ON THE DEFINITION OF ACTUAL CAUSE
(Draft copy – Comments are welcome)

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

August 18, 1998

1 Background

This note proposes a formal explication of the notion of "actual cause" as in, for example, "Socrates drinking hemlock was the actual cause of Socrates death." The philosophical literature has been struggling with this notion at least since Mackie (1965) put forward his famous INUS condition and, as we shall outline shortly, no satisfactory solution has emerged from either the logical, probabilistic, or counterfactual approaches to causality [Sosa and Tooley, 1993].

My interest in this topic has been kindled by a recent paper of Michie (1998) where Good's probabilistic measures of causal tendency (Good, 1961–62) are extended to individual events. Michie argues persuasively that in many legal settings, what need be established (for determining responsibility) is not a counterfactual kind of causation, but "cause in fact." A typical example (introduced by Wright (1988)) considers two fires advancing toward a house. If fire A burned the house before fire B, we (and many juries nationwide) would consider fire A "the actual cause" for the damage, even supposing the house would have definitely burned down by fire B, if it were not for A.

The fact that the standard counterfactual test fails in such examples has prompted me to seek an explication for "actual cause" using the Structural Logic (SL) described in Galles and Pearl (1997, 1998) and Halpern (1988) which is also known as Modifiable Structural Equation Model. This logic has been very successful in explicating a number of causal concepts, for example, unit-potential-response [Rubin, 1974; Robins, 1986], truth and probabilities of counterfactuals, effect of action, direct and indirect effects, confounding [Greenland et al., 1998], exogeneity, causal relevance, and instrumental variables. (The reader is referred to Pearl (1995) and Galles and Pearl (1997, 1998) for explication of these concepts.) SL was proposed as a general language for causal reasoning, capable of representing any concept connected with causation. The notion of "actual cause" has somehow escaped the attention of our research team, probably due to preoccupation with counterfactuals, and this note represents a first attempt at formulating "actual causes" in SL. I would value comments from readers.

This report is based on lecture notes written for CS 262C, Spring 1998, and is organized as follows. Following a review of the SL framework (Section 2) Section 3 provides a comparison to other approaches to causation and suggests an explanation of why the notion of actual cause has encountered difficulties in those approaches. Section 3 defines "actual cause" and illustrates, through examples, how the "probability that event $X = x$ actually caused event $Y = y$" can be computed from a given SL model. Section 4 provides a concise summary of the SL approach and demonstrates, through examples, how effects of actions and probabilities of counterfactuals are defined and computed in SL.

2 The SL Framework (a review)

This section provides a brief review of the basic definitions and assumptions behind the structural logic approach as formulated in [Galles and Pearl, 1998]. Readers familiar with SL can skip directly to section 3.

2.1 Definitions

A causal model is a mathematical object that provides an interpretation (and effective computation) of every causal query about the domain. The causal models used in SL are generalizations of the structural equations used in engineering, biology, and economics.[1]

Definition 1 (causal model) A *causal model* is a triple

$$M = \; < U, V, F >$$

where

(i) U is a set of variables, called *exogenous*, that are determined by factors outside the model.

(ii) V is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called *endogenous*, that are determined by variables in the model.

(iii) F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each f_i is a mapping from $U \cup (V \setminus V_i)$ to V_i such that F defines a mapping from U to V. (i.e., F has a unique solution for each state u in the domain of U). Symbolically, F can be represented by writing

$$v_i = f_i(pa_i, u) \quad i = 1, \ldots, n$$

where pa_i is any realization of the (unique) set of variables PA_i in V/V_i (connoting *parents*) that renders f_i nontrivial.

[1]Similar models, called "neuron diagrams" [Lewis, 1973; Hall, 1998] are used informally by philosophers to illustrate chains of causal processors.

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i. We call such a graph the *causal graph* associated with M. This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i.

Definition 2 (submodel) Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. A submodel M_x of M is the causal model

$$M_x = \quad < U, V, F_x >$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{1}$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of functions $X = x$. Implicit in the definition of submodels is the assumption that F_x possesses a unique solution for every u.

Submodels are useful for representing the effect of local actions and changes. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X. The transformation from M to M_x modifies the algebraic content of F, which is the reason for choosing the name *modifiable structural equations.*

Definition 3 (effect of action) Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. The *effect of action* $do(X = x)$ on M is given by the submodel M_x.[2]

Definition 4 (potential response) Let Y be a variable in V, and let X be a subset of V. The *potential response* of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x.

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form "$do(X = x)$ if $Z = z$" can be formalized using the replacement of equations, rather than their deletion [Pearl, 1994]. We will not consider disjunctive actions, of the form "$do(X = x$ or $X = x')$", since these complicate the probabilistic treatment of counterfactuals.

Definition 5 (counterfactual) Let Y be a variable in V, and let X a subset of V. The counterfactual sentence "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.[3]

---

[2] Readers who are disturbed by the impracticality of some local actions (e.g., creating a world where kangaroos have no tails) are invited to replace the word "action" with the word "modification" (see Leamer, 1985). The advantages of using hypothetical external interventions to convey the notion of "local change" are emphasized in Pearl (1995, p. 706).

[3] The connection between counterfactuals and local actions (sometimes invoking "miracles") is made in Lewis (1973) and is further elaborated in Balke and Pearl (1994) and Heckerman and Shachter (1995).

Two special cases are worth noting. First, if $Y = V_i$ and $X = V \setminus Y$, then $Y_x(u) = f_i(pa_i, u)$ where pa_i is the projection of $X = x$ on PA_i. Thus, each function f_i in M may be given a counterfactual or interventional interpretation; it specifies the potential response of V_i to a hypothetical manipulation of all other variables in V. Second, if Y is included in X and $X = x \Longrightarrow Y = y$, then $Y_x(u) = y$. This means that the potential response of a manipulated variable coincides with the values set by the manipulation.

The formulation above shares many features with that of Simon and Rescher (1966). Both are based on an assembly of autonomous physical mechanisms, represented as a set of equations, and both assume a one-to-one correspondence between equations and variables. Simon and Rescher, however, do not treat counterfactual antecedents as actions and they therefore encounter difficulties handling counterfactuals whose antecedents involve endogenous variables. The SL formulation overcomes these difficulties by representing actions in terms of equation-deletion operators,[4] and defining counterfactuals as solutions to reduced sets of equations. This formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 6 (probabilistic causal model) A probabilistic causal model is a pair

$$< M, P(u) >$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

$P(u)$, together with the fact that each endogenous variable is a function of U, defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) = \sum_{\{u \mid Y(u)=y\}} P(u) \tag{2}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x:

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \tag{3}$$

Likewise a causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u|Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{4}$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \mid Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u). \tag{5}$$

[4]An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995).

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$." Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 1997]. The definition of Y_x and Y'_x in terms of two distinct submodels explains away these objections and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F.

2.2 Example

Next we demonstrate the use of structural logic in reasoning about actions and counterfactuals. Additional applications involving the formalization of causal relevance and the interpretation of causal utterances can be found in Galles and Pearl (1997).

2.2.1 Sprinkler example

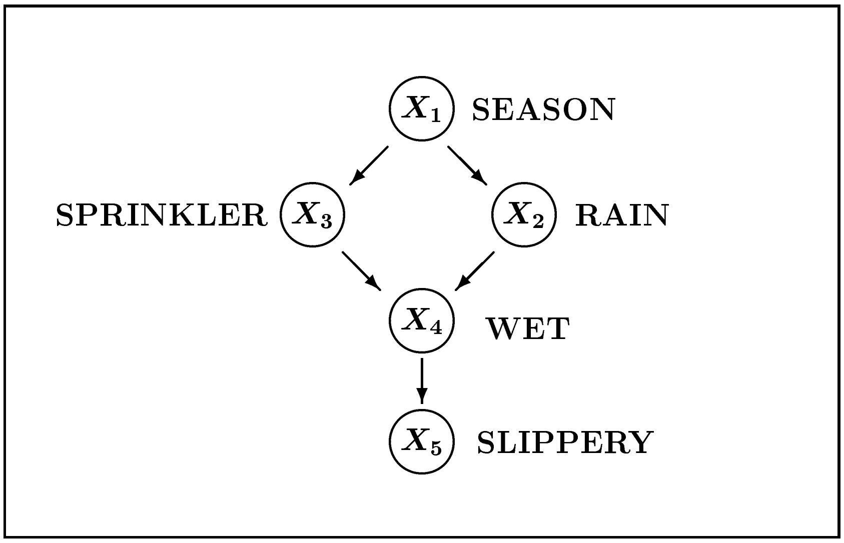

Figure 1: Causal graph illustrating causal relationships among five variables.

Figure 1 is a simple yet typical causal graph used in common sense reasoning. It describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during the season, whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this graph except the root variable X_1 take a value of either "True" or "False." X_1 takes one of four values: "Spring," "Summer," "Fall," or "Winter." Here, the absence of a direct link between, for example, X_1 and X_5, captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$x_1 = u_1$$

$$
\begin{aligned}
x_2 &= f_2(x_1, u_2) \\
x_3 &= f_3(x_1, u_3) \\
x_4 &= f_4(x_3, x_2, u_4) \\
x_5 &= f_5(x_4, u_5)
\end{aligned} \tag{6}
$$

The disturbances $U_1, \ldots, U_5$ are not shown explicitly in Figure 1 but are understood to govern the uncertainties associated with the causal relationships. The causal graph coincides with the Bayesian network [Pearl, 1988] associated with $P(x_1, \ldots, x_5)$ whenever the disturbances are assumed to be independent, $U_i \perp\!\!\!\perp U \setminus U_i$. When some disturbances are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows [Pearl, 1995].

A typical specification of the functions $\{f_1, \ldots, f_5\}$ and the disturbance terms is given by the Boolean model

$$
\begin{aligned}
x_2 &= [(X_1 = \text{Winter}) \vee (X_1 = \text{Fall}) \vee ab_2] \wedge \neg ab'_2 \\
x_3 &= [(X_1 = \text{Summer}) \vee (X_1 = \text{Spring}) \vee ab_3] \wedge \neg ab'_3 \\
x_4 &= (x_2 \vee x_3 \vee ab_4) \wedge \neg ab'_4 \\
x_5 &= (x_4 \vee ab_5) \wedge \neg ab'_5
\end{aligned} \tag{7}
$$

where x_i stands for $X_i = true$, and ab_i and ab'_i stand, respectively, for triggering and inhibiting abnormalities. For example, ab_4 stands for (unspecified) events that might cause the pavement to get wet (x_4) when the sprinkler is off ($\neg x_2$) and it does not rain ($\neg x_3$) (e.g., a leaking water pipe), while $\neg ab'_4$ stands for (unspecified) events that will keep the pavement dry in spite of the rain (x_3), the sprinkler (x_2), and ab_4 (e.g., a pavement covered with a plastic sheet).

To represent the action "turning the sprinkler ON," or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Eq. (6) with the equation $x_3 = \text{ON}$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. Thus, the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution *without* removing the equation for X_3, and therefore may potentially influence (the belief in) every variable in the network. In contrast, the only variables affected by the action $do(X_3 = \text{ON})$ are X_4 and X_5, that is, the descendants of the manipulated variable X_3. This mirrors the difference between *seeing* and *doing*: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the action "turning the sprinkler ON" that a person may consider taking.

This distinction obtains a vivid symbolic representation in cases where the U_i's are assumed independent, because the joint distribution of the endogenous variables then admits the product decomposition

$$
P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{8}
$$

Similarly, the joint distribution associated with the submodel M_x representing the action $do(X_3 = \text{ON})$ is obtained from the product above by deleting the factor $P(x_3|x_1)$ and substituting $X_3 = \text{ON}$

$$P(x_1, x_2, x_4, x_5 | do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, X_3 = \text{ON})\ P(x_5|x_4) \tag{9}$$

The difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$ is thus seen from the corresponding distributions. The former is represented by Eq. (9), while the latter by *conditioning* Eq. (8) on the observation, i.e.,

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

Note that the conditional probabilities on the r.h.s. of Eq. (9) are the same as those in Eq. (8), and can therefore be estimated from pre-action observations, provided $G(M)$ is available. However, the pre-action distribution P together with the causal graph $G(M)$ is generally not sufficient for evaluating counterfactuals sentences. For example, the probability that "the pavement would *continue* to be slippery once we turn the sprinkler off," tacitly presuming that currently the pavement *is* slippery, cannot be evaluated from the conditional probabilities $P(x_i|pa_i)$ alone; the functional forms of the f_i's (Eq. 6) are necessary for evaluating such queries [Balke and Pearl 1994; Pearl 1996].

To illustrate the evaluation of counterfactuals, consider the causal model given by Eq. (7) and assume that all abnormalities vanish and that the only uncertainty in the model lies in the identity of the season, summarized by a probability distribution $P(u_1)$ or $P(x_1)$. We observe the ground slippery and the sprinkler on and we wish to assess the probability that the ground will continue to be slippery had the sprinkler been off. Formally, the quantity desired is given by

$$P(X_{5_{X_3=0}} = 1 | X_5 = 1, X_3 = 1)$$

According to Eq. (4), the expression above is evaluated by summing $P(u)$ over all states of U that are compatible with the sentence to be evaluated. In our example, the only state compatible with the evidence $X_5 = 1$ and $X_3 = 1$ is $X_1 = \text{Summer} \vee \text{Spring}$, and in this state $X_2 = \text{no rain}$, hence $X_{5_{X_3=0}} = 0$. Thus, matching intuition, we obtain

$$P(X_{5_{X_3=0}} = 1 | X_5 = 1, X_3 = 1) = 0.$$

In general, the probability of a counterfactual sentence ($A \Rightarrow B$ given evidence e) can be computed in three steps:

1. **Abduction** – update $P(u)$ by the evidence e available, to obtain $P(u|e)$.

2. **Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.

3. **Deduction** – Use the updated probability $P(u|e)$ in conjunction with M_A to compute the probability of the counterfactual consequence B.

Practical methods of computing probabilities of counterfactuals are presented in Balke and Pearl (1994, 1995).

3 Actual Causes: A Structural-Logic Explication

Assume we are given a causal model $M =< U, V, \{f_i\} >$ and we want to decide whether in a specific state $U = u$ the event $X = x$ was an actual cause of the event $Y = y$, where X and Y are any two members of V. The interventional interpretation of the word "cause" would instruct us to check whether Y would respond to a manipulation that changes x to some other value $x' \neq x$. This leads to the counterfactual criterion $Y_{x'}(u) \neq Y_x(u)$, which tests the overall necessity of $X = x$ in sustaining $Y(u) = y$.

We know, however, that this test is too coarse to detect all cases of causal connections between $X = x$ and $Y = y$. Changing X to x' may interrupt a genuine causal connection and still go undetected, because the new condition $X = x'$ may activate a new mechanism to sustain $Y = y$, a mechanism that was dormant under the condition $X = x$. For example, the switch that currently activates the lamp in my room may be connected in such a way that an auxiliary flush light turns on whenever the switch is turned off. This will keep my room lighted in every switch position; merely testing for the overall effect of the switch on the room light will not reveal the fact that the current switch position is the "actual" cause for the light in the room, since it enables the passage of electric current through the lamp and is in fact the only mechanism currently sustaining light. Likewise, if the room is lighted by two independent light sources, none is truly essential for sustaining light in the room, hence none would meet the counterfactual criterion in isolation. However, we still want to regard each source as a contributory actual cause for the light, because it would become essential, and would meet the counterfactual test, in the hypothetical event that the other is turned off.

Thus, to properly explicate the notion of "actual cause" we must devise a formal way of distinguishing *essential* from *non-essential* mechanisms, and keep the latter *inactive* while testing for the response of Y to a change in X.

3.1 Causal beams: Definitions and implications

We start by defining *essential* variables. Recall that the arguments of the functions $\{f_i\}$ in a causal model were assumed to be essential in some sense, since we have pruned from each f_i all redundant arguments and retained only those called pa_i that render $f_i(pa_i, u)$ nontrivial. However, while in that definition we were concerned with nontriviality relative to all possible u's, further pruning is feasible when we are situated at a particular state $U = u$.

To illustrate, consider the function $f_i = ax_1 + bu_i x_2$. Here $PA_i = \{X_1, X_2\}$, because there is always some value of u_i that would make f_i sensitive to changes in either x_1 or x_2. However, given that we are in a state for which $u_i = 0$, we can safely consider X_2 a trivial argument, replace f_i with $f_i' = ax_1$, and consider X_1 as the only *essential* argument of f_i'. We shall call f_i' the *projection* of f_i on u, and more generally, we will consider the projection of the entire model M by replacing every function in $\{f_i\}$ with its projection relative to a specific u and a specific value of its nonessential part. This leads to a new model which we call *causal beam.*

Definition 7 (causal beam) For model $M =< U, V, \{f_i\} >$ and state $U = u$, a **causal beam** of the event $X = x$ is a new model $M(x, u) = \ < u, V, \{f_i'\} >$, in which the set of functions f_i' is constructed from $\{f_i\}$ as follows:

1. For each variable $V_i \in V$, partition PA_i into two subsets, $PA_i = E \cup NE$, where E (connoting "essential") is any minimal subset of PA_i satisfying[5]

$$f_i(E_x(u), ne, u) = f_i(E_x(u), ne', u) \text{ for all } ne' \tag{10}$$

   In words, E is any minimal set of PA_i sufficient to entail the actual value of $V_i(u)$.

2. For each variable $V_i \in V$, replace $f_i(e, ne, u)$ by its projection $f_i'(e, u)$, given by

$$f_i'(e, u) = f_i(e, NE_{xw}(u), u) \tag{11}$$

   where W is any minimal subset of NE with a value w that renders f_i' non trivial, i.e.,

$$f_i'(e, u) = y' \text{ for some } e \text{ of } E.$$

Thus the new parent set of V_i becomes $PA_i' = E$, and every f' function is responsive to its new parent set E.

Definition 8 (natural beam) A causal beam $M(x, u)$ is said to be *natural*, if condition 2 of Definition 7 is satisfied with $W = \{0\}$, for all $V_i \in V$.

In words, a natural beam is formed by "freezing" the nonessential variables at their actual values, $NE_x(u)$, thus yielding the projection $f_i'(e, u) = f_i(e, NE_x(u), u)$.

Note that there is always a w (W can be empty) that will satisfy condition 2 of Definition 7 for any non-empty E. Note also that when $X = x$ is a natural event (i.e., $X(u) = x$), the subscript x can be removed from E_x and NE_{xw}, because $E_x(u) = E(u)$ and $NE_{xw}(u) = NE_w(u)$. However, when $X = x$ is imposed by external intervention, and $X(u) \neq x$, variables must be set to their corresponding values in the submodel M_x.

Definition 9 (actual cause) We say that event $X = x$ was an **actual cause** of $Y = y$ in a state u (abbreviated "x caused y") iff there exists a natural beam $M(x, u)$ such that:

$$M_x(x, u) \models y \tag{12}$$

$$M_{x'}(x, u) \models y' \neq y \quad \text{whenever} \quad x' \neq x \tag{13}$$

Note that Eq. (12) is equivalent to

$$Y_x(u) = y \tag{14}$$

But Eq. (13) ensures that $Y = y$ is sustained by $X = x$ and not by some other values of X.

Definition 10 (contributory cause) x is a *contributory cause* of y in a state u iff there exists a causal beam, but no natural beam, that satisfies Eqs. (12) and (13).

[5] As usual, we use lowercase letters (e.g., e, ne) to denote specific realizations of the corresponding variables (e.g., E, NE), and $E_x(u)$ to denote the realization of E under $U = u$ and $do(X = x)$.

To summarize, the causal beam can be interpreted as a theory that provides a minimally sufficient and nontrivial explanation for each actual event $X_i = x_i$, under a hypothetical state of the world $(W = w)$ that is minimally removed from the actual state. Using this new theory, we subject the event $X = x$ to a counterfactual test, and check whether Y would change if X were not x. If Y change occurs under a hypothetical state that coincides with the actual state (i.e., $W = \{0\}$), we say that "x was an actual cause of y." If changes occur only under a hypothetical state that is somewhat removed from the actual state (i.e., $W \neq \{0\}$), we say that "x was a contributory cause of y."

When the state u is uncertain, and the uncertainty is characterized by the probability $P(u)$ then, if e is the evidence available in the case, the probability that x caused y can be obtained by summing up the weight of evidence $P(u|e)$ over all states u in which the assertion "x caused y" is true. Formally,

Definition 11 (probability of causes) Let U_{xy} be the set of states in which the assertion "x caused y" is true, and let U_e be the set of states compatible with the evidence e. The probability that x caused y, in light of evidence e, denoted $P(\textit{caused}\ (x, y|e))$, is given by the expression

$$P(\textit{caused}\ (x, y|e)) = \frac{P(U_{xy} \cap U_e)}{P(U_e)} \tag{15}$$

3.2 Examples

Example 1 (disjunctive mechanism) Contributory causation is typified in cases where two actions concur to bring about an event, and either action, operating alone, would have brought about the event absent the other. For example, two factories pouring chemicals into a lake during the same period of time would each be considered a contributing cause to the pollution. In such cases the model consists of just one mechanism which connects the effect E to the two actions through a simple disjunction: $E = A_1 \vee A_2$. There exists no natural beam to qualify either A_1 or A_2 as an actual cause of E. If we fix either A_1 or A_2 at its current value (namely, true), E will become a trivial function of the other action. However, if we deviate minimally from the current state of affairs and let A_2 be false, E would then become responsive to A_1 and will pass the counterfactual test of Eq. (13).

Example 2 (disjunctive normal form) Considering a single mechanism characterized by the Boolean function

$$y = f(x, z, r, s, t, u) = xz \vee rs \vee t$$

where, for simplicity, variables X, Z, R, S and T are assumed to be causally independent of each other (i.e., none is a descendant of another in the causal graph $G(M)$). We next illustrate conditions under which x would qualify as a contributory or actual cause, respectively, for y.

First, consider a state $U = u$ where all variables are true, namely,

$$X(u) = Z(u) = R(u) = S(u) = T(u) = Y(u) = \textit{true}$$

In this state every disjunct represents a set of essential variables. In particular, taking $E = \{X, Z\}$, we find that the projection $f'(x, z, u) = f(x, z, R(u), S(u), T(u)$ becomes trivially true. Thus, there is no natural beam $M(x, u)$, and x could not be an "actual cause" of y. The feasible causal beams correspond to the two minimal choices of w: $w = \{r', t'\}$, or $w = \{s', t'\}$, where primes denote complementation, i.e., $s' \equiv$ "$S = \mathit{false}$." Each of these two choices yields $f'(x, z) = xz$ which renders E nontrivial. Clearly, $M(x, u)$ meets the conditions of Eqs. (12) and (13), thus certifying x as a contributory cause of y.

Using the same argument, it is easy to see that at a state u for which

$$X(u) = Z(u) = \mathit{true} \text{ and } R(u) = T(u) = \mathit{false}$$

a natural beam exists, i.e., a nontrivial projection $f'(x, z) = xz$ is realized by setting the nonessential variables R, S, and T to their actual values in u. Hence, x is qualified as an actual cause of y.

This example illustrates how Mackie's intuition for the INUS condition (Insufficient but Nonredundant part of an Unnecessary but Sufficient condition) can be explicated in the SL framework. It also illustrates the conditions under which a single literal in a disjunction normal form would qualify as a cause — the disjunction formula must represent a genuine mechanism in the sense of Definition 1. The next example illustrates how the INUS condition generalizes to arbitrary Boolean functions, especially those having several minimal conjunctive normal forms.

Example 3 (single mechanism in general Boolean form) Consider the function

$$y = f(x, z, s, u) = xz' \vee x'z \vee xs' \tag{16}$$

which has an equivalent form

$$y = f(x, z, s, u) = xz' \vee x'z \vee zs' \tag{17}$$

Assume, as before, that we consider a state u in which X, Z, and S are true, and that we inquire as to whether the event $x : X = \mathit{true}$ caused the event $y : Y = \mathit{false}$. In this state, the only essential set is $E = \{X, Z, R\}$ because no choice of two variables (valued at this u) would entail $Y = \mathit{false}$ regardless of the third. Since NE is empty, the choice of beam is unique, $M(x, u) = M$, for which $y = f'(x, z, s, u) = xz' \vee x'z \vee xs'$. This $M(x, u)$ passes the counterfactual test of Eq. (13), because $f'(x', z, s) = \mathit{true}$, and we conclude, therefore, that x was an actual cause of y.

Similarly, we can see the event $S = \mathit{false}$ (or s') was an actual cause of $Y = \mathit{false}$. This follows directly from the counterfactual test

$$Y_s(u) = \mathit{false} \text{ and } Y_{s'}(u) = \mathit{true}$$

It is important to note that, because Definitions 9 and 10 rest on semantical considerations, identical conclusions would be obtained from any logically equivalent form of f, as for example the one in Eq. (17). This stands in marked contrast to Mackie's INUS condition which, for want of semantical underpinning, was found to be syntax sensitive [Kim, 1971]. Our explication can therefore be considered the semantical basis behind the INUS intuition.

Example 4 (The desert traveler – after P. Suppes) A desert traveler T has two enemies. Enemy-1 poisons T's canteen, and Enemy-2, unaware of Enemy-1's action, shoots and empties the canteen. A week later, T is found dead and the two enemies confess to action and intention. A jury must decide whose action was the actual cause of T's death.

Let u be the state where the traveler's first need of drink occurred after the shot was fired. Let x be the proposition "Enemy-2 shot".

The original model is characterized by the graph of Figure 2. The causal beam $M(x, u)$ is natural, as depicted in Figure 3. Remarkably, the causal beam $M(p, u)$ formed to test whether p is a cause of y, is identical to $M(x, u)$. Accordingly, we have:

$$\begin{array}{llll} Y_x = 1 & \text{and} & Y_{x'} = 0 & \text{in} \quad M(x, u). \\ Y_p = 1 & \text{and} & Y_{p'} = 1 & \text{in} \quad M(p, u). \end{array} \tag{18}$$

Thus, Enemy-2 shooting at the container (x) was an actual cause of T's death (y), while Enemy-1 poisoning the water (P) was not an actual cause of y.

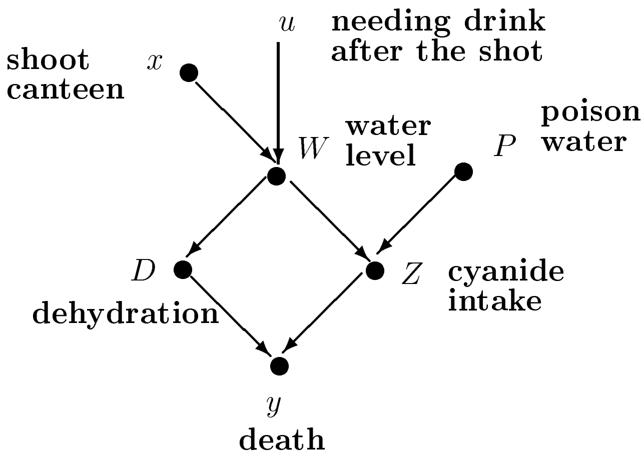

Figure 2:

Next, consider another state, u', standing for the event that our traveler reached for drink before Enemy-2 shot at the canteen. The graph corresponding to $M(x, u')$ is shown in Figure 4:

The causal beam $M(p, u')$ is, again, identical to $M(x, u')$, giving

$$\begin{array}{llll} Y_x = 1 & \text{and} & Y_{x'} = 1 & \text{in} \quad M(x, u') \\ Y_p = 1 & \text{and} & Y_{p'} = 0 & \text{in} \quad M(p, u') \end{array} \tag{19}$$

Thus, in this state of affairs we consider Enemy-1's action to be the actual cause of T's death, while Enemy-2's action is not considered the cause of death.

If we do not know which state prevailed, u or u', we must be satisfied with the probabilistic answer: "The probability that x caused y is $P(u)$." Likewise, if we observe some evidence e reflecting on the probability $P(u)$, such evidence would yield (see Eq. (15):

$$P(caused(x, y|e)) = P(u|e)$$

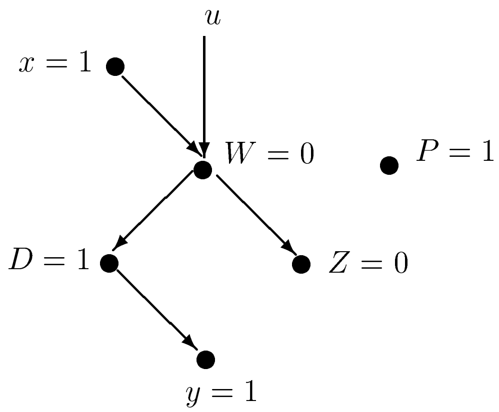

Figure 3:

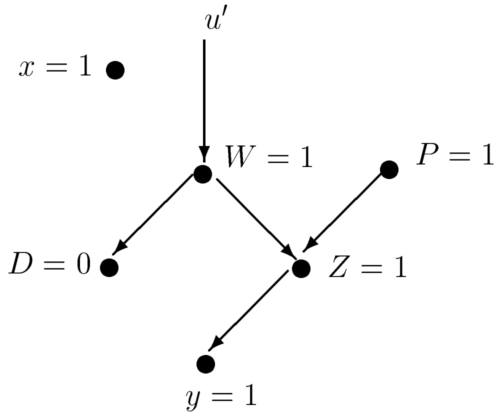

Figure 4:

and

$$P(caused(x', y|e)) = P(u'|e)$$

For example, a forensic report confirming "no cyanide in body" rules out state u' in favor of state u, and the probability of x being the cause of y becomes 100%.

A natural question arises, whether the probability that a traveler could survive poisoning or dehydration (perhaps by the unexpected arrival of a rescue team) should enter the calculations of either $P(caused(x, y|e))$ or $P(caused(p, y|e))$. The details of this consideration are presented in the appendix. It is shown that, because T did not in fact survive, the apriori probabilities of survival plays only a minor role in the calculation. Since chances for rescue are higher under dehydration than under poisoning, T's death provides a weak evidence in favor of poisoning (u'), at the expense of dehydration (u).

Example 5 Let x be the state of a 2-position switch. In position-1 ($x = 1$) the switch turns on ($z = 1$) and turns off a flush light ($w = 0$). In position-2 ($x = 0$) the switch turns on a flush light ($w = 1$) while the lamp is turned off ($z = 0$). Let $Y = 1$ be the proposition that the room is lighted.

The causal beams $M(x, u)$ associated with the state u in which all equipment is working properly are shown in the graphs below (Figure 5):

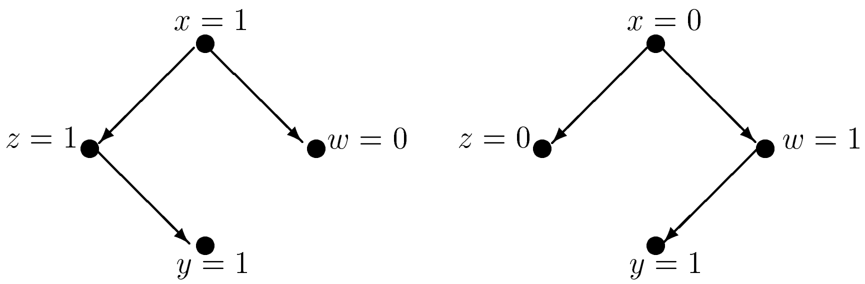

Figure 5:

Here, again, $M(x = 1, u)$ entails $Y_x = 1$ and $Y_{x'} = 0$. Likewise $M(x = 0, u)$ entails $Y_x = 1$ and $Y_{x'} = 0$. Thus both "Switch in position 1" and "Switch in position 2" are considered actual causes for "Room is lighted," although neither is a necessary cause.

4 Comparison to Other Approaches

4.1 The logical approach: INUS, NESS etc.

Attempts to formulate causality in terms of logical sufficiency and necessity encounter insurmountable difficulties [Sosa and Tooley, 1993, p.1-8], partly revolving around J.S. Mill's observation that, in reality, no cause is truly sufficient or necessary [Mill, 1843, p. 425]. Mackie's INUS condition appears to be the earliest attempt to offer an explication of "actual causation" within the logical framework, and it has become quite popular. The basic intuition behind INUS is shared by researchers from a variety of disciplines. Legal scholars, for example, have advocated a relation called NESS [Wright, 1988], standing for "necessary element of sufficient set," which is a rephrasing of Mackie's INUS condition in a simpler mnemonic. In epidemiology, Rothman (1976) has proposed a similar criterion for distinguishing cases where disease is attributable to exposure from those where exposure is merely incidental. These intuitions have all been expressed informally, in a semi-qualitative language of "sufficiency and necessity" but have not been cast in a strictly formal setting, and for a very good reason. Jaegwon Kim (1971), who attempted to provide a careful explication of the INUS condition, admits that logic is not an adequate language for expressing causal concepts. Similar observations are made by Cartwright (1989, pp. 25–34), who resorts to knowledge of mechanisms in order to rule out spurious non-causes from INUS expressions. The weaknesses of logic fall into three main categories:

1. syntactic sensitivity

2. temporal symmetry

3. sensitivity to added variables.

The first two are adequately covered in Kim (1971). I will therefore emphasize the third: causality tests based on logical formulas are not invariant to the addition of fictitious new variables.

Consider the formula

$$E = s \vee p \tag{20}$$

according to which event p is an INUS-cause of effect E. Defining a new event $z = ps'$, and using the identity: $s \vee s'p = s \vee p$, we can rewrite (20) as

$$E = s \vee z \tag{21}$$

Now suppose we know that s is true. That would make z false and we can write (20) as

$$E = s \tag{22}$$

Thus, according to (22) p does not appear as part of any "sufficient set" and, therefore, it is no longer an INUS-cause of E.

What was demonstrated above is merely the phenomenon of preemption in a logical looking glass. (It may be interpreted as a logical attempt to encode the desert-traveler story with E standing for T's death, s and p representing the actions of the shooter and poisoner, respectively, and z standing for "cyanide intake.") This problem does not appear in the SL formulation, because SL does not permit an arbitrary addition of new variables to the model, even when truth values are preserved. The addition of Z would be justified only when there really is an autonomous mechanism involving Z standing on the causal pathway between S and E (as in Figure 2). Standard logic, however, oblivious of such notions as autonomy, pathways, and betweenness, cannot distinguish genuine mechanisms from fictitious definitions.

4.2 The counterfactual approach of Lewis (to be completed)

4.3 Probabilistic causality (to be completed)

Probabilistic causality is a branch of philosophy that attempts to explicate causal relationships in terms of probabilistic relationships. This attempt was motivated by several ideas and expectations. First and foremost, probabilistic causality promises a solution to the centuries-old puzzle of causal discovery, that is, how humans discover genuine causal relationships from bare empirical observations, free of any causal preconceptions. Second, in contrast to deterministic accounts of causation, probabilistic causality offers substantial cognitive economy. Physical states and physical laws need not be specified in minute detail because instead they can be summarized in the form of probabilistic relationships among macro-states so as to match the granularity of natural discourse. Third, probabilistic causality is equipped to deal with the modern (i.e., quantum theoretical) conception of uncertainty, according to which determinism is merely an epistemic fiction, and nondeterminism is the fundamental feature of physical reality.

The formal program of probabilistic causality owes its inception to H. Reichenbach (1956) and I.J. Good (1961–62) and has subsequently been pursued by P. Suppes (1970), B. Skyrms (1980), R. Otte (1981), W. Spohn (1980), and W.C. Salmon (1984). The current state of this program is rather disappointing considering its original aspirations. Salmon has abandoned the effort altogether, concluding that "causal relations are not appropriately analyzable in terms of statistical relevance relations" (1984, page 185); instead, he has proposed an analysis in which "causal processes" are the basic building blocks. More recent accounts by Cartwright (1989) and Eells (1991) have resolved some of the difficulties encountered by Salmon, but at the price of compromising the original goals of the program and invoking deterministic counterfactual relationships. A summary of the major difficulties of probabilistic causality, and a comparison to structural or mechanism-based causality, are given in Pearl (1996b).

Much of human intuition concerning causation rests on the Laplacian illusion that our world is basically deterministic, and that probabilities surface only from our ignorance of the boundary conditions. Probabilistic causality denies determinism a priori, even as a limiting case, and thus deprives its methods and syntax from representing this important component of human cognition. It is not surprising that progress in this approach has been impeded by conceptual and syntactical difficulties. Most problems that plague the logical approach find their counterparts (sometimes disguised) in the probabilistic approach. Legal scholars flatly deny that statistical considerations (e.g., that a certain type of action rarely brings about the effect considered) should enter into judgment of "cause in fact" [Wright, 1988; Robertson, 1997]. Good (1993) admits that not much progress has been made in the past 30 years toward the understanding of actual causes or "causes in fact," and that the parameter χ, which Good has proposed for the degree of actual causation, is still far from adequate explication.

Recent attempt to extend Good's Q measure to singular causes [Michie, 1998] leads to counterintuitive conclusions. In the desert-traveler story, for example, the degree to which shooting the canteen was a sufficient cause for T's death turns out to be a function of the ratio of two small probabilities:

ϵ_1 – the probability that a traveler like T would survive with a poisoned canteen

ϵ_2 – the probability that a traveler like T would survive with an empty canteen.

These probabilities have only minor impact on determining causes in the specific story at hand. Since T was found dead, what we know about survival rates *in general* is of little relevance to events and processes that shaped our story (see the appendix).

Probabilistic methods encounter basic difficulties reconciling evidence (e.g., T's death) that conflicts with the effect of hypothetical actions (e.g., not shooting). Balke and Pearl (1994) have shown that expressions of the type

$$P(T \text{ would have been alive had Enemy-2 not shot} \mid T \text{ is in fact dead})$$

can only be computed if one assumes knowledge of some deterministic mechanism.

Acknowledgment

I am indebted to Donald Michie for calling my attention to the ubiquity of "sufficient causes" in legal discourse. Gary Schwartz referred me to the penetrating papers of [Wright, 1988] and [Robertson, 1997] on causality in the law. Clark Glymour referred me to the paper of [Kim, 1971] and noted that similar ideas were applied in [Glymour and Cheng, 1998] and [Glymour, 1998] for extending the causal power theory of [Cheng, 1997].

This paper is based on lecture outlines for a seminar on causal reasoning given at UCLA (CS 262C, Spring 1998). I thank the seminar participants, Ray Golish, Andrew Lister, Eitan Mendelowitz, Peyman Meshkat, Igor Roizen, and Jin Tian for knocking down two earlier attempts, and for stimulating discussions leading to the current proposal.

References

[Balke and Pearl, 1994] A. Balke and J. Pearl. Counterfactual probabilities: Computational methods, bounds, and applications. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA, 1994.

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

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

[Cheng, 1997] P.W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

[Dawid, 1997] A.P. Dawid. Causal inference without counterfactuals. Technical report, Department of Statistical Science, University College London, UK, 1997.

[Eells, 1991] E. Eells. *Probabilistic Causality*. Cambridge University Press, Cambridge, MA, 1991.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. Technical Report R-250, Department of Computer Science, University of California, Los Angeles, 1998. To appear in *Foundation of Science*, 1998.

[Glymour and Cheng, 1998] C. Glymour and P.W. Cheng. Causal mechanism and probability: A normative approach. Technical report, UCSD, Carnegie Mellon University, and UCLA, 1998. Draft of chapter to appear in Y. Oaksford and V. Chater (Eds), *Rational Models of Cognition*, Oxford, England: Oxford University Press.

[Glymour, 1998] C. Glymour. Psychological and normative theories of causal power and probabilities of causes, 1998. To appear in *Proceedings of the 1998 Conference on Uncertainty in Artificial Intelligence (UAI-98).*

[Good, 1961] I.J. Good. A causal calculus, I. *British Journal for the Philosophy of Science*, 11:305–318, 1961.

[Good, 1962] I.J. Good. A causal calculus, II. *British Journal for the Philosophy of Science*, 12:43–51; 13:88, 1962.

[Good, 1993] I.J. Good. A tentative measure of probabilistic causation relevant to the philosophy of the law. *J. Statist. Comput. and Simulation*, 47:99–105, 1993.

[Greenland et al., 1998] S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. Technical Report R-251, University of California, Los Angeles, January 1998. To appear in *Epidemiology*.

[Hall, 1998] N. Hall. Two concepts of causation. In press, 1998.

[Heckerman and Shachter, 1995] D. Heckerman and R. Shachter. A definition and graphical representation for causality. In P. Besnard and S. Hanks, editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Mateo, CA 1995. Morgan Kaufmann.

[Kim, 1971] J. Kim. Causes and events: Mackie on causation. *Journal of Philosophy*, 68:426–471, 1971. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

[Leamer, 1985] E.E. Leamer. Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy*, 22:255–304, 1985.

[Lewis, 1973] D. Lewis. Causation. *The Journal of Philosophy*, 70:556–567, 1973. Reprinted with postscript in D. Lewis, *Philosophical Papers*, vol. II. New York: Oxford, 1986.

[Mackie, 1965] J.L. Mackie. Causes and conditions. *American Philosophical Quarterly*, 2/4:261–264, 1965. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

[Michie, 1998] D. Michie. Adapting Good's q theory to the causation of individual events. Technical report, University of Edinburgh, UK, 1998. Submitted for publication in *Machine Intelligence 15*.

[Mill, 1843] J.S. Mill. *System of Logic*, volume 1. John W. Parker, London, 1843.

[Otte, 1981] R. Otte. A critque of suppes' theory of probabilistic causality. *Synthese*, 48:167–189, 1981.

[Pearl, 1988] J. Pearl. Embracing causality in formal reasoning. *Artificial Intelligence*, 35(2):259–271, 1988.

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, December 1995.

[Pearl, 1996a] J. Pearl. Causation, action, and counterfactuals. In Y. Shoham, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, pages 51–73. Morgan Kaufmann, San Francisco, CA, 1996.

[Pearl, 1996b] J. Pearl. Structural and probabilistic causality. In D.R. Shanks, K.J. Holyoak, and D.L. Medin, editors, *The Psychology of Learning and Motivation*, volume 34, pages 393–435. Academic Press, San Diego, CA, 1996.

[Reichenbach, 1956] H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

[Robertson, 1997] D.W. Robertson. The common sense of cause in fact. *Texas Law Review*, 75(7):1765–1800, 1997.

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

[Rothman, 1976] K.J. Rothman. Causes. *American Journal of Epidemiology*, 104:587–592, 1976.

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

[Salmon, 1984] W.C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, 1984.

[Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

[Skyrms, 1980] B. Skyrms. *Causal Necessity*. Yale University Press, New Haven, 1980.

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

[Sosa and Tooley, 1993] E. Sosa and M. (Eds.) Tooley. *Causation*. Oxford readings in Philosophy. Oxford University Press, 1993.

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

[Spohn, 1980] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9:73–99, 1980.

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

[Suppes, 1970] P. Suppes. *A Probabilistic Theory of Causality*. North-Holland Publishing Co., Amsterdam, 1970.

[Wright, 1988] R.W. Wright. Causation, responsibility, risk, probability, naked statistics, and proof: Prunning the bramble bush by clarifying the concepts. *Iowa Law Review*, 73:1001–1077, 1988.

Appendix - I (The rescuable desert traveler)

This appendix analyzes a stochastic version of the desert traveler story (Example 4), assuming that travelers, in general, have non-zero probability of surviving infliction such as poisoning and dehydration. (e.g., by unexpected rescue).

To incorporate such considerations, we introduce another state variable u_r as an argument of f_Y,

$$y = f_Y(d, z, u_r)$$

and let u_r attain three possible values

$$u_r = \begin{cases} 0, & \text{no rescue appears} \\ 1, & \text{rescue appears in time to save T from possible dehydration but not from poison} \\ 2, & \text{rescue appears very early, in time to save T from either poison or dehydration.} \end{cases}$$

Let the probabilities associated with these three scenarios be p_0, p_1, and p_2, respectively. Assuming we have no forensic analysis, our only evidence is T's death, namely, $e : Y = y$. This evidence categorically rules out $u_r = 2$, and modifies respectively, p_0, and p_1 into p'_1, $p'_0 = \frac{p_0}{p_o+p_1}$ and $p'_1 = \frac{p_1}{p_0+p_1}$.

We now need to examine four possible states, as shown in the table below:

State	$(u, u_r = 0)$,	$(u, u_r = 1)$	$(u', u_r = 0)$,	$(u', u_r = 1)$
Prior prob	$P(u)p'_0$	$P(u)p'_1$	$P(u')p'_0$	$P(u')p'_1$
State of Y	1	0	1	1
Posterior Prob	$kP(u)p'_0$	0	$kP(u')p'_0$	$kP(u')p'_1$
Cause of y	x	x	p	p

Accordingly, we have

$$P(cause(x, y|y)) = kP(u)p'_0$$

$$P(cause(p, y|y)) = kP(u')(p'_0 + p'_1) = kP(u')$$

Thus, the ratio of $P(cause(x, y|y))/P(cause(p, y|y))$ becomes $\frac{P(u)p'_0}{P(u')}$ which is only a fraction of p'_0 lower than the previous ratio, before considering probabilities of survival. The reason for the change is that T's death provides additional evidence in favor of u', at the expense of u. However, due to the small probability of rescue, p_0 is close to one, and the correction seems negligible small.