? *Epidemiology* **20**, 3–5.

Cole, S.R., L.P. Jacobson, P.C. Tien, L. Kingsley, J.S. Chmiel and K. Anastos (2010). Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death. *American Journal of Epidemiology* **171**, 113-122.

Copas, J.G. (1973). Randomization models for matched and unmatched 2x2 tables. *Biometrika* **60**, 267-276.

Cox, D.R. and N. Wermuth. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*. Boca Raton, FL: CRC/Chapman and Hall.

Dawid, A.P. (2000). Causal inference without counterfactuals (with discussion). *Journal of the American Statistical Association* **95**, 407-448.

Dawid, A.P. (2008). Beware of the DAG! In: *NIPS 2008 Workshop Causality: Objectives and Assessment*. JMLR Workshop and Conference Proceedings.

Duncan, O.D. (1975). *Introduction to Structural Equation Models*. New York: Academic Press.

Fisher, R.A. (1943; reprinted 2003). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association* **38**, 103–104. Reprinted in the *International Journal of Epidemiology* **32**, 692.

Freedman, D.A. and Humphreys, P. (1999). Are there algorithms that discover causal structure? *Synthese* **121**, 29–54.

Glymour, M.M. and S. Greenland (2008). Causal diagrams. Ch. 12 in: Rothman, K.J., S. Greenland and T.L. Lash, eds. *Modern Epidemiology*, 3rd ed. Philadelphia: Lippincott.

Good, I.J. (1983). *Good thinking*. Minneapolis: U. Minnesota Press.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology* **1**, 421-429.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two-contingency tables. *The American Statistician* **45**, 248-251.

Greenland, S. (1993). Summarization, smoothing, and inference. *Scandinavian Journal of Social Medicine* **21**, 227-232.

Greenland, S. (1998). The sensitivity of a sensitivity analysis. In: 1997 Proceedings of the Biometrics Section. Alexandria, VA: American Statistical Association, 19-21.

Greenland, S. (2005a). Epidemiologic measures and policy formulation: Lessons from potential outcomes (with discussion). *Emerging Themes in Epidemiology* (online journal) 2:1–4. (Originally published as "Causality theory for policy uses of epidemiologic measures," Chapter 6.2 in: Murray, C.J.L., J.A. Salomon, C.D. Mathers and A.D. Lopez, eds. (2002) *Summary Measures of Population Health*. Cambridge, MA: Harvard University Press/WHO, 291-302.)

Greenland, S. (2005b). Multiple-bias modeling for analysis of observational data (with discussion). *Journal of the Royal Statistical Society, Series A* **168**, 267–308.

Greenland, S. (2009a). Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. *International Journal of Epidemiology* **38**, 1662–1673.

Greenland, S. (2009b). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science* **24**, 195-210.

Greenland, S. (2009c). Dealing with uncertainty about investigator bias: disclosure is informative. *Journal of Epidemiology and Community Health* **63**, 593-598.

Greenland, S. (2010). The need for syncretism in applied statistics (comment on "The future of indirect evidence" by Bradley Efron). *Statistical Science* **25**, in press.

Greenland, S., J. Pearl, and J.M. Robins (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37-48.

Greenland, S., M. Gago-Dominguez, and J.E. Castellao (2004). The value of risk-factor ("black-box") epidemiology (with discussion). *Epidemiology* **15**, 519-535.

Gustafson, P. (2005). On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). *Statistical Science* **20**, 111-140.

Hajek, P., T. Havranek and R. Jirousek (1992). *Uncertain Information Processing in Expert Systems*. Boca Raton, FL: CRC Press.

Hastie, T., R. Tibshirani and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. New York: Springer.

Hernán, M.A. (2005). Hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology* **162**, 618–620.

Hernán M.A., S. Hernandez-Diaz, M.M. Werler and A.A. Mitchell. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology* **155**, 176–184.

Hernán M.A., S. Hernandez-Diaz and J.M. Robins (2004). A structural approach to selection bias. *Epidemiology* **15**, 615-625.

Jewell, N. (2004). *Statistics for Epidemiology*. Boca Raton, FL: Chapman and Hall/CRC.

Lad, F. (1999). Assessing the foundations of Bayesian networks: A challenge to the principles and the practice. *Soft Computing* **3**, 174-180.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Clarendon Press.

Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Morgan, S.L. and C. Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Neutra, R.R., S. Greenland, and E.A. Friedman (1980). The effect of fetal monitoring on cesarean section rates. *Obstetrics and Gynecology* **55**, 175-180.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika* **82**, 669-710.

Pearl, J. (2000; 2nd ed. 2009). *Causality*. New York: Cambridge University Press.

Pearl, J. and P. Verma (1991). A theory of inferred causation. In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, Ed. J.A. Allen, R. Filkes and E. Sandewall. San Francisco: Morgan Kaufmann, 441-452.

Poole, C. (2001). Poole C. Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology* **12**, 291–294.

Robins, J.M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12**, 313–320.

Robins, J.M. and L. Wasserman (1999a). On the impossibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 305-321.

Robins, J.M. and L. Wasserman (1999b). Rejoinder to Glymour and Spirtes. In: *Computation, Causation, and Discovery*. Glymour, C. and Cooper, G., eds. Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press, pp. 333-342.

Robins, J.M., R. Scheines, P. Spirtes and L. Wasserman (2003). Uniform consistency in causal inference. *Biometrika* **90**, 491-515.

Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* **23**, 3729–3753.

Shafer, G. (2002). Comment on "Estimating causal effects," by George Maldonado and Sander Greenland. *International Journal of Epidemiology* **31**, 434-435.

Spirtes, P., C. Glymour and R. Scheines (1993; 2nd ed. 2001). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Susser, M. (1973). *Causal Thinking in the Health Sciences*. New York: Oxford University Press.

VanderWeele, T.J. (2009). Concerning the consistency assumption in causal inference. *Epidemiology* **20**, 880-883.

VanderWeele, T.J. and J.M. Robins (2007). Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology* **166**, 1096-1104.

Whittemore, A.S. and J.B. Keller (1986). Survival estimation using splines. *Biometrics* **42**, 495-506.

Wright, S., (1934). The method of path coefficients. *Annals of Mathematical Statistics* **5**,161-215.

# 23

# Actual Causation and the Art of Modeling

Joseph Y. Halpern and Christopher Hitchcock

## 1 Introduction

In *The Graduate*, Benjamin Braddock (Dustin Hoffman) is told that the future can be summed up in one word: "Plastics". One of us (Halpern) recalls that in roughly 1990, Judea Pearl told him that the future was in causality. Pearl's own research was largely focused on causality in the years after that; his seminal contributions are widely known. We were among the many influenced by his work. We discuss one aspect of it, *actual causation*, in this article, although a number of our comments apply to causal modeling more generally.

Pearl introduced a novel account of actual causation in Chapter 10 of *Causality*, which was later revised in collaboration with one of us [Halpern and Pearl 2005]. In some ways, Pearl's approach to actual causation can be seen as a contribution to the philosophical project of trying to analyze actual causation in terms of counterfactuals, a project associated most strongly with David Lewis [1973a]. But Pearl's account was novel in at least two important ways. The first was his use of *structural equations* as a tool for modeling causality. In the philosophical literature, causal structures were often represented using so-called *neuron diagrams*, but these are not (and were never intended to be) all-purpose representational tools. (See [Hitchcock 2007b] for a detailed discussion of the limitations of neuron diagrams.) We believe that the lack of a more adequate representational tool had been a serious obstacle to progress. Second, while the philosophical literature on causality has focused almost exclusively on actual causality, for Pearl, actual causation was a rather specialized topic within the study of causation, peripheral to many issues involving causal reasoning and inference. Thus, Pearl's work placed the study of actual causation within a much broader context.

The use of structural equations as a model for causal relationships was well known long before Pearl came on the scene; it seems to go back to the work of Sewall Wright in the 1920s (see [Goldberger 1972] for a discussion). However, the details of the framework that have proved so influential are due to Pearl. Besides the Halpern-Pearl approach mentioned above, there have been a number of other closely-related approaches for using structural equations to model actual causation; see, for example, [Glymour and Wimberly 2007; Hall 2007; Hitchcock 2001; Hitchcock 2007a; Woodward 2003]. The goal of this paper is to look more carefully at the modeling of causality using structural equations. For definiteness, we use the

Halpern-Pearl (HP) version [Halpern and Pearl 2005] here, but our comments apply equally well to the other variants.

It is clear that the structural equations can have a major impact on the conclusions we draw about causality—it is the equations that allow us to conclude that lower air pressure is the cause of the lower barometer reading, and not the other way around; increasing the barometer reading will not result in higher air pressure. The structural equations express the effects of *interventions*: what happens to the bottle if it is hit with a hammer; what happens to a patient if she is treated with a high dose of the drug, and so on. These effects are, in principle, objective; the structural equations can be viewed as describing objective features of the world. However, as pointed out by Halpern and Pearl [2005] and reiterated by others [Hall 2007; Hitchcock 2001; Hitchcock 2007a], the choice of variables and their values can also have a significant impact on causality. Moreover, these choices are, to some extent, subjective. This, in turn, means that judgments of actual causation are subjective.

Our view of actual causation being at least partly subjective stands in contrast to the prevailing view in the philosophy literature, where the assumption is that the job of the philosopher is to analyze the (objective) notion of causation, rather like that of a chemist analyzing the structure of a molecule. This may stem, at least in part, from failing to appreciate one of Pearl's lessons: actual causality is only part of the bigger picture of causality. There can be an element of subjectivity in ascriptions of actual causality without causation itself being completely subjective. In any case, the experimental evidence certainly suggests that people's views of causality are subjective, even when there is no disagreement about the relevant structural equations. For example, a number of experiments show that broadly normative considerations, including the subject's own moral beliefs, affect causal judgment. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Knobe and Fraser 2008].) Even in relatively non-controversial cases, people may want to focus on different aspects of a problem, and thus give different answers to questions about causality. For example, suppose that we ask for the cause of a serious traffic accident. A traffic engineer might say that the bad road design was the cause; an educator might focus on poor driver's education; a sociologist might point to the pub near the highway where the driver got drunk; a psychologist might say that the cause is the driver's recent breakup with his girlfriend.[1] Each of these answers is reasonable. By appropriately choosing the variables, the structural equations framework can accommodate them all.

Note that we said above "by appropriately choosing the variables". An obvious question is "What counts as an appropriate choice?". More generally, what makes a model an appropriate model? While we do want to allow for subjectivity, we need

[1] This is a variant of an example originally due to Hanson [1958].

to be able to justify the modeling choices made. A lawyer in court trying to argue that faulty brakes were the cause of the accident needs to be able to justify his model; similarly, his opponent will need to understand what counts as a legitimate attack on the model. In this paper we discuss what we believe are reasonable bases for such justifications. Issues such as model stability and interactions between the events corresponding to variables turn out to be important.

Another focus of the paper is the use of defaults in causal reasoning. As we hinted above, the basic structural equations model does not seem to suffice to completely capture all aspects of causal reasoning. To explain why, we need to briefly outline how actual causality is defined in the structural equations framework. Like many other definitions of causality (see, for example, [Hume 1739; Lewis 1973b]), the HP definition is based on counterfactual dependence. Roughly speaking, $A$ is a cause of $B$ if, had $A$ not happened (this is the counterfactual condition, since $A$ did in fact happen) then $B$ would not have happened. As is well known, this naive definition does not capture all the subtleties involved with causality. Consider the following example (due to Hall [2004]): Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had Suzy not thrown. Thus, according to the naive counterfactual definition, Suzy's throw is not a cause of the bottle shattering. This certainly seems counterintuitive.

The HP definition deals with this problem by taking $A$ to be a cause of $B$ if $B$ counterfactually depends on $A$ *under some contingency*. For example, Suzy's throw is the cause of the bottle shattering because the bottle shattering counterfactually depends on Suzy's throw, under the contingency that Billy doesn't throw. (As we will see below, there are further subtleties in the definition that guarantee that, if things are modeled appropriately, Billy's throw is not also a cause.)

While the definition of actual causation in terms of structural equations has been successful at dealing with many of the problems of causality, examples of Hall [2007], Hiddleston [2005], and Hitchcock [2007a] show that it gives inappropriate answers in cases that have structural equations isomorphic to ones where it arguably gives the appropriate answer. This means that, no matter how we define actual causality in the structural-equations framework, the definition must involve more than just the structural equations. Recently, Hall [2007], Halpern [2008], and Hitchcock [2007a] have suggested that using defaults might be a way of dealing with the problem. As the psychologists Kahneman and Miller [1986, p. 143] observe, "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it". This intuition is also present in the legal literature. Hart and Honoré [1985] observe that the statement "It was the presence of oxygen that caused the fire" makes sense only if there were reasons to view the presence of oxygen as abnormal.

As shown by Halpern [2008], we can model this intuition formally by combining a well-known approach to modeling defaults and normality, due to Kraus, Lehmann,

and Magidor [1990] with the structural-equation model. Moreover, doing this leads to a straightforward solution to the problem above. The idea is that, when showing that if $A$ hadn't happened then $B$ would not have happened, we consider only contingencies that are at least as normal as the actual world. For example, if someone typically leaves work at 5:30 PM and arrives home at 6, but, due to unusually bad traffic, arrives home at 6:10, the bad traffic is typically viewed as the cause of his being late, not the fact that he left at 5:30 (rather than 5:20).

But once we add defaults to the model, the problem of justifying the model becomes even more acute. We not only have to justify the structural equations and the choice of variables, but also the default theory. The problem is exacerbated by the fact that default and "normality" have a number of interpretations. Among other things, they can represent moral obligations, societal conventions, prototypicality information, and statistical information. All of these interpretations are relevant to understanding causality; this makes justifying default choices somewhat subtle.

The rest of this paper is organized as follows. In Sections 2 and 3, we review the notion of causal model and the HP definition of actual cause; most of this material is taken from [Halpern and Pearl 2005]. In Section 4, we discuss some issues involved in the choice of variables in a model. In Section 5, we review the approach of [Halpern 2008] for adding considerations of normality to the HP framework, and discuss some modeling issues that arise when we do so. We conclude in Section 6.

## 2 Causal Models

In this section, we briefly review the HP definition of causality. The description of causal models given here is taken from [Halpern 2008], which in turn is based on that of [Halpern and Pearl 2005].

The HP approach assumes that the world is described in terms of random variables and their values. For example, if we are trying to determine whether a forest fire was caused by lightning or an arsonist, we can take the world to be described by three random variables:

- $F$ for forest fire, where $F = 1$ if there is a forest fire and $F = 0$ otherwise;
- $L$ for lightning, where $L = 1$ if lightning occurred and $L = 0$ otherwise;
- $ML$ for match (dropped by arsonist), where $ML = 1$ if the arsonist drops a lit match, and $ML = 0$ otherwise.

Some random variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. For example, to model the fact that if either a match is lit or lightning strikes, then a fire starts, we could use the random variables $ML$, $F$, and $L$ as above, with the equation $F = \max(L, ML)$. (Alternately, if a fire requires both causes to be present, the equation for $F$ becomes $F = \min(L, ML)$.) The equality sign in this equation should be thought of more like an assignment statement in programming languages; once we set the values of $F$

and $L$, then the value of $F$ is set to their maximum. However, despite the equality, if a forest fire starts some other way, that does not force the value of either *ML* or $L$ to be 1.

It is conceptually useful to split the random variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. For example, in the forest-fire example, the variables *ML*, $L$, and $F$ are endogenous. However, we want to take as given that there is enough oxygen for the fire and that the wood is sufficiently dry to burn. In addition, we do not want to concern ourselves with the factors that make the arsonist drop the match or the factors that cause lightning. These factors are all determined by the exogenous variables.

Formally, a *causal model* $M$ is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathcal{F}$ defines a set of *modifiable structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for $Y$ (that is, the set of values over which $Y$ *ranges*). $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted $F_X$ such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that $F_X$ determines the value of $X$, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. If there is one exogenous variable $U$ and three endogenous variables, $X$, $Y$, and $Z$, then $F_X$ defines the values of $X$ in terms of the values of $Y$, $Z$, and $U$. For example, we might have $F_X(u, y, z) = u + y$, which is usually written as $X \leftarrow U + Y$.[2] Thus, if $Y = 3$ and $U = 2$, then $X = 5$, regardless of how $Z$ is set.

In the running forest-fire example, suppose that we have an exogenous random variable $U$ that determines the values of $L$ and *ML*. Thus, $U$ has four possible values of the form $(i, j)$, where both of $i$ and $j$ are either 0 or 1. The $i$ value determines the value of $L$ and the $j$ value determines the value of *ML*. Although $F_L$ gets as arguments the vale of $U$, *ML*, and $F$, in fact, it depends only on the (first component of) the value of $U$; that is, $F_L((i, j), m, f) = i$. Similarly, $F_{ML}((i, j), l, f) = j$. The value of $F$ depends only on the value of $L$ and *ML*. *How* it depends on them depends on whether either cause by itself is sufficient for the forest fire or whether both are necessary. If either one suffices, then $F_F((i, j), l, m) = \max(l, m)$, or, perhaps more comprehensibly, $F = \max(L, ML)$; if both are needed, then $F = \min(L, ML)$. For future reference, call the former model the *disjunctive* model, and the latter the *conjunctive* model.

The key role of the structural equations is to define what happens in the presence of external interventions. For example, we can explain what happens if the arsonist

[2]The fact that $X$ is assigned $U + Y$ (i.e., the value of $X$ is the sum of the values of $U$ and $Y$) does not imply that $Y$ is assigned $X - U$; that is, $F_Y(U, X, Z) = X - U$ does not necessarily hold.

does *not* drop the match. In the disjunctive model, there is a forest fire exactly if there is lightning; in the conjunctive model, there is definitely no fire. Setting the value of some variable $X$ to $x$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model denoted $M_{X \leftarrow x}$. In the new causal model, the equation for $X$ is very simple: $X$ is just set to $x$; the remaining equations are unchanged. More formally, $M_{X \leftarrow x} = (\mathcal{S}, \mathcal{F}^{X \leftarrow x})$, where $\mathcal{F}^{X \leftarrow x}$ is the result of replacing the equation for $X$ in $\mathcal{F}$ by $X = x$.

The structural equations describe *objective* information about the results of interventions, that can, in principle, be checked. Once the modeler has selected a set of variables to include in the model, *the world* determines which equations among those variables correctly represent the effects of interventions.[3] By contrast, the *choice* of variables is subjective; in general, there need be no objectively "right" set of exogenous and endogenous variables to use in modeling a problem. We return to this issue in Section 4.

It may seem somewhat circular to use causal models, which clearly already encode causal information, to define actual causation. Nevertheless, as we shall see, there is no circularity. The equations of a causal model do not represent relations of *actual causation*, the very concept that we are using them to define. Rather, the equations characterize the results of *all possible* interventions (or at any rate, all of the interventions that can be represented in the model) without regard to what actually happened. Specifically, the equations do not depend upon the actual values realized by the variables. For example, the equation $F = \max(L, ML)$, by itself, does not say anything about whether the forest fire was actually caused by lightning or by an arsonist, or, for that matter, whether a fire even occurred. By contrast, relations of actual causation depend crucially on how things actually play out.

A sequence of endogenous $X_1, \ldots, X_n$ of is a *directed path* from $X_1$ to $X_n$ if the value of $X_{i+1}$ (as given by $F_{X_{i+1}}$) depends on the value of $X_i$, for $1 = 1, \ldots, n-1$. In this paper, following HP, we restrict our discussion to *acyclic* causal models, where causal influence can be represented by an acyclic Bayesian network. That is, there is no cycle $X_1, \ldots, X_n, X_1$ of endogenous variables that forms a directed path from $X_1$ to itself. If $M$ is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, there is a unique solution for all the equations.

[3] In general, there may be uncertainty about the causal model, as well as about the true setting of the exogenous variables in a causal model. Thus, we may be uncertain about whether smoking causes cancer (this represents uncertainty about the causal model) and uncertain about whether a particular patient actually smoked (this is uncertainty about the value of the exogenous variable that determines whether the patient smokes). This uncertainty can be described by putting a probability on causal models and on the values of the exogenous variables. We can then talk about the probability that $A$ is a cause of $B$.

# 3 The HP Definition of Actual Cause

## 3.1 A language for describing causes

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where $\phi$ is a Boolean combination of primitive events, $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\phi$. The special case where $k = 0$ is abbreviated as $\phi$. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ says that $\phi$ would hold if $Y_i$ were set to $y_i$, for $i = 1, \ldots, k$.

A causal formula $\psi$ is true or false in a causal model, given a context. As usual, we write $(M, \vec{u}) \models \psi$ if the causal formula $\psi$ is true in causal model $M$ given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable $X$ has value $x$ in the unique (since we are dealing with acyclic models) solution to the equations in $M$ in context $\vec{u}$ (that is, the unique vector of values for the endogenous variables that simultaneously satisfies all equations in $M$ with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\phi$ if $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \phi$. We write $M \models \phi$ if $(M, \vec{u}) \models \phi$ for all contexts $\vec{u}$.

For example, if $M$ is the disjunctive causal model for the forest fire, and $u$ is the context where there is lightning and the arsonist drops the lit match, then $(M, u) \models [ML \leftarrow 0](F = 1)$, since even if the arsonist is somehow prevented from dropping the match, the forest burns (thanks to the lightning); similarly, $(M, u) \models [L \leftarrow 0](F = 1)$. However, $(M, u) \models [L \leftarrow 0; ML \leftarrow 0](F = 0)$: if the arsonist does not drop the lit match and the lightning does not strike, then the forest does not burn.

## 3.2 A preliminary definition of causality

The HP definition of causality, like many others, is based on counterfactuals. The idea is that if $A$ and $B$ both occur, then $A$ is a cause of $B$ if, if $A$ hadn't occurred, then $B$ would not have occurred. This idea goes back to at least Hume [1748, Section VIII], who said:

> We may define a cause to be an object followed by another, ..., where, if the first object had not been, the second never had existed.

This is essentially the *but-for* test, perhaps the most widely used test of actual causation in tort adjudication. The but-for test states that an act is a cause of injury if and only if, but for the act (i.e., had the the act not occurred), the injury would not have occurred.

There are two well-known problems with this definition. The first can be seen by considering the disjunctive causal model for the forest fire again. Suppose that the arsonist drops a match and lightning strikes. Which is the cause? According to a naive interpretation of the counterfactual definition, neither is. If the match hadn't dropped, then the lightning would still have struck, so there would have been

a forest fire anyway. Similarly, if the lightning had not occurred, there still would have been a forest fire. As we shall see, the HP definition declares both lightning and the arsonist causes of the fire. (In general, there may be more than one actual cause of an outcome.)

A more subtle problem is what philosophers have called *preemption*, which is illustrated by the rock-throwing example from the introduction. As we observed, according to a naive counterfactual definition of causality, Suzy's throw would not be a cause.

The HP definition deals with the first problem by defining causality as counterfactual dependency *under certain contingencies*. In the forest-fire example, the forest fire does counterfactually depend on the lightning under the contingency that the arsonist does not drop the match; similarly, the forest fire depends counterfactually on the dropping of the match under the contingency that the lightning does not strike.

Unfortunately, we cannot use this simple solution to treat the case of preemption. We do not want to make Billy's throw the cause of the bottle shattering by considering the contingency that Suzy does not throw. So if our account is to yield the correct verdict in this case, it will be necessary to limit the contingencies that can be considered. The reason that we consider Suzy's throw to be the cause and Billy's throw not to be the cause is that Suzy's rock hit the bottle, while Billy's did not. Somehow the definition of actual cause must capture this obvious intuition.

With this background, we now give the preliminary version of the HP definition of causality. Although the definition is labeled "preliminary", it is quite close to the final definition, which is given in Section 5. The definition is relative to a causal model (and a context); $A$ may be a cause of $B$ in one causal model but not in another. The definition consists of three clauses. The first and third are quite simple; all the work is going on in the second clause.

The types of events that the HP definition allows as actual causes are ones of the form $X_1 = x_1 \wedge \ldots \wedge X_k = x_k$—that is, conjunctions of primitive events; this is often abbreviated as $\vec{X} = \vec{x}$. The events that can be caused are arbitrary Boolean combinations of primitive events. The definition does not allow statements of the form "$A$ or $A'$ is a cause of $B$", although this could be treated as being equivalent to "either $A$ is a cause of $B$ or $A'$ is a cause of $B$". On the other hand, statements such as "$A$ is a cause of $B$ or $B'$" are allowed; this is not equivalent to "either $A$ is a cause of $B$ or $A$ is a cause of $B'$".

DEFINITION 1. (Actual cause; preliminary version) [Halpern and Pearl 2005] $\vec{X} = \vec{x}$ is an *actual cause of* $\phi$ *in* $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \phi$.

AC2. There is a partition of $\mathcal{V}$ (the set of endogenous variables) into two subsets $\vec{Z}$ and $\vec{W}$ with $\vec{X} \subseteq \vec{Z}$, and a setting $\vec{x}'$ and $\vec{w}$ of the variables in $\vec{X}$ and $\vec{W}$,

respectively, such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\phi$.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\phi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$, where we abuse notation and write $\vec{W}' \leftarrow \vec{w}$ to denote the assignment where the variables in $\vec{W}'$ get the same values as they would in the assignment $\vec{W} \leftarrow \vec{w}$.

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2.

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of $\phi$ unless both $\vec{X} = \vec{x}$ and $\phi$ actually happen. AC3 is a minimality condition, which ensures that only those elements of the conjunction $\vec{X} = \vec{x}$ that are essential for changing $\phi$ in AC2(a) are considered part of a cause; inessential elements are pruned. Without AC3, if dropping a lit match qualified as a cause of the forest fire, then dropping a match and sneezing would also pass the tests of AC1 and AC2. AC3 serves here to strip "sneezing" and other irrelevant, over-specific details from the cause. Clearly, all the "action" in the definition occurs in AC2. We can think of the variables in $\vec{Z}$ as making up the "causal path" from $\vec{X}$ to $\phi$, consisting of one or more directed paths from variables in $\vec{X}$ to variables in $\phi$. Intuitively, changing the value(s) of some variable(s) in $\vec{X}$ results in changing the value(s) of some variable(s) in $\vec{Z}$, which results in the value(s) of some other variable(s) in $\vec{Z}$ being changed, which finally results in the truth value of $\phi$ changing. The remaining endogenous variables, the ones in $\vec{W}$, are off to the side, so to speak, but may still have an indirect effect on what happens. AC2(a) is essentially the standard counterfactual definition of causality, but with a twist. If we want to show that $\vec{X} = \vec{x}$ is a cause of $\phi$, we must show (in part) that if $\vec{X}$ had a different value, then $\phi$ would have been false. However, this effect of the value of $\vec{X}$ on the truth value of $\phi$ may not hold in the actual context; the value of $\vec{W}$ may have to be different to allow this effect to manifest itself. For example, consider the context where both the lightning strikes and the arsonist drops a match in the disjunctive model of the forest fire. Stopping the arsonist from dropping the match will not prevent the forest fire. The counterfactual effect of the arsonist on the forest fire manifests itself only in a situation where the lightning does not strike (i.e., where $L$ is set to 0). AC2(a) is what allows us to call both the lightning and the arsonist causes of the forest fire. Essentially, it ensures that $\vec{X}$ alone suffices to bring about the change from $\phi$ to $\neg\phi$; setting $\vec{W}$ to $\vec{w}$ merely eliminates possibly spurious side effects that may mask the effect of changing the value of $\vec{X}$. Moreover, when $\vec{X} = \vec{x}$, although the values of variables on the causal path (i.e., the variables $\vec{Z}$) may be perturbed by the change to $\vec{W}$, this perturbation has no impact on the value of $\phi$. If $(M, \vec{u}) \models \vec{Z} = \vec{z}^*$, then $z^*$ is the value of the variable $Z$ in the context $\vec{u}$. We capture the fact that the perturbation has no impact on the value of $\phi$ by saying that if some variables $Z$ on

the causal path were set to their original values in the context $\vec{u}$, $\phi$ would still be true, as long as $\vec{X} = \vec{x}$.

EXAMPLE 2. For the forest-fire example, let $M$ be the disjunctive model for the forest fire sketched earlier, with endogenous variables $L$, $ML$, and $F$. We want to show that $L = 1$ is an actual cause of $F = 1$. Clearly $(M, (1, 1)) \models F = 1$ and $(M, (1, 1)) \models L = 1$; in the context (1,1), the lightning strikes and the forest burns down. Thus, AC1 is satisfied. AC3 is trivially satisfied, since $\vec{X}$ consists of only one element, $L$, so must be minimal. For AC2, take $\vec{Z} = \{L, F\}$ and take $\vec{W} = \{ML\}$, let $x' = 0$, and let $w = 0$. Clearly, $(M, (1, 1)) \models [L \leftarrow 0, ML \leftarrow 0](F \neq 1)$; if the lightning does not strike and the match is not dropped, the forest does not burn down, so AC2(a) is satisfied. To see the effect of the lightning, we must consider the contingency where the match is not dropped; the definition allows us to do that by setting $ML$ to 0. (Note that here setting $L$ and $ML$ to 0 overrides the effects of $U$; this is critical.) Moreover, $(M, (1, 1)) \models [L \leftarrow 1, ML \leftarrow 0](F = 1)$; if the lightning strikes, then the forest burns down even if the lit match is not dropped, so AC2(b) is satisfied. (Note that since $\vec{Z} = \{L, F\}$, the only subsets of $\vec{Z} - \vec{X}$ are the empty set and the singleton set consisting of just $F$.)

It is also straightforward to show that the lightning and the dropped match are also causes of the forest fire in the context where $U = (1, 1)$ in the conjunctive model. Again, AC1 and AC3 are trivially satisfied and, again, to show that AC2 holds in the case of lightning we can take $\vec{Z} = \{L, F\}$, $\vec{W} = \{ML\}$, and $x' = 0$, but now we let $w = 1$. In the conjunctive scenario, if there is no lightning, there is no forest fire, while if there is lightning (and the match is dropped) there is a forest fire, so AC2(a) and AC2(b) are satisfied; similarly for the dropped match.

EXAMPLE 3. Now consider the Suzy-Billy example.[4] We get the desired result—that Suzy's throw is a cause, but Billy's is not—but only if we model the story appropriately. Consider first a coarse causal model, with three endogenous variables:

- $ST$ for "Suzy throws", with values 0 (Suzy does not throw) and 1 (she does);
- $BT$ for "Billy throws", with values 0 (he doesn't) and 1 (he does);
- $BS$ for "bottle shatters", with values 0 (it doesn't shatter) and 1 (it does).

(We omit the exogenous variable here; it determines whether Billy and Suzy throw.) Take the formula for $BS$ to be such that the bottle shatters if either Billy or Suzy throw; that is $BS = \max(BT, ST)$. (We assume that Suzy and Billy will not miss if they throw.) $BT$ and $ST$ play symmetric roles in this model; there is nothing to distinguish them. Not surprisingly, both Billy's throw and Suzy's throw are classified as causes of the bottle shattering in this model. The argument is essentially identical to that in the disjunctive model of the forest-fire example in

[4]The discussion of this example is taken almost verbatim from HP.

the context $U = (1, 1)$, where both the lightning and the dropped match are causes of the fire.

The trouble with this model is that it cannot distinguish the case where both rocks hit the bottle simultaneously (in which case it would be reasonable to say that both $ST = 1$ and $BT = 1$ are causes of $BS = 1$) from the case where Suzy's rock hits first. To allow the model to express this distinction, we add two new variables to the model:

- $BH$ for "Billy's rock hits the (intact) bottle", with values 0 (it doesn't) and 1 (it does); and
- $SH$ for "Suzy's rock hits the bottle", again with values 0 and 1.

Now our equations will include:

- $SH = ST$;
- $BH = \min(BT, 1 - SH)$; and
- $BS = \max(SH, BH)$.

Now it is the case that, in the context where both Billy and Suzy throw, $ST = 1$ is a cause of $BS = 1$, but $BT = 1$ is not. To see that $ST = 1$ is a cause, note that, as usual, it is immediate that AC1 and AC3 hold. For AC2, choose $\vec{Z} = \{ST, SH, BH, BS\}$, $\vec{W} = \{BT\}$, and $w = 0$. When $BT$ is set to 0, $BS$ tracks $ST$: if Suzy throws, the bottle shatters and if she doesn't throw, the bottle does not shatter. To see that $BT = 1$ is *not* a cause of $BS = 1$, we must check that there is no partition $\vec{Z} \cup \vec{W}$ of the endogenous variables that satisfies AC2. Attempting the symmetric choice with $\vec{Z} = \{BT, BH, SH, BS\}$, $\vec{W} = \{ST\}$, and $w = 0$ violates AC2(b). To see this, take $\vec{Z}' = \{BH\}$. In the context where Suzy and Billy both throw, $BH = 0$. If $BH$ is set to 0, the bottle does not shatter if Billy throws and Suzy does not. It is precisely because, in this context, Suzy's throw hits the bottle and Billy's does not that we declare Suzy's throw to be the cause of the bottle shattering. AC2(b) captures that intuition by allowing us to consider the contingency where $BH = 0$, despite the fact that Billy throws. We leave it to the reader to check that no other partition of the endogenous variables satisfies AC2 either.

This example emphasizes an important moral. If we want to argue in a case of preemption that $X = x$ is the cause of $\phi$ rather than $Y = y$, then there must be a random variable ($BH$ in this case) that takes on different values depending on whether $X = x$ or $Y = y$ is the actual cause. If the model does not contain such a variable, then it will not be possible to determine which one is in fact the cause. This is certainly consistent with intuition and the way we present evidence. If we want to argue (say, in a court of law) that it was $A$'s shot that killed $C$ rather than $B$'s, then we present evidence such as the bullet entering $C$ from the left side (rather

than the right side, which is how it would have entered had $B$'s shot been the lethal one). The side from which the shot entered is the relevant random variable in this case. Note that the random variable may involve temporal evidence (if $Y$'s shot had been the lethal one, the death would have occurred a few seconds later), but it certainly does not have to.

## 4 The Choice of Variables

A modeler has considerable leeway in choosing which variables to include in a model. Nature does not provide a uniquely correct set of variables. Nonetheless, there are a number of considerations that guide variable selection. While these will not usually suffice to single out one choice of variables, they can provide a framework for the rational evaluation of models, including resources for motivating and defending certain choices of variables, and criticizing others.

The problem of choosing a set of variables for inclusion in a model has many dimensions. One set of issues concerns the question of how many variables to include in a model. If the modeler begins with a set of variables, how can she know whether she should add additional variables to the model? Given that it is always possible to add additional variables, is there a point at which the model contains "enough" variables? Is it ever possible for a model to have "too many" variables? Can the addition of further variables ever do positive harm to a model?

Another set of issues concerns the values of variables. Say that variable $X'$ is a *refinement* of $X$ if, for each value $x$ in the range of $X$, there is some subset $S$ of the range of $X'$ such that $X = x$ just in case $X'$ is in $S$. When is it appropriate or desirable to replace a variable with a refinement? Can it ever lead to problems if a variable is too fine-grained? Similarly, are there considerations that would lead us to prefer a model that replaced $X$ with a new variable $X''$, whose range is a proper subset or superset of the range of $X$?

Finally, are there constraints on the set of variables in a model over and above those we might impose on individual variables? For instance, can the choice to include a particular variable $X$ within a model require us to include another variable $Y$, or to exclude a particular variable $Z$?

While we cannot provide complete answers to all of these questions, we believe a good deal can be said to reduce the arbitrariness of the choice of variables. The most plausible way to motivate guidelines for the selection of variables is to show how inappropriate choices give rise to systems of equations that are inaccurate, misleading, or incomplete in their predictions of observations and interventions. In the next three subsections, we present several examples to show how such considerations can be brought to bear on the problem of variable choice.

### 4.1 The Number of Variables

We already saw in Example 3 that it is important to choose the variables correctly. Adding more variables can clearly affect whether $A$ is a cause of $B$. When is it

appropriate or necessary to add further variables to a model?[5] Suppose that we have an infinite sequence of models $M^1, M^2, \ldots$ such that the variables in $M^i$ are $X_0, \ldots, X_{i+1}, Y$, and $M^{i+1}_{X_{i+1} \leftarrow 1} = M_i$ (so that $M^{i+1}$ can be viewed as an extension of $M^i$). Is it possible that whether $X_0 = 1$ is a cause of $Y = 1$ can alternate as we go through this sequence? This would indicate a certain "instability" in the causality. In this circumstance, a lawyer should certainly be able to argue against using, say, $M^7$ as a model to show that $X_0 = 1$ is cause of $Y = 1$. On the other hand, if the sequence stabilizes, that is, if there is some $k$ such that for all $i \geq k$, $M^i$ delivers the same verdict on some causal claim of interest, that would provide a strong reason to accept $M^k$ as sufficient.

Compare Example 2 with Example 3. In Example 2, we were able to adequately model the scenario using only three endogenous variables: $L$, $ML$, and $F$. By contrast, in Example 3, the model containing only three endogenous variables, $BT$, $ST$, and $BS$, was inadequate. What is the difference between the two scenarios? One difference we have already mentioned is that there seems to be an important feature of the second scenario that cannot be captured in the three-variable model: Suzy's rock hit the bottle before Billy's did. There is also a significant "topological" difference between the two scenarios. In the forest-fire example, there are two directed paths into the variable $F$. We could interpolate additional variables along these two paths. We could, for instance, interpolate a variable representing the occurrence of a small brush fire. But doing so would not fundamentally change the causal structure: there would still be just two directed paths into $F$. In the case of preemption, however, adding the additional variables $SH$ and $BH$ created an additional directed path that was not there before. The three-variable model contained just two directed paths: one from $ST$ to $BS$, and one from $BT$ to $BS$. However, once the variables $SH$ and $BH$ were added, there were three directed paths: $\{ST, SH, BS\}$, $\{BT, BH, BS\}$, and $\{ST, SH, BH, BS\}$. The intuition, then, is that adding additional variables to a model will not affect the relations of actual causation that hold in the model unless the addition of those variables changes the "topology" of the model. A more complete mathematical characterization of the conditions under which the verdicts of actual causality remain stable under the addition of further variables strikes us as a worthwhile research project that has not yet been undertaken.

### 4.2 The Ranges of Variables

Not surprisingly, the set of possible values of a variable must also be chosen appropriately. Consider, for example, a case of "trumping", introduced by Schaffer [2000]. Suppose that a group of soldiers is very well trained, so that they will obey any order given by a superior officer; in the case of conflicting orders, they obey the

[5] Although his model of causality is quite different from ours, Spohn [2003] also considers the effect of adding or removing variables, and discusses how a model with fewer variables should be related to one with more variables.

highest-ranking officer. Both a sergeant and a major issue the order to march, and the soldiers march. Let us put aside the morals that Schaffer attempts to draw from this example (with which we disagree; see [Halpern and Pearl 2005] and [Hitchcock 2010]), and consider only the modeling problem. We will presumably want variables $S$, $M$, and $A$, corresponding to the sergeant's order, the major's order, and the soldiers' action. We might let $S = 1$ represent the sergeant's giving the order to march and $S = 0$ represent the sergeant's giving no order; likewise for $M$ and $A$. But this would not be adequate. If the only possible order is the order to march, then there is no way to capture the principle that in the case of conflicting orders, the soldiers obey the major. One way to do this is to replace the variables $M$, $S$, and $A$ by variables $M'$, $S'$ and $A'$ that take on three possible values. Like $M$, $M' = 0$ if the major gives no order and $M' = 1$ if the major gives the order to march. But now we allow $M' = 2$, which corresponds to the major giving some other order. $S'$ and $A'$ are defined similarly. We can now write an equation to capture the fact that if $M' = 1$ and $S' = 2$, then the soldiers march, while if $M' = 2$ and $S' = 1$, then the soldiers do not march.

The appropriate set of values of a variable will depend on the other variables in the picture, and the relationship between them. Suppose, for example, that a hapless homeowner comes home from a trip to find that his front door is stuck. If he pushes on it with a normal force then the door will not open. However, if he leans his shoulder against it and gives a solid push, then the door will open. To model this, it suffices to have a variable $O$ with values either 0 or 1, depending on whether the door opens, and a variable $P$, with values 0 or 1 depending on whether or not the homeowner gives a solid push.

On the other hand, suppose that the homeowner also forgot to disarm the security system, and that the system is very sensitive, so that it will be tripped by any push on the door, regardless of whether the door opens. Let $A = 1$ if the alarm goes off, $A = 0$ otherwise. Now if we try to model the situation with the same variable $P$, we will not be able to express the dependence of the alarm on the homeowner's push. To deal with both $O$ and $A$, we need to extend $P$ to a 3-valued variable $P'$, with values 0 if the homeowner does not push the door, 1 if he pushes it with normal force, and 2 if he gives it a solid push.

These considerations parallel issues that arise in philosophical discussions about the metaphysics of "events".[6] Suppose that our homeowner pushed on the door with enough force to open it. Is there just one event, the push, that can be described at various levels of detail, such as a "push" or a "hard push"? This is the view of Davidson [1967]. Or are there rather many different events corresponding to these different descriptions, as argued by Kim [1973] and Lewis [1986b]? And if we take the latter view, which of the many events that occur should be counted as causes of the door's opening? These strike us as pseudoproblems. We believe that questions

[6]This philosophical usage of the word "event" is different from the typical usage of the word in computer science and probability, where an event is just a subset of the state space.

about causality are best addressed by dealing with the methodological problem of constructing a model that correctly describes the effects of interventions in a way that is not misleading or ambiguous.

A slightly different way in which one variable may constrain the values that another may take is by its implicit presuppositions. For example, a counterfactual theory of causation seems to have the somewhat counterintuitive consequence that one's birth is a cause of one's death. This sounds a little odd. If Jones dies suddenly one night, shortly before his 80th birthday, the coroner's inquest is unlikely to list "birth" as among the causes of his death. Typically, when we investigate the causes of death, we are interested in what makes the difference between a person's dying and his surviving. So our model might include a variable $D$ such $D = 1$ holds if Jones dies shortly before his 80th birthday, and $D = 0$ holds if he continues to live. If our model also includes a variable $B$, taking the value 1 if Jones is born, 0 otherwise, then there simply is no value that $D$ would take if $B = 0$. Both $D = 0$ and $D = 1$ implicitly presuppose that Jones was born (i.e., $B = 1$). Our conclusion is that if we have chosen to include a variable such as $D$ in our model, then we cannot conclude that Jones' birth is a cause of his death!

### 4.3 Dependence and Independence

Lewis [1986a] added a constraint to his counterfactual theory of causation. In order for event $c$ to be a cause of event $e$, the two events cannot be logically related. Suppose for instance, that Martha says "hello" loudly. If she had not said "hello", then she certainly could not have said "hello" loudly. But her saying "hello" is not a cause of her saying "hello" loudly. The counterfactual dependence results from a logical, rather than a causal, relationship between the two events.

We must impose a similar constraint upon causal models. Values of different variables should not correspond to events that are logically related. But now, rather than being an *ad hoc* restriction, it has a clear rationale. For suppose that we had a model with variable $H_1$ and $H_2$, where $H_1$ represents "Martha says 'hello'" (i.e., $H_1 = 1$ if Martha says "hello" and $H_1 = 0$ otherwise), and $H_2$ represents "Martha says 'hello' loudly". The intervention $H_1 = 0 \wedge H_2 = 1$ is meaningless; it is logically impossible for Martha not to say "hello" and to say ''hello" loudly.

We doubt that any careful modeler would choose variables that have logically related values. However, the converse of this principle, that the different values of any particular variable *should* be logically related (in fact, mutually exclusive), is less obvious and equally important. Consider Example 3. While, in the actual context, Billy's rock will hit the bottle just in case Suzy's doesn't, this is not a necessary relationship. Suppose that, instead of using two variables $SH$ and $BH$, we try to model the scenario with a variable $H$ that takes the value 1 if Suzy's rock hits, and and 0 if Billy's rock hits. The reader can verify that, in this model, there is no contingency such that the bottle's shattering depends upon Suzy's throw. The problem, as we said, is that $H = 0$ and $H = 1$ are *not* mutually exclusive; there are

possible situations in which both rocks hit or neither rock hits the bottle. In particular, this representation does not allow us to consider independent interventions on the rocks hitting the bottle. As the discussion in Example 3 shows, it is precisely such an intervention that is needed to establish that Suzy's throw (and not Billy's) is the actual cause of the bottle shattering.

While these rules are simple in principle, their application is not always transparent.

EXAMPLE 4. Consider cases of "switching", which have been much discussed in the philosophical literature. A train is heading toward the station. An engineer throws a switch, directing the train down the left track, rather than the right track. The tracks re-converge before the station, and the train arrives as scheduled. Was throwing the switch a cause of the train's arrival? HP consider two causal models of this scenario. In the first, there is a random variable $S$ which is 1 if the switch is thrown (so the train goes down the left track) and 0 otherwise. In the second, in addition to $S$, there are variables $LT$ and $RT$, indicating whether or not the train goes down the left track and right track, respectively. Note that with the first representation, there is no way to model the train not making it to the arrival point. With the second representation, we have the problem that $LT = 1$ and $RT = 1$ are arguably not independent; the train cannot be on both tracks at once. If we want to model the possibility of one track or another being blocked, we should use, instead of $LT$ and $RT$, variables $LB$ and $RB$, which indicate whether the left track or right track, respectively, are blocked. This allows us to represent all the relevant possibilities without running into independence problems. Note that if we have only $S$ as a random variable, then $S = 1$ cannot be a cause of the train arriving; it would have arrived no matter what. With $RB$ in the picture, the preliminary HP definition of actual cause rules that $S = 1$ can be an actual cause of the train's arrival; for example, under the contingency that $RB = 1$, the train does not arrive if $S = 0$. (However, once we extend the definition to include defaults, as we will in the next section, it becomes possible once again to block this conclusion.)

These rules will have particular consequences for how we should represent events that might occur at different times. Consider the following simplification of an example introduced by Bennett [1987], and also considered in HP.

EXAMPLE 5. Suppose that the Careless Camper (CC for short) has plans to go camping on the first weekend in June. He will go camping unless there is a fire in the forest in May. If he goes camping, he will leave a campfire unattended, and there will be a forest fire. Let the variable $C$ take the value 1 if CC goes camping, and 0 otherwise. How should we represent the state of the forest?

There appear to be at least three alternatives. The simplest proposal would be to use a variable $F$ that takes the value 1 if there is a forest fire at some time, and 0 otherwise.[7] But now how are we to represent the dependency relations between $F$

[7] This is, in effect, how effects have been represented using "neuron diagrams" in late preemption

and $C$? Since CC will go camping only if there is no fire (in May), we would want to have an equation such as $C = 1 - F$. On the other hand, since there will be a fire (in June) just in case CC goes camping, we will also need $F = C$. This representation is clearly not rich enough, since it does not let us make the clearly relevant distinction between whether the forest fire occurs in May or June. The problem is manifested in the fact that the equations are cyclic, and have no consistent solution.[8]

A second alternative, adopted by Halpern and Pearl [2005, p. 860], would be to use a variable $F'$ that takes the value 0 if there is no fire, 1 if there is a fire in May, and 2 if there is a fire in June. Now how should we write our equations? Since CC will go camping unless there is a fire in May, the equation for $C$ should say that $C = 0$ iff $F' = 1$. And since there will be a fire in June if CC goes camping, the equation for $F'$ should say that $F' = 2$ if $C = 1$ and $F' = 0$ otherwise. These equations are cyclic. Moreover, while they do have a consistent solution, they are highly misleading in what they predict about the effects of interventions. For example, the first equation tells us that intervening to create a forest fire in June would cause CC to go camping in the beginning of June. But this seems to get the causal order backwards!

The third way to model the scenario is to use two separate variables, $F_1$ and $F_2$, to represent the state of the forest at separate times. $F_1 = 1$ will represent a fire in May, and $F_1 = 0$ represents no fire in May; $F_2 = 1$ represents a fire in June and $F_2 = 0$ represents no fire in June. Now we can write our equations as $C = 1 - F_1$ and $F_2 = C \times (1 - F_1)$. This representation is free from the defects that plague the other two representations. We have no cycles, and hence there will be a consistent solution for any value of the exogenous variables. Moreover, this model correctly tells us that only an intervention on the state of the forest in May will affect CC's camping plans.

Once again, our discussion of the methodology of modeling parallels certain metaphysical discussions in the philosophy literature. If heavy rains delay the onset of a fire, is it the same fire that would have occurred without the rains, or a different fire? It is hard to see how to gain traction on such an issue by direct metaphysical speculation. By contrast, when we recast the issue as one about what kinds of variables to include in causal models, it is possible to say exactly how the models will mislead you if you make the wrong choice.

cases. See Hitchcock [2007b, pp. 85–88] for discussion.

[8] Careful readers will note the the preemption case of Example 3 is modeled in this way. In that model, $BH$ is a cause of $BS$, even though it is the earlier shattering of the bottle that prevents Billy's rock from hitting. Halpern and Pearl [2005] note this problem and offer a dynamic model akin to the one recommended below. As it turns out, this does not affect the analysis of the example offered above.

## 5 Dealing with normality and typicality

While the definition of causality given in Definition 1 works well in many cases, it does not always deliver answers that agree with (most people's) intuition. Consider the following example, taken from Hitchcock [2007a], based on an example due to Hiddleston [2005].

EXAMPLE 6. Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee. Bodyguard puts antidote in the coffee, which would have neutralized the poison had there been any. Victim drinks the coffee and survives. Is Bodyguard's putting in the antidote a cause of Victim surviving? Most people would say no, but according to the preliminary HP definition, it is. For in the contingency where Assassin puts in the poison, Victim survives iff Bodyguard puts in the antidote.

Example 6 illustrates an even deeper problem with Definition 1. The structural equations for Example 6 are *isomorphic* to those in the forest-fire example, provided that we interpret the variables appropriately. Specifically, take the endogenous variables in Example 6 to be $A$ (for "assassin does not put in poison"), $B$ (for "bodyguard puts in antidote"), and $VS$ (for "victim survives"). Then $A$, $B$, and $VS$ satisfy exactly the same equations as $L$, $ML$, and $F$, respectively. In the context where there is lightning and the arsonists drops a lit match, both the lightning and the match are causes of the forest fire, which seems reasonable. But here it does not seem reasonable that Bodyguard's putting in the antidote is a cause. Nevertheless, any definition that just depends on the structural equations is bound to give the same answers in these two examples. (An example illustrating the same phenomenon is given by Hall [2007].) This suggests that there must be more to causality than just the structural equations. And, indeed, the final HP definition of causality allows certain contingencies to be labeled as "unreasonable" or "too farfetched"; these contingencies are then not considered in AC2(a) or AC2(b). As discussed by Halpern [2008], there are problems with the HP account; we present here the approach used in [Halpern 2008] for dealing with these problems, which involves assuming that an agent has, in addition to a theory of causality (as modeled by the structural equations), a theory of "normality" or "typicality". (The need to consider normality was also stressed by Hitchcock [2007a] and Hall [2007], and further explored by Hitchcock and Knobe [2009].) This theory would include statements like "typically, people do not put poison in coffee" and "typically doctors do not treat patients to whom they are not assigned". There are many ways of giving semantics to such typicality statements (e.g., [Adams 1975; Kraus, Lehmann, and Magidor 1990; Spohn 2009]). For definiteness, we use *ranking functions* [Spohn 2009] here.

Take a *world* to be a complete description of the values of all the random variables. we assume that each world has associated with it a *rank*, which is just a natural number or $\infty$. Intuitively, the higher the rank, the less "normal" or "typical" the

world. A world with a rank of 0 is reasonably normal, one with a rank of 1 is somewhat normal, one with a rank of 2 is quite abnormal, and so on. Given a ranking on worlds, the statement "if $p$ then typically $q$" is true if in all the worlds of least rank where $p$ is true, $q$ is also true. Thus, in one model where people do not typically put either poison or antidote in coffee, the worlds where neither poison nor antidote is put in the coffee have rank 0, worlds where either poison or antidote is put in the coffee have rank 1, and worlds where both poison and antidote are put in the coffee have rank 2.

Take an *extended causal model* to be a tuple $M = (\mathcal{S}, \mathcal{F}, \kappa)$, where $(\mathcal{S}, \mathcal{F})$ is a causal model, and $\kappa$ is a *ranking function* that associates with each world a rank. In an acyclic extended causal model, a context $\vec{u}$ determines a world, denoted $s_{\vec{u}}$. $\vec{X} = \vec{x}$ is a *cause of $\phi$ in an extended model $M$ and context $\vec{u}$* if $\vec{X} = \vec{x}$ is a cause of $\phi$ according to Definition 1, except that in AC2(a), there must be a world $s$ such that $\kappa(s) \leq \kappa(s_{\vec{u}})$ and $\vec{X} = \vec{x}' \wedge \vec{W} = \vec{w}$ is true at $s$. This can be viewed as a formalization of Kahneman and Miller's [1986] observation that "an event is more likely to be undone by altering exceptional than routine aspects of the causal chain that led to it".

This definition deals well with all the problematic examples in the literature. Consider Example 6. Using the ranking described above, Bodyguard is not a cause of Victim's survival because the world that would need to be considered in AC2(a), where Assassin poisons the coffee, is less normal than the actual world, where he does not. We consider just one other example here (see [Halpern 2008] for further discussion).

EXAMPLE 7. Consider the following story, taken from (an early version of) [Hall 2004]: Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does *not* treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose that there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday?

Suppose that in fact the hospital has 100 doctors and there are variables $A_1, \ldots, A_{100}$ and $T_1, \ldots, T_{100}$ in the causal model, where $A_i = 1$ if doctor $i$ is assigned to treat Billy and $A_i = 0$ if he is not, and $T_i = 1$ if doctor $i$ actually treats Billy on Monday, and $T_i = 0$ if he does not. Doctor 1 is assigned to treat Billy; the others are not. However, in fact, no doctor treats Billy. Further assume that, typically, no doctor is assigned to a given patient; if doctor $i$ is not assigned to treat Billy, then typically doctor $i$ does not treat Billy; and if doctor $i$ *is* assigned to Billy, then typically doctor $i$ treats Billy. We can capture this in an extended causal model where the world where no doctor is assigned to Billy and no doctor

treats him has rank 0; the 100 worlds where exactly one doctor is assigned to Billy, and that doctor treats him, have rank 1; the 100 worlds where exactly one doctor is assigned to Billy and no one treats him have rank 2; and the $100 \times 99$ worlds where exactly one doctor is assigned to Billy but some other doctor treats him have rank 3. (The ranking given to other worlds is irrelevant.) In this extended model, in the context where doctor $i$ is assigned to Billy but no one treats him, $i$ is the cause of Billy's sickness (the world where $i$ treats Billy has lower rank than the world where $i$ is assigned to Billy but no one treats him), but no other doctor is a cause of Billy's sickness. Moreover, in the context where $i$ is assigned to Billy and treats him, then $i$ is the cause of Billy's recovery (for AC2(a), consider the world where no doctor is assigned to Billy and none treat him).

Adding a normality theory to the model gives the HP account of actual causation greater flexibility to deal with these kinds of cases. This raises the worry, however, that this gives the modeler too much flexibility. After all, the modeler can now render any claim that $A$ is an actual cause of $B$ false, simply by choosing a normality order that assigns the actual world $s_{\vec{u}}$ a lower rank than any world $s$ needed to satisfy AC2. Thus, the introduction of normality exacerbates the problem of motivating and defending a particular choice of model. Fortunately, the literature on the psychology of counterfactual reasoning and causal judgment goes some way toward enumerating the sorts of factors that constitute normality. (See, for example, [Alicke 1992; Cushman 2009; Cushman, Knobe, and Sinnott-Armstrong 2008; Hitchcock and Knobe 2009; Kahneman and Miller 1986; Knobe and Fraser 2008; Kahneman and Tversky 1982; Mandel, Hilton, and Catellani 1985; Roese 1997].) These factors include the following:

- Statistical norms concern what happens most often, or with the greatest frequency. Kahneman and Tversky [1982] gave subjects a story in which Mr. Jones usually leaves work at 5:30, but occasionally leaves early to run errands. Thus, a 5:30 departure is (statistically) "normal", and an earlier departure "abnormal". This difference affected which alternate possibilities subjects were willing to consider when reflecting on the causes of an accident in which Mr. Jones was involved.

- Norms can involve moral judgments. Cushman, Knobe, and Sinnott-Armstrong [2008] showed that people with different views about the morality of abortion have different views about the abnormality of insufficient care for a fetus, and this can lead them to make different judgments about the cause of a miscarriage.

- Policies adopted by social institutions can also be norms. For instance, Knobe and Fraser [2008] presented subjects with a hypothetical situation in which a department had implemented a policy allowing administrative assistants to take pens from the department office, but prohibiting faculty from doing

so. Subjects were more likely to attribute causality to a professor's taking a pen than to an assistant's taking one, even when the situation was otherwise similar.

- There can also be norms of "proper functioning" governing the operations of biological organs or mechanical parts: there are certain ways that hearts and spark plugs are "supposed" to operate. Hitchcock and Knobe [2009] show that these kinds of norms can also affect causal judgments.

The law suggests a variety of principles for determining the norms that are used in the evaluation of actual causation. In criminal law, norms are determined by direct legislation. For example, if there are legal standards for the strength of seat belts in an automobile, a seat belt that did not meet this standard could be judged a cause of a traffic fatality. By contrast, if a seat belt complied with the legal standard, but nonetheless broke because of the extreme forces it was subjected to during a particular accident, the fatality would be blamed on the circumstances of the accident, rather than the seat belt. In such a case, the manufacturers of the seat belt would not be guilty of criminal negligence. In contract law, compliance with the terms of a contract has the force of a norm. In tort law, actions are often judged against the standard of "the reasonable person". For instance, if a bystander was harmed when a pedestrian who was legally crossing the street suddenly jumped out of the way of an oncoming car, the pedestrian would not be held liable for damages to the bystander, since he acted as the hypothetical "reasonable person" would have done in similar circumstances. (See, for example, [Hart and Honoré 1985, pp. 142ff.] for discussion.) There are also a number of circumstances in which deliberate malicious acts of third parties are considered to be "abnormal" interventions, and affect the assessment of causation. (See, for example, [Hart and Honoré 1985, pp. 68ff.].)

As with the choice of variables, we do not expect that these considerations will always suffice to pick out a uniquely correct theory of normality for a causal model. They do, however, provide resources for a rational critique of models.

## 6 Conclusion

As HP stress, causality is relative to a model. That makes it particularly important to justify whatever model is chosen, and to enunciate principles for what makes a reasonable causal model. We have taken some preliminary steps in investigating this issue with regard to the choice of variables and the choice of defaults. However, we hope that we have convinced the reader that far more needs to be done if causal models are actually going to be used in applications.

**Acknowledgments:** We thank Wolfgang Spohn for useful comments. Joseph Halpern was supported in part by NSF grants IIS-0534064 and IIS-0812045, and by AFOSR grants FA9550-08-1-0438 and FA9550-05-1-0055.

## References


Adams, E. (1975). *The Logic of Conditionals*. Dordrecht, Netherlands: Reidel.

Alicke, M. (1992). Culpable causation. *Journal of Personality and Social Psychology 63*, 368–378.

Bennett, J. (1987). Event causation: the counterfactual analysis. In *Philosophical Perspectives, Vol. 1, Metaphysics*, pp. 367–386. Atascadero, CA: Ridgeview Publishing Company.

Cushman, F. (2009). The role of moral judgment in causal and intentional attribution: What we say or how we think?". Unpublished manuscript.

Cushman, F., J. Knobe, and W. Sinnott-Armstrong (2008). Moral appraisals affect doing/allowing judgments. *Cognition 108*(1), 281–289.

Davidson, D. (1967). Causal relations. *Journal of Philosophy LXIV*(21), 691–703.

Glymour, C. and F. Wimberly (2007). Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 43–67. Cambridge, MA: MIT Press.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR '08)*, pp. 198–208.

Halpern, J. Y. and J. Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Hansson, R. N. (1958). *Patterns of Discovery*. Cambridge, U.K.: Cambridge University Press.

Hart, H. L. A. and T. Honoré (1985). *Causation in the Law* (second ed.). Oxford, U.K.: Oxford University Press.

Hiddleston, E. (2005). Causal powers. *British Journal for Philosophy of Science 56*, 27–59.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy XCVIII*(6), 273–299.

Hitchcock, C. (2007a). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Hitchcock, C. (2007b). What's wrong with neuron diagrams? In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 69–92. Cambridge, MA: MIT Press.

Hitchcock, C. (2010). Trumping and contrastive causation. *Synthese*. To appear.

Hitchcock, C. and J. Knobe (2009). Cause and norm. *Journal of Philosophy*. To appear.

Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Reprinted by Open Court Press, LaSalle, IL, 1958.

Kahneman, D. and D. T. Miller (1986). Norm theory: comparing reality to its alternatives. *Psychological Review 94*(2), 136–153.

Kahneman, D. and A. Tversky (1982). The simulation heuristic. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment Under Incertainty: Heuristics and Biases*, pp. 201–210. Cambridge/New York: Cambridge University Press.

Kim, J. (1973). Causes, nomic subsumption, and the concept of event. *Journal of Philosophy LXX*, 217–236.

Knobe, J. and B. Fraser (2008). Causal judgment and moral judgment: two experiments. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 2: The Cognitive Science of Morality*, pp. 441–447. Cambridge, MA: MIT Press.

Kraus, S., D. Lehmann, and M. Magidor (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*, 167–207.

Lewis, D. (1973a). Causation. *Journal of Philosophy 70*, 113–126. Reprinted with added "Postscripts" in D. Lewis, *Philosophical Papers*, Volume II, Oxford University Press, 1986, pp. 159–213.

Lewis, D. (1986a). Causation. In *Philosophical Papers*, Volume II, pp. 159–213. New York: Oxford University Press. The original version of this paper, without numerous postscripts, appeared in the *Journal of Philosophy* **70**, 1973, pp. 113–126.

Lewis, D. (1986b). Events. In *Philosophical Papers*, Volume II, pp. 241–270. New York: Oxford University Press.

Lewis, D. K. (1973b). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Mandel, D. R., D. J. Hilton, and P. Catellani (Eds.) (1985). *The Psychology of Counterfactual Thinking*. New York: Routledge.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin CXXI*, 133–148.

Schaffer, J. (2000). Trumping preemption. *Journal of Philosophy XCVII*(4), 165–181. Reprinted in J. Collins and N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, MIT Press, 2002.

Spohn, W. (2003). Dependency equilibria and the causal structure of decision and game situations. In *Homo Oeconomicus XX*, pp. 195–255.

Spohn, W. (2009). A survey of ranking theory. In F. Huber and C. Schmidt-Petri (Eds.), *Degrees of Belief. An Anthology*, pp. 185–228. Dordrecht, Netherlands: Springer.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford University Press.