Causal Reasoning with Causal Models

Kevin B. Korb, Charles R. Twardy,*
Toby Handfield and Graham Oppy†

November 28, 2005

Abstract

We introduce and discuss the use of Bayesian networks for causal modeling. Despite their growing popularity and utility in this application, numerous objections to it have been raised. We address the claims that Chickering's arc reversal rule undermines a causal interpretation and that failures of Reichenbach's Common Cause Principle, or again failures of faithfulness, invalidate causal modeling. We also argue against Pearl's deterministic interpretation of causal models. Against these objections we propose new model-building principles which evade some of the difficulties, and we put forward a concept of causal faithfulness which holds when faithfulness simpliciter fails. Finally, we particularize our account of type causal relevance to token causal relevance, providing an alternative to the recent deterministic accounts of token causation due to Hitchcock and Halpern & Pearl.

Keywords: causal models, type and token causality, causal faithfulness, Bayesian networks, probabilistic causality, causal processes, Common Cause Principle.

1 Introduction

While still within a few decades of its origin, Bayesian network technology has achieved prominence as a tool for probabilistic reasoning in artificial intelligence, and philosophers of science have begun to adopt the technology for reasoning about causality and methodology (e.g., Bovens and Hartmann, 2002; Hitchcock, 2001a). Here we introduce some of the basic ideas and features of Bayesian networks and defend their interpretation in causal terms and, particularly, in the terms of a probabilistic theory of causality. This interpretation underwrites much of the reasoning behind causal discovery algorithms, although we won't be examining those here.[1] In the process of developing our causal interpretation we will take issue with some skeptical assessments of causal modeling with Bayesian networks, including their common 'unfaithfulness' to reality, their deterministic interpretation, and some of the difficulties raised in reference to Reichenbach's Common Cause Principle. We will frequently adopt the 'point of view' of causal discovery algorithms, since their success is what is driving much of the current debate about causal interpretation; that is, we will often assess matters in terms of whether a generic problem is raised, or not, for algorithms

*School of Computer Science & Software Engineering; Monash University; Clayton, Victoria; Australia.
†School of Philosophy and Bioethics; Monash University; Clayton, Victoria; Australia.

[1] See instead, *inter alia,* Heckerman, 1998; Spirtes et al., 2000; Cooper and Glymour, 1999; Neapolitan, 2004; Korb and Nicholson, 2004.

that can learn causal models from data. We advance the view that many of the problems encountered in these disputes are properly addressed by developing well-founded model-building rules, in particular rules for individuating variables and for employing arcs. We offer a tentative set of rules aimed at addressing these problems. They also appear to be helpful in developing a unified account of type and token causality, towards which we make an initial effort here.

2 What Is a Causal Model?

We begin by introducing terms and concepts underlying both the technical and philosophical discussion of Bayesian networks; readers familiar with these will want to begin one section forward. A **Bayesian network** is a directed acyclic graph (DAG) over a set of variables V with associated conditional probabilities which represent a probability distribution over the joint states of the variables. For a simple example, see Figure 1. *Rain* and *Sprinkler* are the **root** nodes (equivalently, the **exogenous** variables), which report whether there is rain overnight and whether the automatic sprinkler system comes on. The **endogenous** (non-root) variables are *Lawn*, which describes whether or not the lawn is wet, and *Newspaper* and *Carpet*, which respectively describe the resultant sogginess and muddiness when the dog retrieves the newspaper. Note that we shall vary between talk of variables and their values and talk of **event types** and their corresponding **token events** without much ado.

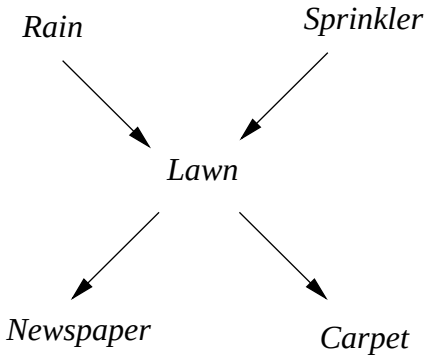

Figure 1: A simple Bayesian network.

Probabilistic reasoning is computationally intractable (NP-hard; Cooper, 1990); likewise, causal discovery (Chickering et al., 1995). The substantial advantage Bayesian networks offer for probabilistic reasoning is that, if the probability distribution can be represented with a sparse network, the computations become practicable. In order for these savings to be real, the lack of an arc between two variables must be reflected in a probabilistic independence of the system being modeled. Thus, in a simple two-variable model with nodes X and Y, a missing arc implies that X and Y are probabilistically independent. If they are not, then the Bayesian network simply fails to be an appropriate model. Thus, *Rain* and *Sprinkler* must be independent of each other; if the sprinkler system is turned off in rainy weather, then Figure 1 is simply the wrong model.

X and Y being probabilistically independent just means that $P(X = x_i|Y = y_j) = P(X = x_i)$ for any two states x_i and y_j. **Conditional independence** generalizes this to cases where a third variable may induce an independence between the first two variables. Philosophers, following Reichenbach, have tended to call this relationship **screening off**. For example, *Rain* and *Newspaper* are presumably dependent in Figure 1; however, if we

hold fixed the state of the lawn — say, we already know it is wet — then they are no longer probabilistically related: $P(Newspaper|Lawn, Rain) = P(Newspaper|Lawn)$. Given these and like facts for other variables), Figure 1 is said to have the **Markov property**: that is, all of the conditional independencies implied by the Bayesian network are true of the actual system (or, equivalently, it is said to be an **independence map (I-map)** of the system).

In the opposite condition, where all apparent dependencies in the network are realized in the system, the network is called a **dependence-map (D-map)** of the system; alternatively, the network is said to be **faithful** to the system. A network which both satisfies the Markov property and is faithful is said to be a **perfect map** of the system. There is no general requirement for a Bayesian network to be faithful in order to be considered an adequate probabilistic model. In particular, arcs can always be added which do nothing — they can be parameterized so that no additional probabilistic influence between variables is implied. Of course, there is a computational cost to doing so, but there is no misrepresentation of the probability distribution. What we are normally interested in, however, are I-maps that are **minimal**: i.e., I-maps such that if any arc is deleted, the model is no longer an I-map for the system of interest. A minimal I-map need not necessarily also be a perfect map, although they typically are; in particular, there are some systems which have multiple distinct minimal I-maps.

Causal models are Bayesian networks each of whose arcs (arrows) can be given a causal interpretation. Each arc represents a direct causal relation between its parent and its child. In general, the set of causal models true of a system will be a subset of the set of Bayesian networks true of the system. For one thing, a system should have only (at most) one true causal model.[2] For another, we can find other Bayesian networks that can represent the same probability distributions by employing Chickering's arc-reversal rule (Chickering, 1995).

So, an arc in a causal model represents a direct causal relation. In order to make sense of the causal interpretation of Bayesian networks, we also need to consider indirect causal connections. These occur across paths. A **path** is a sequence of nodes which can be visited by traversing arcs in the model in either direction (i.e., either with or against the causal flow), in which no node is visited twice. A **directed path** (or **causal path**) is one in which the traversal is entirely in the causal direction. A path between X and Y is **blocked** in case, taking that path in isolation, an observation of one of the variables has no probabilistic impact upon the other. Such questions are usually raised in some observational context, represented, say, by the set $\mathbf{Z}$ of observed variables. In this case, asserting that a path Φ between X and Y is blocked by $\mathbf{Z}$ is equivalent to asserting that X and Y are conditionally independent given $\mathbf{Z}$ (assuming no other paths between them are present).

For a causal model satisfying the Markov property, the graph-theoretic correlate to blocking (generalized across all paths between X and Y) is called **d-separation** (direction-dependent separation). That is, X and Y are d-separated given $\mathbf{Z}$ (for any subset $\mathbf{Z}$ of variables not including X or Y) if and only if each distinct path Φ between them is cut by one of the graph-theoretic conditions:

1. Φ contains a chain $X_1 \longrightarrow X_2 \longrightarrow X_3$ and $X_2 \in \mathbf{Z}$.

2. Φ contains a common causal structure $X_1 \longleftarrow X_2 \longrightarrow X_3$ and $X_2 \in \mathbf{Z}$.

[2]One of the philosophical debates in this area is whether or not there are causal systems which have *no* causal models. We shall enter into this debate below.

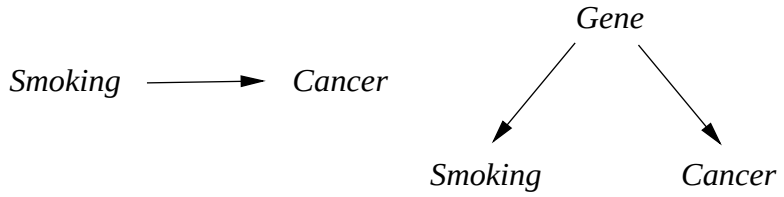

Figure 2: Fisher's smoking example.

3. Φ contains a common effect structure $X_1 \longrightarrow X_2 \longleftarrow X_3$ (i.e., a **collider**) and neither X_2 nor any descendant of X_2 is in $\mathbf{Z}$.

The idea is simply that dependencies can be cut by observing intermediate variables or common causes, on the one hand, and induced by observing common effects (or their descendants), on the other. As for the latter, if we assume per above that the automated sprinkler and rain are independent in Figure 1, they will not remain so if we presuppose knowledge of the state of the lawn. For example, if the lawn is wet, something must explain that, so if we learn that there was no rain overnight, we must *thereby* increase our belief that the sprinkler system came on.

A more formal version of the **Markov property** is then: a model has the Markov property relative to a system if the system has a conditional independence corresponding to every d-separation in the model. The causal Markov property is a close relative, which explicitly invokes a causal interpretation of Bayesian networks. The **causal Markov property** holds for a model if and only if every variable X is independent of every set of variables $\mathbf{S}$ (exclusive of X and its descendants) when conditioning upon the direct causes (parents) of X.

The opposite condition to d-separation is **d-connection**. A related, but non-equivalent, concept is that of an active path. We use active paths to consider the probabilistic impact of observations of some variables upon others: a path between X and Y is **active** in case an observation of X can induce a change in the probability distribution of Y. The concepts of d-connected paths and active paths are not equivalent because a d-connected path may be inactive due only to its parameterization.

Bayesian networks which are causal models represent most plainly **type causality**. That is, the causal relations they encode directly are causal relations between types of events, in contrast to assertions about specific events, or **token causality**. We consider the exact relation between type and token causality in §8 below, where we develop an explicit criterion for token causality.

3 Are Bayesian Networks Causal Models?

There has been debate from the beginning over the status of causal interpretations of Bayesian networks. In Figure 1 the arcs seem to be causal; plausibly, *Sprinkler* $\longrightarrow$ *Lawn means* that the sprinkler being on or off causes the lawn to get wet or not, under some circumstances. Exactly what the causal relation means will be developed throughout, but we start with what appears to be a universal intuition about causality: if we intervene upon a cause C of E, then, at least in some circumstances, the probability distribution for E will change. R.A. Fisher's alternative hypotheses relating smoking and cancer (Figure 2) provide a clear example (Fisher, 1957). In the first model, interventions on smoking have a direct effect on cancer. On Fisher's second model they have none: the lack of an effectual intervention would provide evidence against a direct causal connection.

This gives us our first model-building principle:

Principle 1 (Intervention Principle) *Variables in a causal model must be intervenable.*[3]

In ordinary usage, an intervention represents an influence on a causal system which is external to that system. It's common in the current philosophy and AI literature to consider only *perfect* interventions, that is, interventions which impact on exactly one variable and which infallibly set that variable to exactly one state. Such interventions are, in fact, what Pearl has formalized in his *do-calculus* (Pearl, 2000). Spirtes et al. (2000) also restrict themselves to perfect interventions by choosing to represent interventions by arc-cutting in Bayesian nets: we can represent an intervention upon C by setting it to the given value, cutting all in-coming arcs from its parents, and then applying ordinary Bayesian net propagation. An alternative we prefer is to *augment* the given model with intervention variables: a variable C to be intervened upon receives a new parent I_C. A perfect intervention can then be modeled by adopting the conditional probability $P(C = c|I_{C=c}) = 1$, allowing no other connections between I_C and the model, and using ordinary Bayesian net propagation. The advantage of using augmented models is that various *imperfect* interventions can then also be modeled — for example, interventions which interact with other parents of C, as an antidote would with a poison, or interventions which are not necessarily effective, such as a doctor's attempt to reduce a patient's smoking habit. Although, for the sake of simplicity, many of the examples we will consider below involve perfect interventions, we will not, as others do, rely upon that perfection in developing our theory. Theories which cannot generalize to deal with antidotes, doctors and the world as we know it are, we take it, of very limited use.[4]

All of this leaves open the question of what *interventions* are. We propose the following working definition of how they might be represented.[5]

Definition 1 *An intervention on a variable C in a causal model M transforms M into the augmented model M' which adds $I_C \longrightarrow C$ to M where:*

1. *I_C is exogenous in M'.*

2. *I_C directly causes (is a parent of) C.*

[3]In adopting this principle, and in similar moves throughout, we explicitly rule out non-causal models, of course. It is no part of our intention to deny the value of such models and the techniques being developed to employ them; however, our focus here is to understand those models which are specifically causal.

The possibilities invoked in our model-building principles — here, the possibility of intervention — are, of course, subject to alternative interpretations, ranging at least from practical possibility to logical possibility. In this work, we assume that physical possibility, what is consistent with the laws of nature for our world, is the appropriate concept.

Incidentally, it will occur to some that presupposing physical possibility presupposes causal structure, which may tend towards a circular analysis of causal relationships. However, we, along with much of the probabilistic causality community, reject the idea of reducing the concept of causality to non-causal notions (see, e.g., Irzik, 1996). Our goal here is to relate token and type causality to each other and to further concepts, but not to eliminate causal language.

[4]For discussion of some of the different types of causal intervention see Korb et al. (2004). We note in passing that by adding new intermediate variables imperfect interventions can be emulated by systems with exclusively perfect interventions. But in that case insistence upon perfection both complicates and misleads.

[5]This may be compared with many in the literature, but again those we know of are limited to perfect interventions. For example, definition M of Woodward and Hitchcock (2003) adopts such a restriction in its conditions 2 and 4.

Most generally we allow I_C to have children additional to C, as an additional form of imperfection. We suggest that talk of an intervention will only be legitimate when there is a possible causal process, in the sense of Salmon (1984) and Dowe (2000), connecting the intervention variable and its child, which we will discuss further below.

Given this much of an account of what causal models are, we can consider the question: is there some sense in which a causal interpretation of Bayesian networks is fundamental? The common argument against such a claim is: Bayesian networks were built for, and are used for, representing probability distributions; even when we're given a model which is known to represent exclusively causal connections, we can use Chickering's (1995) arc-reversal rule to find other, *anti-causal*, Bayesian networks that represent the very same probability distribution; therefore, those networks which do represent causal structure only do so by accident.

The premises of this argument are correct; however, the conclusion is false. Chickering's transformation rule allows us to traverse a sequence of Bayesian networks all of whose later networks are capable of representing the probability distributions of the earlier networks, although typically not vice-versa. The rule is:

Rule 1 (Chickering's Transformation Rule) *The transformation of M to M', where $M = M'$ except that, for some variables C and E, $C \longrightarrow E \in M$ and $C \longleftarrow E \in M'$ (and excepting any arc introduced below), will allow the probability distribution induced by M to be represented via M' so long as:*

> *if any collider is introduced or eliminated, then a covering arc is added (e.g., if $A \longrightarrow C \longrightarrow E \in M$ then A and E must be directly connected in M').*

Clearly, this rule only *introduces* arcs and never eliminates any. Thus, when starting with a causal model and applying a sequence of Chickering transformations to find additional models capable of representing the original probability distribution, we can only start from a simpler model and reach (monotonically) ever more complex models. For example, we can apply this rule to our *Sprinkler* model in order to find that the alternative of Figure 3 can represent the probabilities just as well. Except that, this model clearly does not represent the probabilities *just as well* — it is far more, needlessly, complex. The more complex networks introduced by Chickering transformations fail to represent the very same causal models, and under a causal interpretation they will imply falsehoods about the consequences of interventions on its variables.

It is worth noting in passing that when a Chickering transformation from M to M' fails to introduce any arc, then the two models are said to be **statistically indistinguishable**: any set of observational evidence will support one to the same degree that it supports the other (that is, when maximizing the likelihoods of the two models, the result is the same). Of relevance to our discussion will be the further point that statistical distinguishability is very much weaker than empirical distinguishability (which we prove in Appendix B).

These considerations suggest a conjecture:

Conjecture 1 (Causal Simplicity) *The true causal model for a physical system is always the structurally simplest among those capable of representing the system's probability distribution.*

By the structurally simplest network we just mean the sparsest network, where density is measured by the number of arcs present divided by the number possible, which latter is just the number of pairs of nodes — $\binom{N}{2}$ — where N is the number of nodes.

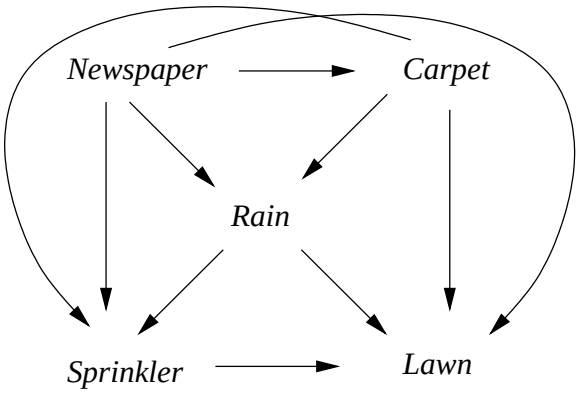

Figure 3: A less simple Bayesian network.

As a conjecture we think this is highly meritorious. To be sure, it is false — a common drawback to conjectures. Indeed, we will demonstrate its falsehood below. However, it leads to a later, related conjecture, which we think may be true, since it has not yet been falsified. The path required takes us through some of the skeptical literature, which forces us to find some necessary qualifications.

4 Indeterministic Causal Models

Wesley Salmon, when developing his version of probabilistic causality in the 1980s, considered it something of a scandal that nearly a century after the development of quantum mechanics and its fundamental commitment to indeterminism (under the usual interpretation of the theory) philosophers of causality were still almost exclusively entertaining deterministic analyses of causality (Salmon, 1984). His basic point was that, since both deterministic and indeterministic universes are consistently describable, it cannot be a matter for a priori philosophizing to decide the issue. A philosophy of causality committed to determinism is deciding upon matters that only an empirical investigation of the world can do. Indeterministic analyses of causation have the distinct advantage of being agnostic about determinism, since stochastic relations can accommodate any kind of universe.

It is a curiosity, then, of the Bayesian network collaboration between AI and philosophy of science that many involved agree to interpret these causal networks as deterministic, despite the fact that they are explicitly probabilistic models. Pearl, for one, is adamant that a deterministic conception of causality is required and for three reasons (Pearl, 2000, pp. 26-7):[6]

1. Determinism is intuitive.

2. Counterfactuals and causal explanation can only be made sense of given a deterministic interpretation.

3. The deterministic interpretation is more general, since any indeterministic model can be transformed into a deterministic model.

Whether determinism is intuitive or not, we shall leave to the reader to ponder. It seems to us that the most plausible approach to giving a fundamental causal interpretation to the

[6]For a philosopher voicing the same opinion see, for example, Cartwright (2001). Like-minded researchers include Joseph Halpern and Dan Hausman.

individual arcs in Bayesian networks is to understand them as indeterministic propensities to bring about states of affairs, something along the lines discussed by Gillies (2000).[7] In any case, for our part, we share Wesley Salmon's intuition, that the deterministic accounts of causation have passed their use-by date.

Regarding the second point, Pearl has been collaborating with Joseph Halpern in developing a deterministic account of token causation and causal explanation (Halpern and Pearl, 2005). We hope, however, that an indeterministic account of token causation (and, ultimately, causal explanation) is not impossible, since we present one below, in §8.

So Pearl's case seems to stand or fall on his claim of generality. It is undeniable that any Bayesian network can be converted into a deterministic model. The point, however, is empty, since equally every deterministic system can be represented as an indeterministic system. Even were things otherwise, it would remain deniable that the deterministic version is the proper vehicle for interpreting the original. We illustrate with a simple three-variable model which is linear (the simplest kind of Bayesian network). Structurally we have: $X \longrightarrow Z \longleftarrow Y$. The more common way to write linear models is with equations of this type:

$$Z = a_1 X + a_2 Y + U$$

Here, a_1 is a coefficient representing the degree of dependency of Z upon X and a_2 the dependency of Z upon Y. But, Z is not a strict function of any of X or Y or the combination of the two: there is a residual degree of variation, described by U. U is variously called the residual, the error term, the disturbance factor, etc. Whatever it's called, once we add it into the model, the model is deterministic, for Z certainly is a function — a linear function, of course — of the combination of X, Y and U. Does this make the physical system we are trying to model with the equation (or, Bayesian network) deterministic? Well, only if as a matter of fact U describes a variable of that system. Since as a matter of actual *practice* U is typically identified only in negative terms, as what is 'left over' once the influences of the other parents of Z have been accounted for, and since in that typical practice U is only ever measured by measuring Z and computing what's left over after our best predictions using X and Y, it is simply not plausible to identify this as a variable of the system. What is represented by U is everything that either is unknown about this system or else is *unknowable* about this system, the ineradicable indeterminism in its fundamental relationships. Any justification for bundling all such unknowns and unknowables into a 'known' variable can only lie in an a priori argument for determinism. But since indeterministic worlds are describable and, for all we can see, consistent, such an a priori argument would be ruling out a posteriori possibilities, which is something any

[7] See Popper (1959) for the original description of a propensity theory of probability. Some seem to think that Paul Humphreys' (1985) objection is fatal to propensity theory. That objection is that the concept of a propensity is inherently forward looking (in time or causality), whereas probabilities can be computed in any temporal direction; therefore, to talk about a propensity for a cause given an effect, as a full-blooded propensity interpretation of the probability of a cause given an effect must, reveals a deep confusion. This objection, however, is mere pedantry: it wants to tie the metaphysics of probability to our linguistic habits in using the term 'propensity'. Given a complete set of forward-looking propensities, it is trivial that we can compute all other probabilities, backwards looking or sideways looking (the common methods of updating Bayesian networks by sampling top-down — cf. Henrion (1988) — make this clear enough!). If Humphreys is uncomfortable calling the latter propensities, then let us agree to call them limit frequencies in a counterfactually infinite sequence of samples from a single chance setup, the physical probabilities determined by the propensities; to simplify our language, we can just call them all 'propensities'. It is perhaps worth noting that in this sense Richard von Mises (1957) was probably a propensity theorist — but then Popper was explicit in acknowledging von Mises as the source of his theory, proposing propensities as a means of applying von Mises' ideas to single case probabilities.

reasonable a priori theory should not aspire to do. In short, the identification of causal models with their deterministic counterparts has been achieved only by presumption.[8]

So, given an indeterministic model we can trivially turn it into a deterministic model, by computing a term to cover all and only residual variances in endogenous variables. Or, given a deterministic model we can render it indeterministic by re-introducing residual variances. But, although the two models are equivalent in the sense that they make the same predictions when conditionalizing on any subset of measured variables, they are not equally meritorious. In particular, the residual variables of the deterministic models are not in any interesting sense observable — they are only measured by their silhouettes, the amount of variation in an endogenous variable which is left over after all the variation predicted by its parents is accounted for; but what is not even in-principle observable is not intervenable in principle, we could only substitute blundering for manipulation in such cases. Hence, Pearl's deterministic models do not even count as causal models. These considerations lead us to a derivative principle, which we dedicate to the memory of logical positivism:

PRINCIPLE 1(A). (Positivism). *Variables in a causal model must be observable, at least in principle.*

5 Reichenbach's CCP

In *The Direction of Time* Hans Reichenbach (1956) introduced his famous (or infamous) **Common Cause Principle** (CCP). CCP asserts that, for any two variables X and Y which are probabilistically dependent, one of the following obtains:

1. X causes Y, directly or indirectly.
2. Y causes X, directly or indirectly.
3. X and Y have a common cause (ancestor) Z such that X and Y are independent conditional upon Z.

Reichenbach claimed that these are the only plausible causal explanations available for the dependence between X and Y, so if none of the above conditions hold, then we are left with magic. The CCP can be interpreted as the methodological principle that we should assume, or search for, one of the three causal explanations for any observed probabilistic dependence. And that can further be understood as asserting that for any system of probabilistic dependencies there will be some causal model which has the Markov property:[9] CCP claims that any real dependency is to be modeled by causal structure, whereas the Markov property asserts the contrapositive, that any model independence will be reflected in a real independence.

[8]To convert any deterministic system into an indeterministic system, simply remove the error terms. On the other hand, should a deterministic model need no error terms, then of course the system is indeed deterministic, but that is no bar to *representing* it with an indeterministic system having only extreme probability parameters (i.e., without the unneeded error terms).

[9]See Gyenis and Redei (2004) and Hofer-Szabo and Redei (2004) for interesting work where this question is explored mathematically. That is, it attempts to answer the question whether this principle is possibly true for all possible systems of dependencies. That question has yet to be answered positively, although the results so far are promising.

There has been some controversy over this principle, with a recent flurry of contributions, presumably encouraged by the new activity in causal discovery algorithms. All causal discovery algorithms depend upon some version of this principle. Here we canvass some of the more prominent objections to CCP.

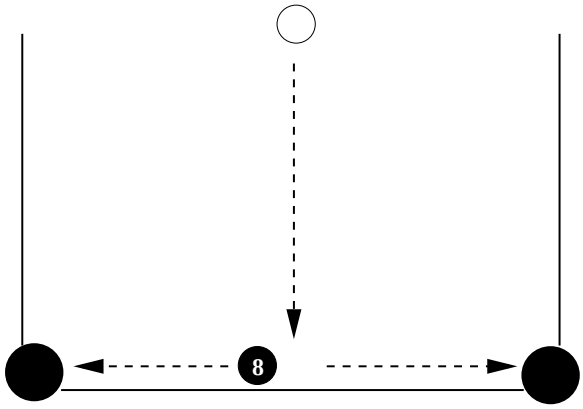

Figure 4: An interactive fork.

Wesley Salmon, a student of Reichenbach, already raised a central objection and formalized it in his concept of *interactive forks*, which have the same structure as common causes but which violate Reichenbach's probabilistic conditions by introducing dependencies between the effects which are *not* screened off by the common cause (Salmon, 1980). Salmon asked us consider a pool shot such as sketched in Figure 4. The setup is supposed to be one where the cue ball will drop the 8 ball in one corner pocket if and only if the cue ball itself drops into the other corner pocket. In that case, if that condition is perfectly instantiated, then nothing we can learn about the striking of the cue ball itself, the common cause, can screen off one effect from the other (ignoring the possibility of the common cause being deterministic). And yet, obviously, the dropping of neither ball is directly or indirectly causing the dropping of the other. We think the most plausible response to this particular example is to point out both that the condition is unlikely to be *perfectly* instantiated in an indeterministic world and that relevant refinements in the description of the common cause — such as describing the spin, momentum, etc. imparted to the cue — are available. However, Salmon's example was intended simply as an intuition-building case: the real point was that in some quantum mechanical setups, such as the Einstein-Podolsky-Rosen (EPR) paradox (Einstein et al., 1935), which has just the structure of the pool example, such refinements are unavailable (van Fraassen, 1982). Our response to this is more or less to despair. If quantum mechanics violates Reichenbach's principle and if causal discovery relies upon Reichenbach's principle, then causal discovery will be unable to cope with quantum mechanics. The only source of consolation is that humans also seem to be unable to cope with quantum mechanics: there is certainly no agreement about how to interpret quantum mechanics, and the causal interpretation of the EPR paradox in particular remains unclear. Complaints about the limits of our causal discovery algorithms which imply that our algorithms deserve credit only if they can *outperform* humans are complaints we need not take too dear. In short, we can abandon the CCP as some kind of *metaphysical* principle describing the nature of the universe, while retaining it as a fruitful *methodological* principle of inductive method.

Still, some of Salmon's cases arose from there simply being constraints on the system, such as the applicability of conservation laws. Salmon presented such an example depending upon the conservation of energy. A simple example of Cartwright (1983) is to suppose that

we are given \$10 to spend on two items in a store; then the two purchase prices are effects of a common cause, but the prices are strict functions of each other and are not screened-off by the common cause. An example Williamson (2001) offers is of a Bayesian network containing two nodes reporting on a binomial process, one the sample mean and the other the sample variance. But for binomials, expected mean and variance are strict functions of each other, so nothing will screen these variables off (assuming the binomial variable itself is not represented!). Another possible complaint, which takes the constraint-based examples to an extreme, is to point out that logically related variables may not be screened-off by common causes. For example, if one variable represents $A \vee B$ and another $A \vee C$, they will be interdependent and we may have no way of disentangling them via common causes. Or, we may model a person's sex with two binary variables *Male* and *Female*. Or, even more trivially, we may simply duplicate a variable. But, of course, Reichenbach's principle was never meant to deal with such matters; logico-semantic relations, and constraints which define the problem under study, are properly handled *prior* to any application of the CCP.

In order to rule out the introduction of variables with such illegitimate ties to other variables in a model we adopt the suggestion of Hausman and Woodward (2004) that the relevant criterion of distinctness of two variables is that it should be physically possible to intervene upon one without directly affecting the other. That is, if every possible physical intervention upon either variable affects the probability distribution over the other, then the variables lack distinctness.[10]

Principle 2 (Variable Distinction) *Every pair of variables in a causal model must have a physically possible intervention which, in some physically possible context, affects the distribution of one variable without affecting that of the other.*

The resolution of a violation of the principle would typically be to combine the offending variables in a single variable. For example, $A \vee B$ and $A \vee C$ could be replaced by a variable representing the joint states of *A, B, C* (of course, there are other possibilities). To put matters crudely, we are responding to the constraint-driven 'anomalies' to the CCP by rewriting the CCP so as to exclude them (i.e., by Lakatosian 'monster-barring'). But we think this is well within the spirit of Reichenbach's principle.

Elliott Sober championed a different kind of counterexample to the CCP (Sober, 1988): the price of bread in London and the water level in Venice have been rising jointly for a long time; we are justified in believing this to be no mere coincidence, rather that they are aspects of the development of two systems. Yet, we cannot reasonably assert that there is some causal factor ancestral to both and screening them off from each other. There is, however, a measurable variable which in fact screens the one off from the other, what we typically use to measure a system's evolution: *Time*. If we insert a measurement of time, readings off of a clock, as a common cause in the model, conditional independence is obtained. Nevertheless, from the causal interventionist point of view, there is a strong case against *Time*: namely, it is a variable upon which we cannot even in principle intervene. Obviously, pushing the clock hands around the dial will do nothing of interest to bread

[10] Hausman and Woodward (2004) claim that in the particular case of the EPR paradox the entangled states of the two space-wise separated effects are not separately intervenable in this sense and hence are not describable by distinct variables. But what Einstein et al. (1935) showed is that quantum theory requires that the two effects are not separately *observable*, rather than separately intervenable. We can certainly alter the spin states of two separated electrons independently of each other, for example by capturing one of them with a proton and forming a hydrogen atom! Of course, the states of those processes will then no longer be quantum mechanically entangled, so if we *define* the two states as entangled, their claim can be sustained.

prices or water levels, but by-passing the measuring device leaves us with no means of intervention. The objection is not simply that humans lack the physical means of intervention; after all, black holes are causal factors, yet we cannot physically intervene upon them. But black holes are themselves caused by other events, such as super-massive star formation, and these other events may be causally arranged in the form of intervention variables in an augmented model. Regardless of merely human limits, we can make sense of such a model. But what causes *Time*? Williamson (2001) points out that *Time* is too diffuse for there to be a determinate story about intervening upon it. We suspect that the idea of intervening upon *Time* is simply incoherent. But, if we cannot intervene upon *Time*, then on our definition, it cannot be a variable in a causal model.

There is another limitation of the CCP that lands somewhere between the correlations induced by system evolution and systemic constraints: two aspects of a single system will typically be coordinated with each other, so long as the single system maintains its integrity; however, it may be difficult or impossible to identify their coordination with some prior common cause. Arntzenius (1999) points out that the coordinated flight of flocks of birds provide an example of this. A conceivable common cause would be a flock-leader; but a better explanation seems to be that each bird follows an elastic rule keeping its distance between adjacent members of the flock within an interval. Thus, when one bird changes direction, its neighbors will change nearly simultaneously, resulting in coordinated group flight (Reynolds, 1987). Every organism will provide innumerable other examples. Every pair of your cells are coordinated in space-time. While we can offer events which are ancestral to both, with ease, it would only be with extraordinary, or superhuman, difficulty that we could identify some complex joint prior event which truly *screens off* the spatiotemporal location of one cell from another.

We think there is a natural two-fold response to this complaint. On the one hand, it is a kind of pedantry. If we are, say, trying to model a player's performance on the soccer pitch, we are not then also trying to model the player's individual cells, so questions about their coordination do not arise. If, however, we are modeling the mutagenic origin of cancers, then we may well be modeling individual cells and questions about the coordination of their states of health — if not their spatiotemporal locations — might be welcome. The other half of our response is to concede — we concede that causal discovery algorithms which start their work with a completed set of measurable variables are radically incomplete.[11] Not only do they need to be fitted out with mechanisms for discovering hidden variables, including hidden common causes, as Spirtes et al. (2000) point out, but rather more fundamentally they need to be fitted out with mechanisms for basic concept formation. They need to be able to decide whether two observable features belong to a common object, so that any additional explanation of their coordination becomes otiose, perhaps employing methods like those of Holland et al. (1986).

So, we agree that CCP has limits — namely the limits of our modeling and learning methods that determine what variables and arcs can be introduced into causal models. Some limits simply reflect our local modeling interests. Others reflect our understanding of what it is to model causally. For example, variables which cannot be intervened upon cannot be introduced. It follows that time-dependent correlations of the Sober-type cannot find their explanation in a common cause. Of course, it also follows that a causal discovery algorithm which forces such correlations to be explained causally will be mistaken. Our causal discovery methods, to be complete and correct, need to be enhanced with methods of concept formation and also with the meta-ability to recognize when the search for a

[11] Let alone the many which demand a complete chronological ordering of those variables!

common cause is pointless. This is not easy for humans, and so is also not likely to be easy for our machines. It may be a long time before our AI systems can do without the guidance and support of an experienced metaphysician.

6 Faithfulness and Causal Faithfulness

In Suppes' account of probabilistic causation the true causes of E are found via a two-step process: first identify the prima facie causes of E and then filter out the spurious causes (Suppes, 1970). Spurious causes are those which can be screened off by a common ancestor of the prima facie cause and the purported effect. The candidates — the **prima facie causes** — are just those event types which are positively related to the effect (ignoring an additional temporal precedence requirement); i.e., those C such that $P(C|E) - P(C) > 0$. The essence of this is the identification of potential causes by way of probabilistic dependence. Every method of automated causal discovery in some way takes probabilistic dependence as its starting point: they all assume that if all probabilistic dependencies are causally explained, then there is no more work for causality to do. In other words, they assume the *converse* of the Markov property, that causal relations cast an observable probabilistic shadow of some sort, so where there is no shadow, no causal structure should be invoked. In the case of the Markov property the question was whether it is reasonable to expect that a real independence should correspond to every implied independence in the model. That follows Reichenbach's idea that correlations are not sustained by nothing, but only by some explanatory causal structure. Here, the question is whether it is legitimate, given a lack of probabilistic dependence in the system, to assume a lack of causal relation in an explanatory model for that system. This assumption is precisely what is wrong with causal discovery, according to Nancy Cartwright (2001).

This question has generally been assumed to be equivalent to the question whether our causal models should be faithful to reality: that corresponding to every d-connection in a causal model (and so also to every arc) there must be a probabilistic dependence. Let us call this reliance on probabilistic dependence the **naive principle of faithfulness**. We go this far in endorsing the naive principle: other things being equal, we shall prefer faithful to unfaithful models.

In the simplest case, where $C \dashrightarrow E$ is the *only* causal process, it is hard to see how a presupposition of faithfulness can be contested. Such a simple causal process which leaves no probabilistic shadow is as supernatural as a probabilistic shadow cast by no process. In any case, insisting upon the possibility of an unfaithful structure between understanding and reality *here* leaves inexplicable the ability to infer the causal structure in any circumstance.

But there are many situations where 'other things' are *not* equal, where we should and must prefer an unfaithful model. One kind of unfaithfulness is where transitivity fails. If causality is somehow based upon the kind of causal processes investigated by Salmon (1984) and Dowe (2000), processes which are capable of carrying information from one space-time region to another (Salmon-Dowe processes for short), then it seems somehow causality *ought* to be transitive. While it may be unclear exactly what Salmon-Dowe processes are, and so exactly how to individuate them, it does seem clear that they are composable: when ball A strikes B and B strikes C, the subprocesses composed form a larger process from A to C. No doubt this kind of Newtonian example lies behind the widespread intuition that causality must be transitive. We share the intuition that causal processes are somehow foundational for causal relevance, and that they can be composed transitively; unfortunately for any simple analysis, causal relevance itself is not transitive. One of many examples of Hitchcock

(2001a) will make this clear: suppose there is a hiker on a mountain side and at just the wrong time a boulder dislodges and comes flying towards her; however, observing the boulder, she ducks at the right moment, and the boulder sails harmlessly past; the hiker survives. This is represented graphically in Figure 5. We are to suppose that if the boulder dislodges, the hiker will duck and survive, and that if the boulder doesn't dislodge, she will again survive. In this case, there is no sense in which the boulder makes a difference to survival. It would be perverse to say that the boulder has caused the hiker to survive, or to generalize and assert that in this and relevantly similar cases boulders cause hikers to survive. While causal processes, and their inherent transitivity, are one part of the story of causality, making a difference, probabilistic dependence, is equally part of that story, a part which here fails dramatically.

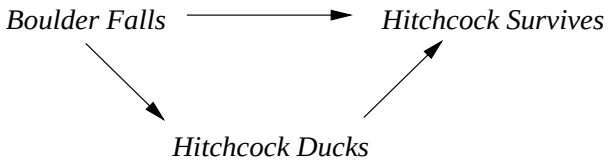

Figure 5: The hiker surviving.

Hiddleston (2004) claims that intransitivity in all cases can be attributed to the fact that there are multiple causal paths, when we look at component causal effects in isolation, such things cannot happen. He is mistaken. An example of Richard Neapolitan (2003) makes this clear: finesteride reduces DHT (a kind of testosterone) levels in rats; and low DHT can cause erectile dysfunction. However, finesteride doesn't reduce DHT levels sufficiently for erectile dysfunction to ensue (in at least one study). There is no direct relation between finesteride and erectile dysfunction. Graphically, this is simply represented by:

$$Finesteride \longrightarrow -DHT \longrightarrow Dysfunction$$

So, we can have a failure of transitivity in a simple chain.[12]

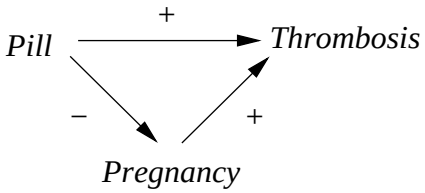

Figure 6: Neutral Hesslow.

These cases of unfaithfulness due to intransitivity pose no problems for causal discovery algorithms; they will find the intransitive causal chains so long as the separate links within them are discoverable via their distinct dependencies. Far more problematic are examples of unfaithfulness generated by multiple paths, where *individual arcs* can fail the faithfulness test. These include 'Simpson's paradox' type cases, where two variables are directly

[12]Hiddleston's mistake lies in an overly-simple account of causal power, for in linear models, and the small generalization thereof that Hiddleston addresses, causality is indeed transitive. Such models are incapable of representing threshold effects, as is required for the finesteride case. We shall consider this further in a future paper on causal power.

We note also that some would claim that all cases of intransitive causation will be eliminated in some future state of scientific understanding: by further investigation of the causal mechanisms, and consequent increase in detail in the causal model, all apparently intransitive causal chains will turn into (sets of) transitive causal chains. This may or may not be so. Regardless, it is an a posteriori claim, and one which a theory of causality should not presuppose.

related, but also indirectly related through a third variable. The difficulty is most easily seen in linear models, but generalizes to discrete models. Take a linear version of Hesslow's example of the relation between the *Pill*, *Pregnancy* and *Thrombosis* (Figure 6). In particular, suppose that the causal strengths along the two paths from *Pill* to *Thrombosis* exactly balance, so that there is no net correlation between *Pill* and *Thrombosis*. And yet, the above model, by stipulation, is the true causal model. Well, then we have a failure of faithfulness, since we have a direct causal arc from *Pill* to *Thrombosis* without any correlation wanting to be explained by it. In fact, causal discovery algorithms in this case will generally not return Figure 6, but rather the simpler model (assuming no temporal information is provided!) of Figure 7. This simpler model has all and only the probabilistic dependencies of the original, given the scenario described.

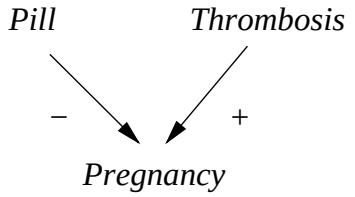

Figure 7: Faithful Hesslow.

A plausible response to this kind of example, the response of Spirtes et al. (2000), is to point out that it depends upon a precise parameterization of the model. If the parameters were even slightly different, a non-zero correlation would result, and faithfulness would be saved. In measure theory (which provides the set-theoretic foundations for probability theory) such a circumstance is described as having *measure zero* — with the implication that the probability of this circumstance arising randomly is zero (see Spirtes et al., 2000, Theorem 3.2). Accepting the possibility of zero-probability Simpson-type cases implies simply that we *can* go awry, that causal discovery is fallible. But no advocate of causal discovery can reasonably be construed as having claimed infallibility.[13] The principle at issue, naive or not, is not a metaphysical claim that faithfulness *must* always be maintained — otherwise, how could we ever come to admit that it had been violated? Rather, it is a methodological principle that, until we have good reason to come to doubt that we can find a faithful model, we should assume that we can. In the thrombosis case, with precisely counterbalancing paths, we know the true model is not faithful, so we are not obliged to adhere to the naive principle of faithfulness.

Cartwright objects to this kind of saving maneuver. She claims that the 'measure zero' cases are far more common than this argument suggests. In particular, she points out that many systems we wish to understand are artificial, rather than natural, and that in many of these we specifically want to cancel out deleterious effects. In such cases we can anticipate that the canceling out will be done by introducing third variables associated with both cause and effect, and so introducing *by design* a 'measure-zero' case. In addition, there are many natural cases involving negative feedback where we might expect an equilibrium to be reached in which an approximate probabilistic independency is achieved. For example, suppose that in some community the use of sun screen is observed to be unrelated to skin cancer. Yet the possible causal explanation that sun screen is simply ineffective may be implausible. A more likely causal explanation could be that there is a feedback process such that the people using the sun screen expose themselves to more sunlight, since their

[13]Cartwright notwithstanding: "Bayes-net methods... will bootstrap from facts about dependencies and independencies to causal hypotheses—and, claim the advocates, *never get it wrong*" (Cartwright, 2001, p. 254; italics ours). Here, Cartwright's straw-man has it wrong.

skin takes longer to burn. If the people modulate their use of sun screen according to their exposure to the sun, then their total UV exposure would remain the same. Again, Steel (2004) has pointed out that there are many cases of biological redundancy in DNA, such that if the allele at one locus is mutated, the genetic character will still be expressed due to a backup allele; in all such cases the mutation and the genetic expression will fail the faithfulness test. As Steel emphasizes, the point of all these cases is that the measure-zero premise fails to imply the probability zero conclusion: the system parameters have not been generated 'at random' but under intentional or evolutionary control, leading to unfaithfulness.

If these cases posed some insurmountable burden for causal discovery algorithms, this would surely justify the skepticism of Cartwright and others, for obviously we humans have no insurmountable difficulties in learning that the sun causes skin cancer, etc., even if these relations are also not easy to learn. But, regardless of the frequency or importance of these cases, they all concern only the naive principle of faithfulness.

The **sophisticated principle of faithfulness**, on the other hand, is this: causal models satisfy *causal faithfulness*, rather than simple faithfulness. We offer the following definition:

Definition 2 (Causal Faithfulness) *A model M is causally faithful to a system if and only if its fully augmented model M' is faithful to the fully augmented system under perfect interventions.*

A **fully augmented** model is one in which each original variable has an intervention variable added.[14]

Causal faithfulness is not tested by the probabilistic dependencies demanded by the naive principle. Indeed, not even probabilistic dependency under intervention is demanded by the naive principle, although that also would not suffice. In the neutral Hesslow case, for example, intervention by forcing the pill on subjects would have no net effect on thrombosis. This does not mean, however, that the neutral Hesslow model is causally unfaithful. The causal faithfulness test, unlike the original test over ordinary probabilistic dependencies, corresponds to what we can observe under all possible experimental interventions. In any fully augmented model every endogenous variable will become the center of at least one new collider. Colliders, together with undirected arc structure, determine statistical distinguishability (Verma and Pearl, 1991); that is, by introducing new colliders augmentation enables new conditional dependency tests and so new opportunities to distinguish empirically between the faithful and faithless causal models.[15] For example, under imperfect interventions the fully augmented true Hesslow model (top, Figure 8) implies a dependency between *Pill* and I_T, given a fixed value for *Thrombosis*, which the fully augmented faithful model (bottom, Figure 8) does not. Some may find this an odd case to think about, where we are unsure whether an intervention has been made; but we can surely be unsure about interventions made by *others*. In any case, we can also consider perfect interventions. Under perfect intervention, the named test will fail, since I_T will fully determine the value of *Thrombosis*. But in that case an intervention on *Pregnancy*

[14] Of course, we could recursively continue the process since the intervention variables must themselves be intervenable. No such recursion is required to make sense of causal faithfulness.

[15] Spirtes et al. (2000) also considered fully augmented models in the context of statistical indistinguishability. In particular, they showed (in effect, via a minor extension of their Theorem 4.6) that any two distinct models (and so also those which are statistically indistinguishable) will, when fully augmented, be statistically distinguishable. In their language, no two distinct models are 'rigidly' statistically indistinguishable.

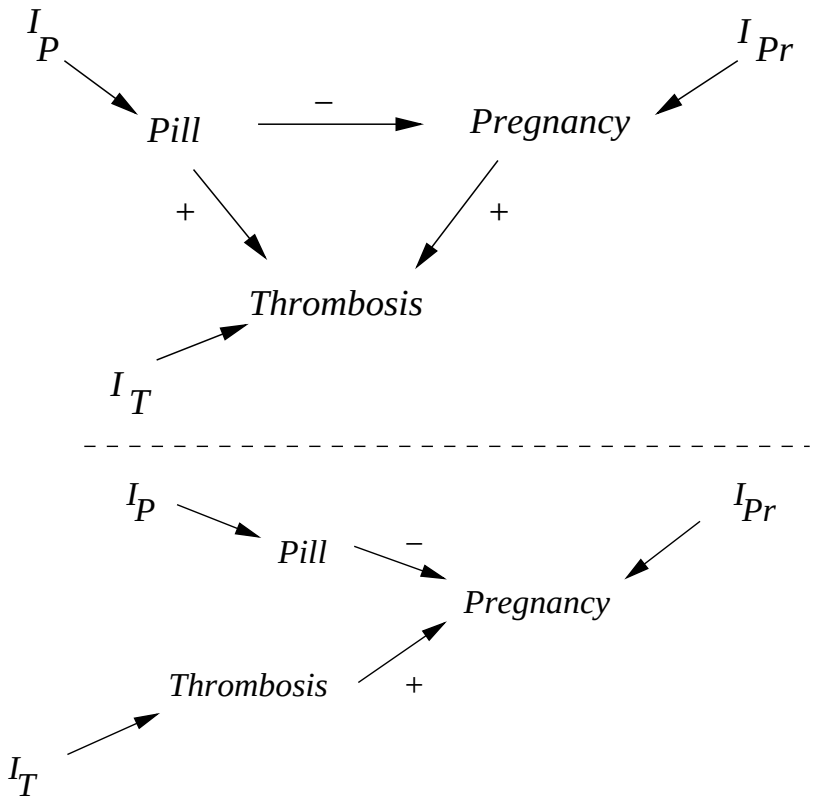

Figure 8: Fully augmented Hesslow models: the faithless, but true, causal model (top); the faithful, but false, model (bottom).

will induce a dependency between *Pill* and *Thrombosis* in the true model, but not in the simpler model. Such discrepancies between implied dependencies can be tested for directly.

In summary, augmenting models works statistically by introducing new conditional dependency structures. When a faithless and a faithful model coincide in the probability distributions which they induce over the original variables, they will not coincide in the probability distributions they induce over the augmented set of variables. The proof of this for perfect interventions (Theorem 1) and for imperfect interventions (Theorem 2) upon linear models is in Appendix B. Since the causal model, by stipulation (by being the source of the data!), is the true model, the data will favor the causally faithful model, by showing a dependency structure which the simpler model cannot represent or cannot easily represent.

There are two ways in which faithfulness can fail: an isolated chain can fail due to non-linearities, such as threshold effects; or, a combination of causal paths can lead to cancellations. The first case is non-problematic, in the sense that there is no simpler model which is causally faithful to confuse the learning process. The second case certainly can lead to a confused learning process, but the confusion can be removed by augmenting the data, leading to disambiguation of which model, the faithless original and the faithful imposter, is the true source of the data.

Faithfulness is a property tested by *observed* probabilistic dependencies — i.e., dependencies under samples of observed variables. Causal faithfulness is a property tested by experimental interventions and experimental data. The common restriction to observational methods is perfectly understandable in the applied AI literature, where observational data come more readily to hand. Even there the restrictions are slowly being relaxed, as experimental methods and their data begin to be taken more seriously (e.g., Tong and Koller, 2001; Murphy, 2001). However, disputes and arguments which depend upon an in-

principle restriction to observational data just miss the point: causal modeling is actually about causality, and causality is actually about intervention. When we allow our algorithms to benefit from the same kind of information-gathering processes from which we benefit, in particular from experimental data, the objections of Cartwright and others that true models are often unfaithful evaporate.

Returning to our conjectured relation between simplicity and causal structure (Conjecture 1), we note that the neutral Hesslow case and others like it render the conjecture untenable: the simplest model for representing the probabilities is often wrong. But that does not mean that the simplest causally faithful model is often wrong, or even ever wrong. All of these considerations suggest a revised conjecture:

Conjecture 2 (Augmented Causal Simplicity) *A causally faithful model for a physical system is always the structurally simplest among those capable of representing the system's fully augmented probability distribution.*[16]

7 Causal Relevance

The story thus far is that causal models — that is, causally interpreted Bayesian networks — are plausible candidate representations of causal systems. In order to ensure that our models are causal, we must enforce certain model-building principles which exclude relationships between variables that are non-causal. Furthermore, we need to be prepared to employ experimental data in order to learn our causal models. The question now arises, given that we have a causal model, how do we identify the causal relationships which such a model represents? How do we specify criteria for causality relative to that model?

7.1 Relevance, Role and Process

A preliminary distinction, emphasized by Hausman (2005), should be drawn between causal relevance and causal role. "Smoking causes cancer" is normally an assertion of **causal role**, that smoking *raises* the probability of cancer. "Exercise prevents heart disease" is also a role assertion, in the opposite direction. Causal role attributions are always of a causal factor (variable) taking a specific value or set of values. The variable *Smoking*, considered across all possible values from, say, zero cigarettes a day to some possible maximum, does not, of course, raise the probability of cancer. Raising (lowering) an effect has to be in contrast to some state which does not raise (lower) the effect. If we wish to generalize from a variable having some specific causal roles to a variable having any causal role, then we are talking about the **causal relevance** of the variables which play those roles under different values. C will be causally relevant to E if, under some circumstances (made precise below), setting C to a particular value will either raise or lower the probability of E taking some value.[17]

A causal model encodes information about both role and relevance, of course. And the distinctions between causal role and relevance and that between type and token causality are orthogonal, so we could consider any of the quadrants in the table:

[16] The story to and beyond Conjecture 2 represents our train of thought and motivates the investigations of this paper. We do not, however, attempt to resolve the status of this Conjecture in this paper; that is the subject of another paper in press.

[17] Note that, although the concept of relevance in general is symmetric, we are here using 'causal relevance' asymmetrically for type causality *directed* from a cause to an effect.

	Type	Token
Relevance	Causal Factor	Instantiated Factor
Role	Promotion (Prevention)	Attribution, Blame (Accident, ± Fortune)

However, we shall focus upon the top row, developing candidate criteria for type and token causal relevance; in effect, they are our proposals for how to decode a causal model's representations of causal relevance. We will not provide any explicit criterion for causal role. Assertions about role end up involving questions about conversational pragmatics, which we prefer not to address. We shall nevertheless see examples of this, since many of the intuitive examples raised in the literature concern issues of token causal role. And many of these raise intuitions of intention, blame and responsibility, which further complicate matters. Thus, our concern in this paper will *not* be to provide definite accounts of these examples. Nevertheless, we do want to be satisfied that our criteria can somehow accommodate them, that they can be articulated with future pragmatic, ethical, legal, etc. theories which deal with the examples.

7.2 Dependence and Process Accounts

Two main approaches dominate the recent history of attempts to come to grips with the notion of causality. One is the attempt to locate a supervenience base for causal relationships in an underlying metaphysics of process, initiated by Salmon (1984) and furthered by Dowe (2000). Processes are contiguous regions of space extended through some time interval — i.e., spacetime 'worms'. Of course, they can't be just any such slice of spacetime; most such slices are causal junk (Kitcher, 1989). The Salmon-Dowe research program is largely aimed at coming up with clear criteria that rule out junk, but rule in processes which can sustain causal relationships. Intuitively, we can say legitimate processes are those which can carry information from one spacetime region to another ('mark transmission' is what Salmon called this; Dowe calls it 'conserving physical quantities'). Examples are ordinary objects (balls carry around their scratches) and ordinary processes (recipes carry their mistakes through to the end). Non-examples are pseudo-processes and pseudo-objects (e.g., Platonic objects, shadows, the Void of Lewis, 2004). Hitchcock (2004b) rightly points out that, thus far, this account leaves the metaphysics of causal processes unclear. The Salmon-Dowe research program is incomplete. But we know of no reason to believe it is not completable, so for purposes of our discussion we shall describe as 'Salmon-Dowe processes' those which fall under some future completed analysis of this type.

If it is causal processes which ground the probabilistic dependencies between variables, then it must be possible to put the variables within a single model into relation with one another via such processes. This suggests a natural criterion of relevance to require of variables within a single model:

Principle 3 (Causal Processes) *If two variables appear in a causal model, there must be a sequence of possible or actual causal processes connecting them.*

This contrasts with Hitchcock (2001a), who vaguely requires that pairs of variables not "be too remote" from each other. Note that we do not demand a possible sequence of causal processes between any two variables, but a sequence of possible processes: it may be, for

example, that two events are spacewise separated, yet mediated by a common third event. Nor, of course, do we demand actual processes between any event types in the model. Probabilistic dependency is founded upon possibilities, realized and unrealized.[18]

The second prominent approach to the philosophy of causation that we draw upon is the attempt to find criteria for causal relations in some form of probabilistic dependence, particularly in the 'probabilistic causality' camp of Reichenbach (1956), Suppes (1970), Salmon (1980) and also in the counterfactual dependency analysis of Lewis (1986, 2000). Roughly, probabilistic causality identifies C causing E with C raising the probability of E under some further conditions, such as not being screened off by a common cause. Lewis's counterfactual dependency theory asserts that C causes E if and only if had C not occurred, E would not have occurred. The counterfactual has to be assessed without 'backtracking' — that is, when considering the circumstance where C counterfactually fails, we are not allowed to go backwards in time and explain this failure by considering counterfactual values for ancestors of C. If we did allow that, then this criterion would countenance the assertion that C causes E if and only if it countenanced the assertion that E causes C.

Both of these accounts are oriented towards accommodating questions of causal role, the first explicitly and the second by implication, since it addresses causal relations between events (the values of variables) rather than between the variables as such. Regardless of this orientation, the ideas also inspire our account of causal relevance.

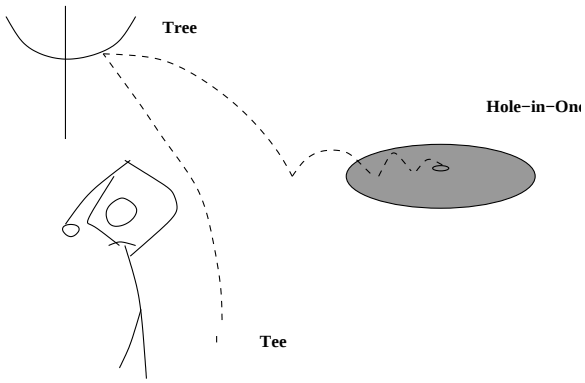

Figure 9: Rosen's hole-in-one.

The two approaches, dependency and process, have disparate strengths and weaknesses. This disparity has led some to suggest that there is no one concept of causality and that attempts to provide a unified account are confused.[19] While we agree that there may well be various distinct concepts of causality, we are unconvinced that the particular disparity between the dependence and process analyses argues for two concepts of causality: rather, we believe the disparity argues for a unification of the two analyses, a unification that uses the strengths of the one to combat the weaknesses of the other.

Dependency accounts characteristically have difficulties dealing with negative relevance, that is, causes (promoters, in role language) which in some token cases are negatively relevant to the effect (i.e., prevent it), or vice versa. Deborah Rosen (1978) produced a nice

[18]That there are causal processes behind the arcs of causal models suggests the answer to one of the concerns about causal modeling that Nancy Cartwright raises, namely that causal reality may not be made up of discrete token events, but perhaps continuous processes instead (Cartwright, 2001). Well, we expect that reality is made up of token processes, whether discrete or continuous. Discrete Bayesian networks are a convenient way of modeling them, and the variables we choose are convenient and useful abstractions. They need to be tied to the underlying reality in certain ways — and we suggest Principle 3 as one of them — but they certainly do not need to be exhaustive descriptions of that reality.

[19]Hitchcock has suggested this, e.g., in Hitchcock (2004a, b); see also Hall (2004).

example of this in response to Suppes (1970). In Figure 9 Rosen has struck a hole-in-one, but in an abnormal way. In particular, by hooking into the tree, she has *lowered* her chance of holing the ball, and yet this very chance-lowering event is the proximal cause of her getting the hole-in-one. The only hope of salvaging probability-raising here, something which all of the dependency accounts mentioned above wanted, is to refine the reference class from that of simply striking the tree to something like striking the tree with a particular spin, momentum, with the tree surface at some exact angle, with such-and-such wind conditions, etc. But the idea that we can ever refine this reference class in enough detail to recover a chance-raising reference class is far-fetched. It is what Salmon (1980) described as *pseudo-deterministic faith.*[20] In any case, as Salmon also pointed out, we can always generate chance-lowering causes in games, or find them in quantum-mechanical scenarios, where there is no option for refinement. Take Salmon's 'cascade', where a quantum-mechanical system undergoes state changes with the probabilities indicated in Figure 10. Every time such a system reaches state d via state c rather than state b, it has done so through a chance-lowering event of transition into state c. By construction (in case this is a game, by nature otherwise) there is no refinement of the intermediate state which would make the final transition to state d more probable than the untaken alternative path through b; hence, probability-raising alone cannot account for causality here.

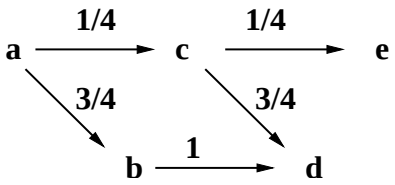

Figure 10: Salmon's quantum-mechanical cascade.

Salmon's way out was to bite the bullet: he asserted that the best we can do in such cases is to locate the event we wish to explain causally (transition to d) in an objectively homogeneous reference class. If that reference class happens to raise the probability of the outcome relative to its alternatives, then we have a contrastive causal explanation — that is, we can account for token causal role by reference to the type causal role. But, if it doesn't raise the probability, then there is no contrastive explanation. Insisting on the *universal* availability of contrastive explanations is tantamount to the pseudo-deterministic faith he denounced (Salmon, 1984, chapter 4).

Of course, biting the bullet here doesn't answer the question, Given the lack of a contrastive explanation, why are we inclined to believe that the transition from a to c to d counts as causal? The answer Salmon (1984) gave was that we have a Salmon-Dowe causal process leading from each state to the next. This seems the only available move for retaining irreducible state transitions within the causal order.

Assuming, per above, that the metaphysics of process has been completed, the problem remains for Salmon's move that it is insufficient. For one thing, as we saw above, causal processes are composable: if we can carry information from one end to the other along two processes, then if we connect the processes, we can carry the (or, at any rate, some)

[20]Note that the escape by contrasting striking the tree with missing it fails on at least two counts. Of course, missing the tree, given the hook, is a contrast class with a *lower* probability of success than hitting the tree. But we are attempting to understand causality *relative* to a given causal model. And this maneuver introduces a new variable, namely *how* the ball is hit (or, perhaps, its general direction), so the maneuver is strictly evasive. Secondly, if we are going to countenance new variables, we can just introduce a local rule for this hole: just behind the tree is a large net; landing in the net also counts as holing the ball.

information from the composite beginning to the composite end. But the many cases of end-to-end probabilistic independency need to be accommodated; the possibility of causal intransitivity needs to be compatible with our criteria of causality. Hence, invoking causal processes cannot suffice.

The over-inclusiveness of a pure causal process account is far broader than problems with intransitivity. Whereas contrastive probability (probability raising) clearly itself is too weak a criterion, missing minimally every least probable outcome within a non-trivial range of outcomes, simply invoking causal process is clearly too strong a criterion. In some sense the Holists are right that everything is connected to everything else; at any rate, everything within a lightcone of something else is likely to have a causal process or potential process relating the two. But while it makes sense to assert that the sun causes skin cancer, it makes less sense to say that the sun causes recovery from skin cancer. Yet from the sun stream causal processes to all such events, indeed to every event on Earth. Salmon's account of 1984 lacked distinction.

It is only in adding back probabilistic dependencies that we can find the lacking distinction. Positive dependencies, of course, have difficulties dealing with negative relevance; processes do not. Processes alone cannot distinguish relevant from irrelevant connections; probabilistic dependencies can. Plausibly what is wanted is an account of causal relevance in terms of processes-which-make-a-relevant-probabilistic-difference.[21] Here we present a criterion which does just this. We present it in brief, since we have recently presented it elsewhere in more detail (Twardy and Korb, 2004), but we extend those ideas here by applying the combination of dependence and process to token causation, in §8.

7.3 A Criterion of Type Causality

Our criterion is for C being type causally relevant to E relative to an observational context O specified for a causal model M (or, simply, relative to M/O, where O are variables of M held to some fixed values). An obvious initial attempt is to ask what happens probabilistically when we *intervene upon* the causal variable C. In experimental science, one concern with such a test is that the experimental intervention may be correlated with other causes of E, confounding the result. Such concerns lead to randomized designs, blind controls, etc. Since we have here *defined* interventions as exogenous, we have ruled out such confusing possibilities in advance. However, other causes of E may well *interact* with C in arbitrary ways. One such possibility is that in some background contexts an intervention setting $C = c$ increases the probability that $E = e$ while others reduce that probability, resulting in no net effect. Yet we do not want to say that C is therefore not causally relevant to E. As Hitchcock (2001b) emphasizes, there is a distinction between net effects and component effects, but in analysis we should certainly give preference to component effects since they are the building blocks of net effects. So, we should be considering causal relevance in distinct objectively homogeneous background contexts separately. Following Eells and Sober (1983), the plausible approach to identifying such contexts is to fix all variables other than C at the time of intervention upon C to some given value; the result we subsequently read out at E will be the causal effect of C relative to a homogeneous background $O' \supset O$.

Unfortunately, as has been clear for some time now, this is too simple a picture. Multiple causal paths relating C and E can, even in homogeneous contexts, result in hiding the effect

[21] Menzies' "Difference-making in Context" (2004) presents an analysis in many ways anticipating ours, not just the connecting requirement of intersecting processes, but also, for example, the relativity of causality to context and model.

of C on E, as Cartwright (1989) pointed out. The neutral version of Hesslow's example (Figure 6) suffices for this. In neutral Hesslow any intervention on *Pill* will have a null effect on *Thrombosis*, despite its causing *Thrombosis*.

What is required in judging whether C is type causally relevant to E in M/O is that we fix not only a *background* context, but also a *foreground* context. In particular, we examine each distinct causal path from C to E independently of the others — this indeed is what Hitchcock means by looking at component effects. So: if any individual causal path carries probabilistic dependence from interventions on C to E given M/R, where R extends O to an objectively homogeneous context where all but one causal path between C and E is blocked,[22] then we have established causal relevance. Formally, we can describe our criterion as:

Wiggle Logic: C is causally relevant to E with respect to M/O if and only if

$$\exists R \supseteq O \text{ s.t. :} \quad \exists! \Phi \in \text{Paths}(C, E) \text{ s.t. } \Phi \text{ is active in } M/R \text{ and } \exists c \; P_{M/R}(E|I_{C=c}) \neq P_{M/R}(E)$$

where O provides an objectively homogeneous context for C with respect to E and where $\exists! \Phi \in \text{Paths}(C, E)$ means there is a unique directed path Φ from C to E (see Figure 11). This uniqueness is enforced via R.[23]

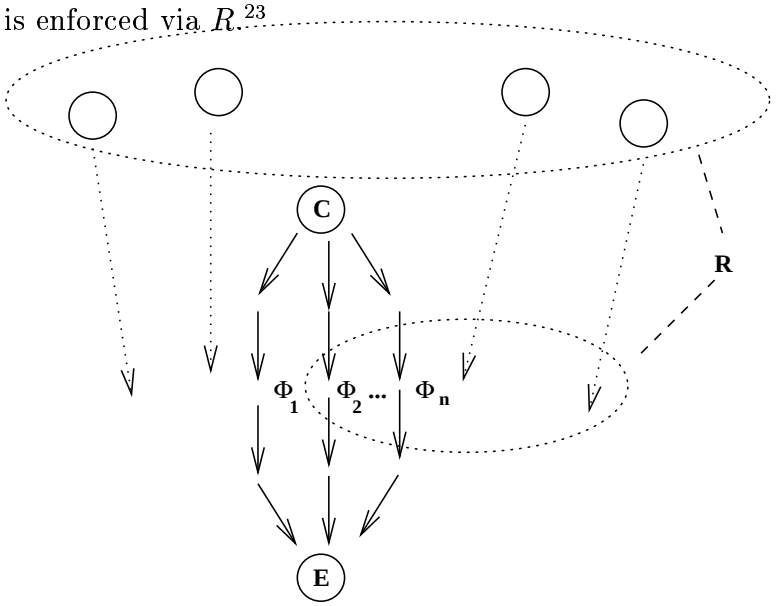

Figure 11: Wiggle Logic. R fixes a background and foreground wherein exactly one path allows an intervention on C to make a probabilistic difference.

Our type causality criterion, then, just combines causal paths with probabilistic dependence under intervention, but in such a way that any neutralizing alternative pathways are discounted.

Note that our Wiggle Logic criterion is clearly aimed at causal relevance, rather than causal role. No attention is paid to the question of whether the probability of an effect goes up or down. This is, we claim, as it should be. Type causality is fundamentally a relation

[22] Details of how to do this are in Twardy and Korb (2004).

[23] Note that this criterion does not deny causal relevance to variables connected by multiple paths; it will deny relevance in such cases only when the end-to-end probabilistic dependence is not decomposable.

between types of events, and event types are better thought of in terms of the different ways in which a variable might be instantiated, rather than in terms of propositional attitudes. In particular, there may be many, even uncountably many, potential ways of describing the state of a causal process. Concerns about causal role inevitably constrain one to thinking simply about some preferred set of states and its complement set — but the black-and-white view of reality is not normally fundamental.

Again, our account is also clearly aimed at component causal effects in the first instance, rather than at net effects. As many have pointed out, considering net effects is important for decision making when the information needed to identify a particular effect's objectively homogeneous reference class may be unavailable, in which case we are doing the best we can by considering net effects across some epistemically homogeneous reference class. However, the metaphysical basis for such decision making is one of causal paths and component effects.[24]

There has been a fair amount of agitation over the possibility, or not, of the causal efficacy of absences. Some become concerned when causal efficacy is attributed to an absence — say, when Mr Dolittle's recovery from lung disease is attributed to his cessation of smoking. Perhaps the worry is that, while smoking seems to be a determinate type of event, not smoking can be done in too many distinct ways to be considered a natural kind. But what the natural kinds may be is here irrelevant. What is relevant is the causal power of the attribution. Either Dolittle's behavior is being claimed to be a member of some specific objectively homogeneous reference class of behaviors which has causal power for recovery (which seems an unlikely gloss in this case) or (here more likely) the claim is based upon epistemic homogeneity and net causal effects. Describing the reference class in terms of an absence of smoking makes it no more suspect than the reference class of smoking, which itself is a composite reference class. That the reference class for any particular Dolittle could in principle be identified more precisely, and perhaps even to objective homogeneity, does not undermine the meaningfulness of ascribing causal efficacy to the absence of smoking; on the contrary, it underwrites that meaningfulness. Lewis also emphasized that absences (or negative events, or voids) are causally efficacious (Lewis, 2004). Unfortunately, he concluded from this that causality is not a relation, since relations require relata and an absence implies the non-existence of a relatum. However, under a causal attribution, a variable fails to instantiate some value, or range of values, only by instantiating some *other* value and not by, say, remaining uninstantiated.[25]

Absence, then, is short-hand for a net causal factor. In a causal theory for the omniscient, no doubt they could be done away with, since net causal factors could be done away with, given the omnipresent accessibility of objectively homogeneous reference classes. Causal theory for the rest of us should not be so demanding, however.

8 Token Causality

Type causality and token causality are distinct. This is the settled verdict of the probabilistic causality research program, and certainly a reasonable one. Smoking causes cancer in general, but in many particular cases there will be no cancer. But equally clearly type and token causality are related, at least roughly in the form of general to particular. Most of the

[24] We plan to give an account of causal power in a future paper. In such an account we shall need to generalize across all composites of component effects.

[25] To be sure, in some token cases it is the causal process itself which is absent, rather than a particular description of an existing process. We adjourn this discussion to the next section.

thought experiments which philosophers of causality have employed to support one theory or undermine another are examples of token causality, which is natural since our better intuitions are about concrete instances rather than abstract concepts. Here we present a particularization of our account of type causal relevance, which puts type and token causality into the relation of general to particular,[26] and which does as good a job as any we know of in handling the common examples. We will develop our account by contrast with those of Christopher Hitchcock (2001a) and David Lewis (1973, 1986, 2000, 2004).[27]

Consider again the case of Hitchcock's hiker (Figure 5). Clearly what we want to say is that boulders do cause death in such circumstances, if only because human responses are fallible, so the type relations are right in that model — each arc corresponds to a causal relevance that manifests itself in a probabilistic dependency under fitting Wiggle Logic.[28] But in the particular case — to be sure, idealistically (deterministically) described — the boulder's fall does not affect survival in any way, because there is no probabilistic dependency between the two.

Hitchcock (2001a) describes two plausible criteria for token (actual) causality. Both of them look at component effects, by isolating some causal path of interest. The first is very simple. Let's call it **H1** (following Hiddleston, 2004).

> **H1:** $C = c$ **actually caused** $E = e$ if and only if both $C = c$ and $E = e$ occurred and when we iterate through all $\Phi_i \in \text{Paths}(C, E)$, for some such Φ_i if we block all the alternative paths by fixing them at their actually observed values, there is a probabilistic dependence between C and E.

In application to Hitchcock's hiker's survival, this works perfectly. Considering the direct path *Boulder* → *Survival*, we must fix *Duck* at true, when there is no probabilistic dependency. The second path (through *Duck*) doesn't need to be considered, since there is no variable mediating the path alternative to it, so there is no question of blocking it.

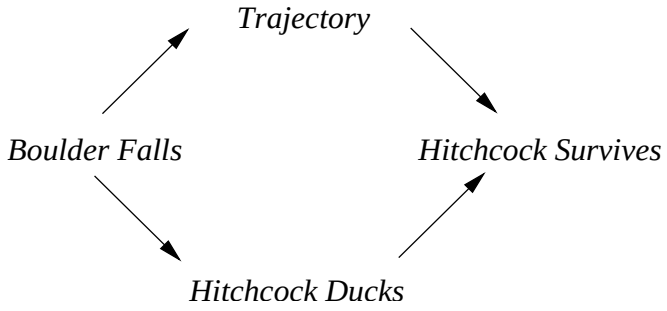

Figure 12: The hiker surviving some more.

The second path could, of course, be considered if we embed the model of Figure 5 in a larger model with a variable that mediates *Boulder* and *Survival*. We could model the trajectory of the boulder, giving us Figure 12. In this case, H1 appears to get the wrong answer, since we are now obliged to fix *Trajectory* and discover that there is now a probabilistic dependency between *Boulder* and *Survival*. In particular, if the boulder *doesn't* fall, but somehow regardless achieves its original trajectory, then the hiker won't have ducked

[26] With an exception in the treatment of context, discussed below.

[27] Hitchcock's is a simplification of Halpern and Pearl (2005), which is arguably superior in some ways but more complex than we care to deal with here. The differences between our account and Hitchcock's carry through transitively to that of Halpern and Pearl.

[28] Of course, we would never say "Boulders falling cause survival." But that's because in our speech acts causal role ordinarily leaks into causal attributions. We are not here interested in a theory which generates all and only ordinary language utterances about causality.

and will end up dead. Hitchcock's response to this possibility is to say that the introduction of *Trajectory* requires a 'sophisticated philosophical imagination' — we have to be able to imagine the boulder miraculously appearing on collision course without any of the usual preliminaries, such as being dislodged — and so an account of actual causation for the ordinary world needn't be concerned with it. Hiddleston objects to this as an ad hoc maneuver: he suspects that variables will be called miraculous when and only when they cause trouble for our analysis. However, he is mistaken. Our Intervention Principle makes perfectly good sense of Hitchcock's response. Either *Trajectory* is intervenable (independently of *Boulder*) or it is not. If it is not, then modeling it is a mistake, and H1's verdict in that case is irrelevant. If it is intervenable, then there must be a possible causal process for manipulating its value. A possible example would be: build a shunt aimed at the hiker through which we can let fly another boulder. For the purposes of the story, we can keep it camouflaged, so the hiker has no chance to react to it. All of this is possible, or near enough. But it also shows itself to be entirely irrelevant: in order to introduce this variable, making it intervenable, we have to alter the original story in dramatic ways, so that it is no longer recognizable as the original story. Hitchcock's criterion, just as much as ours, is model-relative. The fact that it gives different answers to different models is unsurprising; the only relevant question is what answer it gives to the right model.

This last consideration was something of an aside, but it reveals some of the useful work our model-building principles do in accounting for actual causation, even before considering the details of any explicit criterion.

H1 handles a variety of examples without difficulty. For example, it copes with the ordinary cases of pre-emption which cause problems for dependency theories. Thus, in Figure 13 if the supervisor fires at the victim if and only if the trainee assassin doesn't fire and, idealistically again, neither the trainee nor supervisor can miss, then an account requiring end-to-end dependency, such as Lewis's original counterfactual analysis of causation (Lewis, 1973), fails. In particular, should the trainee fire, this action will not be considered the cause of the victim's death, since there is no dependency. Lewis turned to a step-wise dependency of states of the bullet as it traverses the distance to the victim. Although there is no end-to-end dependency, if we take the transitive closure of step-by-step dependencies, we find end-to-end causation. We find this objectionable on two counts: first, as we have seen, causation is not transitive; second, finding the intermediate dependencies requires generating intermediate variables, and so altering the causal story in unacceptable ways. Hitchcock's H1, on the other hand, has it easy here: we simply observe that the supervisor did not fire and that under this circumstance there is a dependency between the trainee's action and the victim's health.

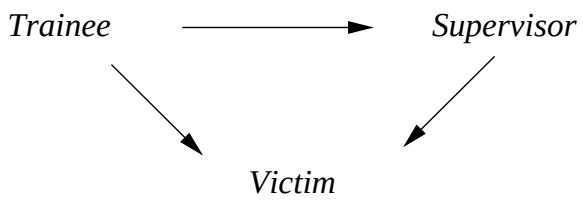

Figure 13: Pre-emptive assassination.

Lewis called the assassin case pre-emption by 'early cutting'. Pre-emption can also occur through late cutting. If Billy and Suzy are each throwing a rock at a bottle (and, as usual, they cannot miss) and if Suzy throws slightly earlier than Billy, then Suzy causes the bottle to shatter and Billy does not (see Figure 14), again despite the fact that there

is no end-to-end dependency. In this case, however, there is also no step-wise dependency for Lewis to draw upon: at the very last step, where the bottle shatters, the dependency will always fail, because Billy's rock is on its way. To deal with this Lewis turned to 'quasi-dependence' (Lewis, 1986). By examining the event chain from Suzy's throw to the bottle smashing under *counterfactual* circumstances surrounding it, in particular when Billy's throw is missing, then the *intrinsic* properties of the chain must be unchanged. If Suzy's throw in one case is causal, then it must be so in the other, since Lewis assumed that causality is an intrinsic property of such chains. But, clearly, the throw is causal absent Billy, so it must also be causal given Billy. Although there is no step-wise dependence in that case, Lewis concluded that the counterfactual step-wise dependence, quasi-dependence, sufficed for causality.

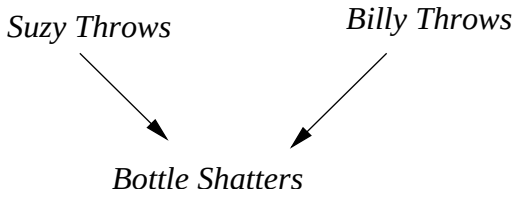

Figure 14: Pre-emptive bottle smashing.

Hitchcock's H1 also fails with Suzy's throw, and for the same reason, that the dependency fails under the actual circumstances; and Hitchcock also resorts to counterfactual contexts to cope (in this he is following the account of Halpern and Pearl, 2005). The idea appears to be the same as that of Lewis: using non-factual circumstances to reveal intrinsic causal properties. However, Hitchcock is more exact about the circumstances that are allowable. In particular, he will not allow any counterfactual circumstances which would change the values of any of the variables on the causal path under consideration. Any variable off that path will have a range of values which have no impact on the causal path, minimally that value which it actually took. Such values are said to be in the **redundancy range** (RR) for that path. Then the new criterion, **H2**, is:

> **H2:** $C = c$ **actually caused** $E = e$ if and only if both $C = c$ and $E = e$ occurred and when we iterate through all $\Phi_i \in \text{Paths}(C, E)$, for some such Φ_i there is a set of variables $\mathbf{W}$ s.t. when fixed at values in their redundancy ranges relative to Φ_i, there is a probabilistic dependence between C and E.

Since actual values are trivially within the RR, the prior (positive) successes of H1 remain successes for H2. With Suzy and Billy, it's clear that Billy's throwing or not are both within the redundancy range, and the dependency upon Suzy's throw reappears when we consider what happens when Billy's throw is absent. This seems a very tidy solution.

Unfortunately, Hiddleston (2004) raises a simple objection to which we know no decisive rebuttal. We have here a criterion which is counting $C \longrightarrow E$ as a path bearing actual causation because of what would be happening under *non-actual* circumstances. Lewis was surely right that counterfactual contexts that don't impact upon a causal chain cannot alter the intrinsic properties of that chain, but it is far from clear that causation is actually intrinsic. Lewis (2004) himself pointed out that both counterfactual and at least some causal relationships are extrinsic, being dependent upon other features of the world (as, for example, in the Backup Fielder case of the next section). Whether or not there is some fundamental, metaphysically hard-core, intrinsic supervenience base for causation, which some call 'biff' for short, many causal relations built upon biff are not that, but are engaged with the rest of the world, and so supervene on more than biff. And these broader

causal relations are what causal discovery, and our analysis, is aimed at. So, Lewis's earlier argument for the introduction of counterfactual circumstances is unsatisfactory. Hitchcock has no tale to tell beyond the success, or lack of it, of the criterion offered (and likewise Halpern and Pearl).

However, Hiddleston offers an example which H2 cannot handle, as usual an example concerning potential violent death. Suppose the king's guard, fearing an attack, pours an antidote to poison in the king's coffee. The assassin, however, fails to make an appearance; there is no poison in the coffee. The king drinks his coffee and survives. Did the antidote cause the king to survive? That is no more plausible than the claim that the boulder falling has caused the hiker to survive; however, H2 makes this claim, since *Poison* being true is in the redundancy range.[29] Interestingly, H1 gets this story right, since the poison is then forced to be absent, when the dependency of survival on antidote goes away. Hiddleston suggests that H1 was just the right criterion all along, but needed to be supplemented with Patricia Cheng's (and Clark Glymour's) theory of causal models and process theory (Cheng, 1997; Glymour, 1998, 2002). We agree with the general idea: examining dependencies under actual instantiations of context variables is the right way to approach actual causality. Cheng's causal model theory, however, is far too restrictive. As the precursor to applying a token causality criterion we need a full causal model theory (incorporating process theory in its model-building rules), whereas Cheng's models are minor extensions to linear models and so too weak to represent the varieties of causation we are interested in.

8.1 An Algorithm for Assessing Token Causation

We now present our alternative account of actual causation in the form of an 'algorithm' for assessing whether $C = c$ actually caused $E = e$, given that both events occurred. Our steps are hardly computationally primitive, but opportunities for refining them and making them clearer are surely available.

Step 1 Build the right causal model M.

Of course, this is a complicated step, possibly involving causal discovery, expert consultation, advancing the science of relevant domains, and so forth. All of the model-building rules presented in this paper (before and after this point) apply. More specifically, our account of type causal relevance applies, providing the test for whether a causal model is the right one. So, as earlier suggested, this account of token causality starts from the type causal model and gets more specific from there.

As we have already made clear, this step (applying the model-building rules) circumvents a number of problems that have arisen in the literature. For example, we know that *Time* should not be invoked as a variable and that problem-defining constraints should likewise be excluded. We also know that the imaginative introduction of intermediate events to save some kind of step-wise dependency across a causal chain is (normally) illegitimate. So, despite being open-ended, this 'step' is not vacuous.

[29] It might be pointed out that the model here is incomplete, and an intermediate node which registers the combined state of poison and antidote would push *Poison=1* out of the redundancy range. But that's an ineffective response, unless in fact *no* model of the structure offered by Hiddleston is possible. However, we can always construct such a model, as a game, for example.

Step 2 Select and instantiate an actual context O.

Typically, this involves selecting a set O of variables in M and fixing them at their observed values. Often the causal question itself sets the context for us, when selecting the context is trivial. For example, someone might ask, "Given that no poison was added to the coffee, did the antidote cause the king's survival?" Indeed, that appears to be exactly what Hiddleston was asking in the above example. It would be a different problem were we asked, "Given nothing — that we don't know whether or not there was poison in the coffee, did the antidote cause survival?" We shall give our answer to *that* question later.

Instead, consider Glue Girl. Glue Girl is about to step inattentively onto the street, onto the path of an on-coming truck. Her father is next to her, grabs her hand and holds her in place. There is no problem understanding that he saved her life. But what if Glue Girl, true to her name, had her shoes firmly glued to the pavement and never would have succeed in stepping onto the street? Surely, in that case the father's action is redundant and non-causal. And yet in the latter case, on the face of it, the two potential causes are in a symmetric position: each one cancels the dependency of survival upon the other, making this a case of symmetric overdetermination. But, although symmetric overdetermination is an important issue (which we consider briefly below), we would claim this is not such a case. Under the description given, the event of gluing the girl's shoes to the pavement must have occurred first, making it more salient than the father's action and, typically, putting it (if only implicitly) in the context for the problem. With Glue Girl's shoes firmly glued to the pavement, the father's action carries no probabilistic dependency. Context selection matters, and, in ordinary discourse, timing matters for context selection. (Here, there may also be an asymmetry in the effectiveness of the two events in preventing death.)

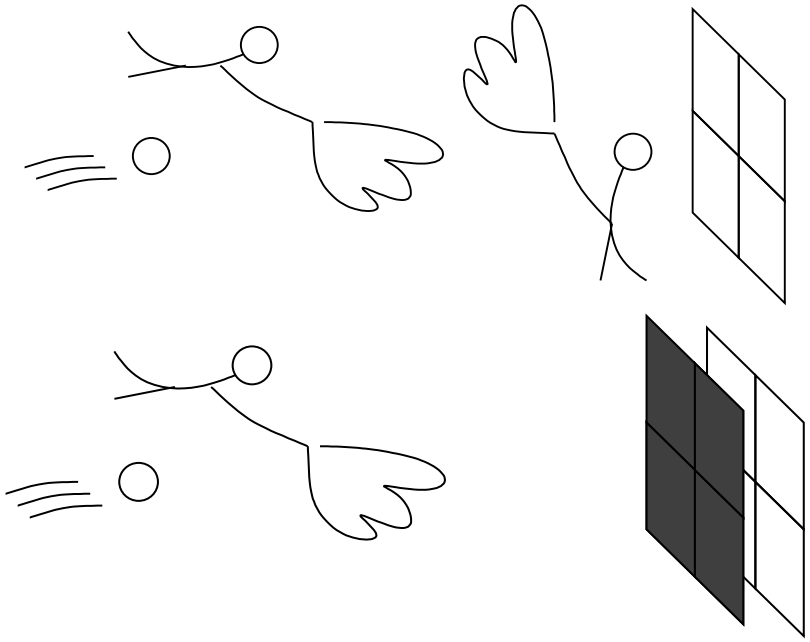

Figure 15: Backup fielders, imperfect and perfect.

An even more striking example of context-setting is this (from Collins, 2000). Suzy and Billy are playing left and center field in a baseball game. The batter hits a long drive between the two, directly at a kitchen window. Suzy races for the ball. Billy races for the ball. Suzy's in front and Billy's behind. Suzy catches the ball. Did Suzy prevent the window from being smashed? Not a hard question. However, now suppose we replace Billy with a movable metal wall, the size of a house. We position the wall in front of the window. The

ball is hit, Suzy races and catches it in front of the window. Did Suzy prevent the window from being smashed? This is no harder than the first question, but produces the opposite answer. The question, of course, is how can our criterion for actual causation reproduce this switch, when the two cases are such close parallels of each other. What changes is the context in which the causal question gets asked. Here the changed context is not reflected in the values of the variables which get observed — neither the wall nor Billy are allowed to stop the ball; rather, the changed context is in the probability structure.[30] Billy is a fallible outfielder; the wall, for practical purposes, is an infallible outfielder. The infallibility of the wall leaves no possible connecting process between Suzy's fielding and the window.[31]

The idea of a connecting process arose already in Principle 3, in that a sequence of connecting processes must be possible between any two variables in the model. Given our discussion of type causal relevance, we can now say we require of a **connecting process** between C and E both that it be a Salmon-Dowe process and that C be type causally relevant to E — i.e., that there be an intervention on C that makes a probabilistic difference to E. The baseball example is a case where an *individual* arc fails to have a corresponding connecting process for any value of its causal variable. Such arcs we call **wounded** and require their removal:[32]

Principle 4 (Connecting Process) *For every arc $C \dashrightarrow E$ in M/O there must be a possible value for C such that there is a connecting process between C and E.*

Step 3 Delete wounded arcs, producing M'.

In the second baseball case the arc *Suzy Catches* $\dashrightarrow$ *Window Shatters* starts out wounded: it's an arc that should never have been added and which no causal discovery algorithm would add. But many cases of wounding arise only when a specific context is instantiated. For example, in bottle smashing (Figure 14), if Suzy doesn't throw, there's nothing wrong with the causal process carrying influence from Billy's throw to the bottle. If we ask about Billy's throw, however, specifically in the context of Suzy having thrown first, then there is no connecting process. The arc *Billy Throws* $\dashrightarrow$ *Bottle Shatters* is **vulnerable** to that context, and, given the context, is wounded.

Until now, in Bayesian network modeling two kinds of relationship between pairs of variables $< C, E >$ have been acknowledged: those for which there is always a probabilistic dependency regardless of context set, when a direct arc must be added between them;[33] and those pairs which are screened off from each other by some context set (possibly empty). But vulnerable arcs are those which are sometimes needed, they connect pairs of the first type above, but also they are sometimes not needed; when they are wounded, the arcs can

[30] The causal question being relative to an observational context presupposes that it is also relative to the causal model, including parameters, in which the context is set.

[31] An alternative analysis of this would be to say that since the wall *is* infallible, it effectively has only one state, and so is not a variable at all. Whether we really should remove the variable depends upon whether or not we wish to take seriously the possibility of it altering state, even if only by an explicit intervention. Regardless of whether we deal with the wall in parameters or structure, there will remain no possible dependency between Suzy's catch and the window.

[32] For a more careful discussion of the metaphysics of process and wounding we refer the reader to our companion paper (Handfield et al., 2005).

[33] This means, for every possible set of context variables O there is *some* instantiation of the context $O = o$ such that $P(E|C, O = o) \neq P(E|O = o)$. This is not to be confused with requiring that for all possible context sets, and *all possible instantiations thereof*, C and E are probabilistically dependent, which is a mistake Cartwright (2001) makes.

mediate no possible probabilistic dependency, since they cannot participate in any active path. We might say, the arcs flicker on and off, depending upon the actual context.

Strictly speaking, deleting wounded arcs is not necessary for the assessment of token causality, since the next step applies a dependency test which is already sensitive to wounding. Removing the wounded arc simply makes the independency graphic, as faithful models do.

Step 4 Apply a restricted causal relevance criterion (Wiggle Logic) to C and E in M'/O.

The causal relevance criterion is restricted specifically in that the move in Wiggle Logic to extend the context O in order to intercept alternative causal paths from C to E should be disabled.[34] In type causation our interest is in identifying whether or not there is a direct or indirect causal path which potentially *can* make a difference given O; to answer that question we must cut through the strands of causal influence which may hide such difference making. In token causation our interest is in identifying whether C actually *does* make a difference within context O; to answer this question we must allow that alternative strands of influence may nullify the affect. Thus, in questions of token causation we should allow for neutralizing alternatives. In neutral Hesslow, for example, we would not attribute token causality to the pill, whether or not thrombosis ensued — unless, of course, the woman's state of pregnancy were fixed as part of the context. Neutrality relative to the type question is not an option, however, since the type question, indicates a desire to know whether there is an available extension to context which yields a difference-making intervention, which there is in neutral Hesslow. So, in this way, our token causality criterion differs from a straightforward particularization of our type criterion, by being bound to the given context M'/O.

Since we are not focused here on causal role, the probabilistic difference identified in Step 4 might well be to *reduce* the probability of E; hence, rather than saying that C actually caused E, it might be better simply to say that C is actually causally relevant to E. As the context in which the token causal question gets raised, O, is enlarged, this criterion becomes more particular to the historical circumstances; as the context O shrinks, this criterion more closely resembles our criterion for type causal relevance, with the exception noted above.

8.2 Remarks

8.2.1 Relevance versus Role

As the last step of the algorithm falls back upon a version of our type causality criterion, which is explicitly aimed at relevance rather than role, our token causality criterion likewise aims at relevance rather than role. On the other hand, many of the cases of which we are asked our intuitive judgments are clearly oriented towards causal roles. So how do these facts fit together?

Well, in part, they just don't. As we have said, we are not here engaged in ordinary language philosophy nor the pragmatics of causal attribution. Our criterion will end up endorsing token causality for an event which is type-wise a preventative, but which fails in some case to prevent. This seems fine in Rosen's golf ball example: we intuitively judge

[34] Note that, as with our type criterion, we are assuming that the context O is already (idealistically) objectively homogeneous. If not, then either criterion may misfire with either a falsely negative or a falsely positive claim of causation.

that hitting the tree in this case is an actual cause of holing the ball. In the case of the king's coffee, if the poison is added and the antidote is added and the king nevertheless dies, then the antidote will also be a token cause of the king's death, since under the context of poison, it will be causally relevant. But it seems unnatural to say that the antidote caused the king's death, since it is relevant because it lowers the king's chances of dying.

The answer to this issue appears to be complex, and we will give only part of it. First, our criterion surely doesn't warrant the ordinary language assertion of the form "A actually caused B" (let alone "A was responsible for B" or its relatives). Such warrant is a difficult matter of pragmatics. A more plausible English rendition of what our criterion might warrant would be the more obscure "A was actually causally relevant to B". Thus, the golf ball should count as less of a success for our account, since we do not explicitly treat role issues, and the king's death as more of a success, for the same reason.

But actual causal role is not only a matter of pragmatics; what we offer here is a framework for building such a pragmatic account. Also needed is an assessment of the impact or power of the various actual causes of an effect. If there are two or more actual causes of an effect, conversational implicature plausibly requires picking out the more powerful of them for the actual effect value, at least in the first instance. In the king's case, no doubt, the antidote has some negative causal power for death and the poison some positive causal power. We expect that a complete account of actual causation covering the warrant for assertions would need to draw upon these disparate causal powers.

8.2.2 Ossification

Some might object to our blending of type relevance into our account of token causality, on the ground that when dealing with *actual* causality, matters are necessarily historical, when all the values of variables have been fixed. If all the variables are fixed, then the context O must be all inclusive and there is no longer any question of establishing causal relevance according to Wiggle Logic, because there is no wiggle left. But if we take this point seriously, then the effects and causes are themselves saturated — a fully ossified (saturated) model is one in which we cannot find causal powers, and so we cannot find causal explanations either. Nor can we escape this by observing that, while clearly the candidate cause and effect must themselves be unfixed in order to find any dependency, we can unfix *only* the effect and cause: in a deterministic world that would have no impact unless the cause and effect are directly related. As we have argued, the analysis of causality should depend upon neither determinism nor indeterminism, so this won't do.

In any case, it is clear that ordinary judgments of actual causality do not presuppose that all variables are fixed. There is clearly a difference in our judgments, for example, of the danger posed by a falling rock when we are, or are not, given that the hiker ducked. Although accounting for all variations in our intuitive judgments is beyond our scope here, an account of actual causation which requires judgments always to depend upon fully saturated contexts could never accommodate them. Our proposal is that contexts come in degrees of saturation and that judgments of token causality vary accordingly.

8.2.3 Process Theory and Absences

We noted in §7.3 that absences, and their causal impacts, are not particularly problematic. This is clear when we are modeling causal structure with Bayesian networks, when we are always supposing the variables take *some* value, even if the value is unknown (the variable unobserved), or again if the value is only known to within an interval, or, by negation, to

within two intervals. In all such cases we can consider "causation by absence" as shorthand for causation by presence, where the exact value present is unknown.

When C and E are type related, the relation (dependency) may be borne by a variety of Salmon-Dowe-like processes, and which process is realized is determined by which value C takes, perhaps in conjunction with some background context. But one variety of such a 'like process' is the non-process! As we mentioned earlier, causal dependence (and probabilistic dependence) is a function of the structure of possible causal processes, realized or not. Given that poison has been poured into the king's coffee, applying an antidote initiates a clear Salmon-Dowe process which neutralizes the poison and saves the king; equally clearly, however, failing to apply the antidote initiates no process at all, *which failure* allows the king to die. The token causal efficacy of a non-process is here plain. Thus, it would be a mistake to identify either variables or arcs with actual causal processes.[35] We claim that in any case where a non-process is causally efficacious it is precisely because there is a physically possible process which could be realized in its stead which would have a different outcome (i.e., induce a distinct probability distribution over the effect).

Lewis averred (Lewis, 2004, p. 282) "the problem of missing relata hits any relational analysis of causation," but he was wrong. This is plainly so for relational analyses of type causation, when relata are always present. But it is equally so for token causation. A causal process may fail to be present, of course. However, the causal relata of interest are not the processes, but the instantiated variables, states of affairs, or events (whichever you prefer) which supervene upon the complex of processes, and missing processes. Processes may collectively absent themselves, but states of affairs may not. And whatever state obtains at a time — whatever joint variable state describes reality — may be put into causal relations with effect variables of a later time, whether this is done by present processes or absent ones.[36]

This is somehow the *opposite* of wounding. When wounded, the (type) causal arc between two variables can no longer mediate a causal dependence, because no difference-making process can connect the two. When Suzy throws, Billy's throw no longer matters, because no connecting process is even possible. For the king, however, difference-making is supported specifically by the possibility or *not* of instantiating a process. In both cases, there may be no process mediating the two variables of interest. The difference is that in one case, that of the king, there is a *possible* connecting process that remains unactualized, whereas in the case of the rock there is no possible connecting process, and hence no probabilistic dependency (in the context of Suzy having thrown).[37]

The temptation to deny the existence of causation by omission seems to arise from the identification of causation with the ineffable biff. For biff is causal glue, and you can't glue something to nothing. The denial of absent causes is taking deep causal metaphysics for

[35]This is a temptation to which Woodward and Hitchcock (2003) succumb in their first (informal) definition of intervention.

[36]Note that this also supplies the answer to cases of double prevention, where some action prevents a preventer from interfering with another process. For example, if the guard kills the assassin before he can tamper with the the king's coffee, the action prevents the poison from interfering with the king's continued survival. Those who worry about the token causality of this kind of action are overly fixated on the fact that there is no causal process from the guard's action to the king's bodily health; but it is the *absence* of a deadly process which is telling here.

[37]We note for the metaphysically inclined that in this paper we have often avoided metaphysical commitments where they might have been made. In particular, we do not here consider precisely what it means for a causal process to connect variables, events, event types, etc.; nor do we consider precisely what it means to *be* a causal process. Indeed, these two questions are interrelated. We simply trust that sensible accounts of these matters can be given.

the whole of the causal story, when it is at best only a part, and at worst a part which will never be told.

In short, absent token processes can be part of the causal story, type or token, and for the same reason that present token processes are: they can make a difference in context. Absent difference-making, absent causality.

8.3 Token Cases

Finally, in light of the above remarks, we apply our algorithmic criterion in reviewing various cases of token causality — those presented above, as well as several additional ones taken from the literature — to see how it handles them, plausibly or not.

To reprise:

Token Causality Algorithm

Step 1 Build the right causal model M.

Step 2 Select and instantiate an actual context O.

Step 3 Delete wounded arcs, producing M'.

Step 4 Apply the restricted causal relevance criterion (Wiggle Logic) to C and E in M'/O — i.e., determine whether the actual context allows for C to make a difference to E.

In applying the algorithm it needs to be borne in mind that our criterion is *model relative*; this is the intention behind Step 1. It often makes a difference to the verdict whether the models are stochastic or deterministic. It might be thought that under an indeterministic interpretation just about *everything* is causally relevant to everything else! Given contexts that screen things off, that is not so, as we see in §8.3.8. With the null context the complaint is closer to the truth, if sometimes an exaggeration. This suggests that divergence between our criterion of token causality and intuition will often be based upon the difference between a *low degree* of relevance and *no degree* of relevance. But in that case we must beg the reader's indulgence: our account is only a beginning, and specifically does not include any benefit an analysis of causal power might confer.

8.3.1 Hiking

The first step of the algorithm is always to find the right causal model. Therefore, we assume we have Figure 5 to work with. The algorithm will now give different answers, depending upon whether the model is understood deterministically or indeterministically.

- If the model is deterministic, then *Boulder* does not cause survival given an empty context: whatever value it takes, survival ensues.
- If the model is indeterministic, then in the null context *Boulder* will be token causally relevant to *Survival*. We should expect this also in the context *Duck* = *true*, on the grounds that the trajectory of the boulder is uncertain.

8.3.2 Trainee Assassin

The model is given in Figure 13. Assuming idealistic determinism, if we are given that the supervisor did not fire, we get the same answer as Hitchcock's H1, that the trainee caused the victim's death. In a null context, the answer is that there is no connecting process from trainee to victim; there is a process, of course, but it fails to make a difference to the outcome. This answer relies upon our account of the token question being more firmly tied to context than the type question, disallowing the consideration of any elaborations of the (null) context to fix the supervisor's state. Given indeterminism, on the other hand, plausibly the supervisor's abilities are imperfect, when the trainee's firing will always be token causally relevant, as will be the supervisor's action.

8.3.3 Hiddleston's Antidote

Hiddleston begins his story by affirming that no poison is added to the king's coffee; this can only mean that that fact is part of the given context. Hence, on our algorithm the antidote makes no difference and is not a token cause of the king's survival.

8.3.4 Fielder

We already analysed the case of Figure 15 in the text, with the answer being that discrepancies in the intuitive judgment between the two cases can be understood in terms of discrepancies between two contexts, in a sense of context that extends beyond a set of observations to include also conditional probability relations between variables.

8.3.5 Bottle Smashing

An overly simple model of the story is that of Figure 14. Given that Suzy's rock was always going to arrive at the bottle first, there is no connecting process between *Billy Throws* and *Bottle Shatters* — the arc between them is wounded; hence, Suzy's action is causal and Billy's is not. The same answer holds for more complex models which reflect the temporal asymmetry between Billy's and Suzy's actions explicitly.

8.3.6 Matches

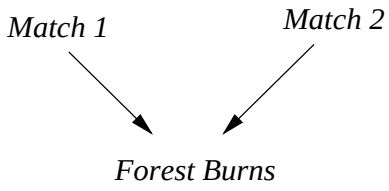

Figure 16: Symmetric matches.

Matters are more difficult for the truly symmetric overdetermination case of Matches, in Figure 16. Here, we are to suppose that both matches are lit, either is alone sufficient to burn the forest, and the forest has burned down. Since it is symmetric, answering for one match answers for both. The general intuition is that both matches caused the forest to burn.

Our algorithm gives the same answer, assuming a null context. On the other hand, if you force the other match's state into the context, then the arc from the candidate match to the forest burning is wounded, and so not causal. You may be inclined to say "that

match was causal *even if* the other was lit." We believe such an intuition is built upon a more complex, and more realistic, model of forest fires — one where each match is going to burn *some* of the forest regardless of what happens with the other match. In such a model, token causation of *some* of the forest's burning will always be attributable to a lit match. When we simplify matters by moving to a three-variable model our intuitions founded upon a more complex understanding of the system may well carry over; but if we really want an answer for the more complex reality we are then simply running the algorithm on the wrong model.[38]

8.3.7 Trumping Prevention

Schaffer (2000) introduced the idea of trumping prevention as a problem for dependency accounts of causation. A sergeant and a major simultaneously shout orders to the troops. Were the major not to make an order, the troops would follow the sergeant's; however, the major's orders trump those of the sergeant's when they are in conflict. In this case, they *don't* conflict. So, which order was causally efficacious? The intuitive answer is, of course, that the major's order took precedence and was efficacious, while the sergeant's was not. The curious thing about this example is that both processes of issuing-order-to-soldier-decision-making can be traced from start to finish and, unlike Billy's rock-to-bottle process, they both somehow complete. Any simple dependency account will have trouble because neither order is necessary for the effect. Our account notes that the story of the sergeant's order is intended to be understood within the context of the major having issued an order, when the lack of a connecting process is plain, since there is no probabilistic dependency between sergeant and soldier in the given context.

8.3.8 The Stochastic Assassin

It may seem that under indeterministism everything is causally relevant to everything else; however, that impression is mistaken.

Consider, for example, our trainee assassin, but in a situation where the supervisor may, or may not, take *Aim* in preparation for providing the backup, as in Figure 17 (as suggested by Hiddleston, 2004). When the supervisor does aim, it is somewhat more likely that her weapon will fire independently of her trainee, and the victim die. In actual fact

[38]The classic symmetric overdetermination case of the firing squad might be held to cause us greater problems. It makes sense to talk of burning down part of a forest, but obviously not to talk of killing part of a prisoner. While our intuitions of responsibility may continue when expanding a context to include the actions of others, we seem to be unable to expand the causal model to incorporate partial death. But here we simply hold firm: either the simple three-variable model (or its analog) is the correct model of what we know of the firing squad or it is not. If it is correct, then given the deadly context of some soldier's firing, another's action can make no difference, and so cannot be token causal. If we must make further moves to accommodate intuition (which is not clear), then there are two routes open to us. First, intuitions might well be based upon opinions about moral or legal responsibility; but such opinions are based upon much more than the basic causal story, and we make no pretence to be offering such elaborations. Second, reference to more detailed causal models may well be appropriate, after all. Although no soldier will be killing a part of a prisoner, different soldiers may well be causing different varieties of damage to the prisoner. For example, an autopsy might reveal that the actions of some soldier pre-empted what anyone else did, or were themselves pre-empted. The possibility of an autopsy reveals that any real case will be much richer than any three-variable model. Both our intuitive judgments and the judgments of our analysis would differ from those produced by analysing the original model. The fact that our analysis is actually sensitive to such considerations is a considerable advantage over any which is invariant to them.

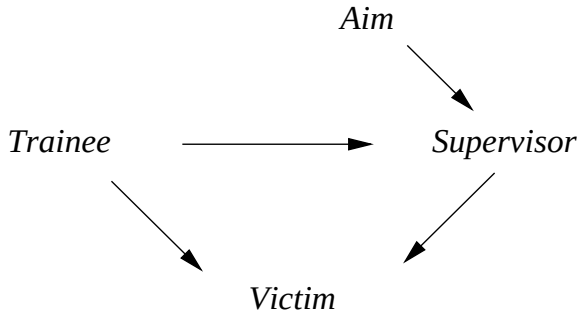

Figure 17: The stochastic assassin.

(i.e., in the actual case of causation we are dealing with), however, the supervisor did not fire, the trainee fired and the victim died.

But on our algorithm it seems that *Aim* remains an actual cause of the victim's death, since the trainee's firing doesn't screen off *Aim* from *Victim*. Our response is that this is an illusion cast by the neglect of context. Since it is an explicit part of the story that the supervisor didn't fire, we put that in the context, at which point *Aim* is indeed screened off from *Victim* and the only cause remaining is the trainee's firing. If we are to consider causality when the supervisor's actions are unknown, then the stochastic causal model tells us that they were relevant to the victim's health whatever the trainee did — precisely as in the simple case of § 8.3.2.

8.3.9 Menzies' Alarm

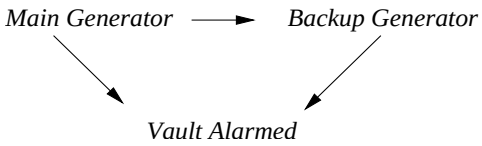

Figure 18: Fail-safe alarm.

Menzies (2002) asserts that in a default context, holding fixed nothing special, the alarm model of Figure 18 should lead to the conclusion that *Main* does not cause the alarm to be powered, since under ordinary circumstances *Backup* is powered, so wiggling *Main* can make no difference. But this is an unhappy interpretation of the case. If this reasoning goes through it seems it should also go through in the case of the trainee assassin, leaving the trainee off the hook!

A plausible argument by cases is: either *Backup* is a separately intervenable variable or it is not. If not, if we are to think of the interrelation between the two generators as unbreakable, then Menzies has given us the wrong model. If they are separable, then in the null context we need to take seriously the *possibility* that *Backup* will be intervened upon. For example, we can consider the unlikely, but possible, state where thieves have sabotaged the backup generator. Having a null context means that we do *not* fix the state of *Backup*, after all, either to active or inactive, which means taking the full range of possibilities seriously. But in that case, it's clear that *Main* does make a difference to whether the alarm is active. Another way of putting this is: staying with the implausibly idealized, deterministic model in this case misleads intuition; viewing the relations as stochastic and the context as empty, we readily see the probabilistic difference wiggling *Main* will make.

Of course, if we were to interpret the alarm system as deterministic and ideal, so that *whenever Main* failed *Backup* would certainly take over, then the above argument would not

work, since wiggling *Main* would make no difference. So our interpretation relies upon an indeterministic understanding of the model. But the deterministic interpretation is inferior in this case. Presumably the electrical engineers must have sometime taken the possibility of the main generator failing seriously, since they went to the trouble to design and build a backup generator. But, under determinism, we are now being asked to not take seriously the possibility of the backup generator failing! That is certainly no good way to do risk analysis! The asymmetry in treating the two generators, the fallible main and the infallible backup, is driven by nothing beyond a deterministic bias and highlights how unreasonable it is to rely upon determinism, and intuitions built upon determinism, in judging causal relationships.

9 Conclusion

We have sought to provide an initial analysis of type and token causation relative to causal models. We suppose that a large number of different causal questions may raised, and that they often deserve different answers. Some part of that variation is captured in considering distinct models, perhaps with some models being supermodels of others, showing a finer degree of resolution. Another part of that variation in question-raising is handled by the various contexts that causal questions assume. Contexts can be explicitly set by the question, or implicitly set by the temporal precedence of some events or conversational implicature. What answer you get to what causes or caused E ought to depend upon what causal model E is operating in; and it ought also to depend upon one's assumptions about what should be held fixed. Our proposal makes these dependencies explicit.

Ours is a level of analysis which can make sense of inductive processes learning about the world, whether realized in humans or machines. Such causal relations supervene upon, but not only upon, causal processes active in the world. A common intuition is that these deeper processes are intrinsically causal. But, it seems to us that little progress has been made in accounting for these deep causal processes — and dubbing them 'biff' is a kind of confession of this. Our analysis does not depend on mining such deep metaphysics. Nor does it depend upon any specific way of soaring into the pragmatics of causal talk: just how the plethora of causal intuitions and locutions are to be accommodated remains an open box of troubles. We are happy if we have found a Daedalian solution to accounting for causation as we actually model it.

Appendix A: Model Building Rules

Our proposed model building rules are:

Principle 1 (Intervention) Variables in a causal model must be intervenable.

Principle 2 (Variable Distinction) Every pair of variables in a causal model must have a physically possible intervention which, in some physically possible context, affects the distribution of one variable without affecting that of the other.

Principle 3 (Causal Processes) If two variables appear in a causal model, there must be a sequence of possible or actual causal processes connecting them.

Principle 4 (Connecting Process) For every arc $C \longrightarrow E$ in M/O there must be a possible value for C such that there is a connecting process between C and E.

We have presented various reasons for adopting these principles, although we offer them only tentatively and as needing elaboration and refinement. We do think they are an improvement on Hitchcock's alternative rules (Hitchcock, 2001a, notation omitted):

> What makes a causal model appropriate? There are at least three requirements. The first two are objective: the equations must entail no false counterfactuals, and they must not represent counterfactual dependence relations between events that are not distinct. The third component is pragmatic: [the model] should not contain variables whose values correspond to possibilities that we consider too remote.

Hitchcock's first requirement is tantamount to the requirement that the stochastic laws encoded in a causal model should be true.[39] We make the same assumption, but have not made it explicit in a principle. Treating it in any detail would take us into a study of causal discovery algorithms and outside this paper. The second requirement we adopt, and we believe that the Hausman and Woodward (2004) account of variable distinction in terms of separate intervenability provides its best interpretation. Our principle 3 provides a more exact requirement of locality than Hitchcock's third rule. Of course, remoteness may be less of a matter of space-time and more a matter of causal, or decision-theoretic, relevance. Treating such issues must wait upon a proper theory of causal power, however. Our fourth principle goes beyond Hitchcock's recommendations in pointing out that it is not just variables which require interpretation, but arcs as well. Salmon-Dowe process theory makes its contribution here, in providing us with a metaphysical basis for connecting processes. Since such processes are also required to make sense of interventions, process theory actually lies behind each of our principles.

Appendix B: Empirical Distinguishability via Interventions

Here we prove that any faithful and faithless models of a single probability distribution — and so sharing a common conditional dependency structure — will have distinct conditional dependency structures when fully augmented. Our proof is for linear path models, however it should generalize since linear models induce the same kinds of dependency structures as Bayesian networks generally. There is one respect in which our second proof does not so generalize, which we note below.[40]

Since we are only interested in the faithful-faithless distinction arising from multiple paths from cause to effect, we can restrict our proof to models structurally isomorphic to the neutral Hesslow case. In other words, the Hesslow case models all varieties of faithlessness which might cause difficulties for causal discovery algorithms. In particular, we consider models M_1 and M_2 of Figure 19. The p_i report the path coefficients (direct linear causal powers, if you like), indexed according to Wright's convention of naming effects first. The relevant fact about path models needed for our proofs is that the correlation between any two variables is equal to the sum of the products of the linear path coefficients across all

[39] Curiously, Woodward and Hitchcock (2003) take the view that the relations encoded in causal models aren't lawlike. But their view is motivated by models that only incompletely model the causal relations between their variables. When treating metaphysics we prefer to idealize in considering only complete models, or at least complete submodels.

[40] Also, in the special case of interventions I_X which directly impact on more than X in the model, either proof may fail.

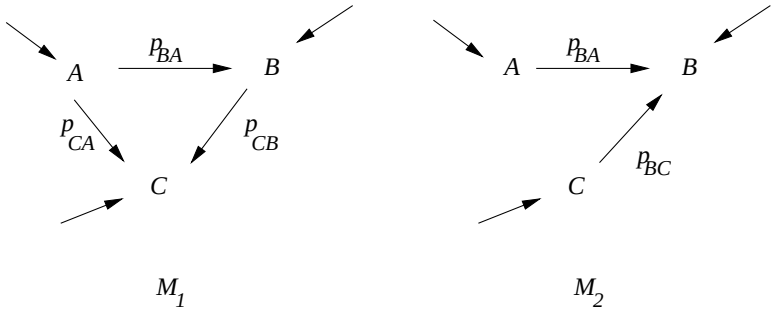

Figure 19: M_1 and M_2.

d-connecting paths between them. We call this Wright's rule (see Wright, 1934, for details). In a formula:

$$r_{XY} = \sum_i \prod_{WZ} p_{ZW}$$

where i ranges over the paths $\Phi_i(X, Y)$ d-connecting X and Y and WZ ranges over pairs of variables directly connected within each such path.

Theorem 1 (Distinguishability under Perfect Intervention) *If the unaugmented path models M_1 and M_2 represent the same probability distribution, then when fully augmented with perfect interventions, they represent distinct distributions.*

Proof. By assumption

(1) $$r_{AC} = 0$$

in both models. Indeed, all correlations, conditional and marginal, are shared between the unaugmented models. For M_1, using Wright's rule, this implies:

(2) $$p_{CA} = -p_{BA}p_{CB}$$

whereas for M_2 we have:

(3) $$p_{CA} = 0 \neq -p_{BA}p_{CB}$$

By Wright's rule, the following are true of both models:

(4) $$r_{AB} = p_{BA}$$

(5) $$r_{BC} = p_{BA}p_{CA} + p_{CB}$$

From (2), (4) and (5) it follows algebraically for M_1 that

(6) $$p_{CA} = \frac{-r_{AB}r_{BC}}{1 - r_{AB}^2}$$

Under a perfect intervention upon B, for M_1 we derive a new correlation between A and C, namely:

(7) $$r'_{AC} = p_{CA} = \frac{-r_{AB}r_{BC}}{1 - r_{AB}^2}$$

where primed correlations represent those under the intervention and plain correlations refer to those determined by the original model. For M_2 we get:

(8) $$r'_{AC} = 0$$

QED.

Theorem 2 (Distinguishability under Imperfect Intervention) *If the unaugmented path models M_1 and M_2 represent the same probability distribution, then when fully augmented with imperfect interventions, they represent distinct distributions.*

Proof. First, note that all of the formulas above for the models prior to intervention (1-6) hold. It suffices to find one conditional dependency under augmentation in M_1 which fails to hold under augmentation in M_2. The following conditional independency ('vanishing partial correlation') holds in M_2 by d-separation (the Markov property):

$$r_{AI_C.CB} = 0 \tag{9}$$

In M_1 we can see that A and I_C are d-connected given $\{C, B\}$. It does not immediately follow that there is a conditional dependency, since faithfulness is at issue. However, by Wright's rule we have

$$r_{AI_C.CB} = -p_{CA}p_{CI_C} \tag{10}$$

But this cannot also be zero, since neither term on the right-hand side can be zero. In particular, $p_{CA} \neq 0$ by construction of the problem, while I_C, being an intervention variable on C, must somehow affect C. QED.

We note that the very last step of this proof implicitly takes advantage of the linearity of the models: I_C is required to affect C regardless of the observed state of B. This will be true of any linear model, however some non-linear models may have interacting causes where states of B nullify attempted interventions on C. In that case, the proof does not hold, although we should expect other interventions on the model nevertheless to lead to experimentally distinguishable outcomes between M_1 and M_2.

Acknowledgements

We thank Christoper Hitchcock, Dan Hausman, Erik Nyberg, Philip Dawid, Jon Williamson, Colin Howson, John Worrall, Luc Bovens, Stephan Hartmann and various participants in colloquia at Konstanz, UCL and LSE during 2004. A Monash Research Fund grant and a Monash Arts/IT grant supported this work.

References

Arntzenius, F.: 1999, 'Reichenbach's Common Cause Principle'. In: *Stanford Encyclopedia of Philosophy*. Stanford. http://plato.stanford.edu.

Bovens, L. and S. Hartmann: 2002, 'Bayesian Networks and the Problem of Unreliable Instruments'. *Philosophy of Science* **69**, 29–72.

Cartwright, N.: 1983, *How the Laws of Physics Lie*. Oxford University.

Cartwright, N.: 1989, *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Cartwright, N.: 2001, 'What Is Wrong with Bayes Nets?'. *The Monist* **84**, 242–64.

Cheng, P. W.: 1997, 'From covariation to causation: A causal power theory'. *Psychological Review* **104**, 367–405.

Chickering, D. M.: 1995, 'A Tranformational Characterization of Equivalent Bayesian Network Structures'. In: P. Besnard and S. Hanks (eds.): *11th Conference on Uncertainty in AI*. San Francisco, pp. 87–98.

Chickering, D. M., D. Geiger, and D. Heckerman: 1995, 'Learning Bayesian Networks Is NP-Hard'. In: *5th Conference on AI and Statistics*. pp. 112–28.

Collins, J.: 2000, 'Preemptive Prevention'. *Journal of Philosophy* **97**, 223–234.

Cooper, G. F.: 1990, 'The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks'. *Artificial Intelligence* **42**, 393–405.

Cooper, G. F. and C. Glymour (eds.): 1999, *Computation, Causation and Discovery*. MIT Press.

Dowe, P.: 2000, *Physical Causation*. New York: Cambridge University.

Eells, E. and E. Sober: 1983, 'Probabilistic Causality and the Question of Transitivity'. *Philosophy of Science* **50**, 35–57.

Einstein, A., B. Podolsky, and N. Rosen: 1935, 'Can Quantum-Mechanical Descriptions of Physical Reality Be Considered Complete?'. *Physical Review* **47**, 777–80.

Fisher, R.: 1957, 'Letter'. *British Medical Journal* pp. 297–8.

Gillies, D.: 2000, 'Varieties of Propensity'. *British Journal for the Philosophy of Science* **51**, 807–35.

Glymour, C.: 1998, 'Psychological and Normative Theories of Causal Power and the Probabilities of Causes'. In: G. F. Cooper and S. Moral (eds.): *14th Conference on Uncertainty in Artificial Intelligence*. pp. 166–72.

Glymour, C.: 2002, *The Mind's Arrows*. MIT Press.

Gyenis, B. and M. Redei: 2004, 'When Can Statistical Theories Be Causally Closed?'. *Foundations of Physics*. forthcoming.

Hall, N.: 2004, 'Two Concepts of Causation'. In: J. Collins, N. Hall, and L. A. Paul (eds.): *Causation and Counterfactuals*. MIT Press, pp. 225–76.

Halpern, J. Y. and J. Pearl: 2005, 'Causes and Explanations, Part I'. *British Journal for the Philosophy of Science*. Forthcoming.

Handfield, T., G. Oppy, C. Twardy, and K. B. Korb: 2005, 'Probabilistic Process Causality'. in preparation.

Hausman, D. M.: 2005, 'Probabilistic Causality and Causal Generalizations'. In: E. Eells and J. H. Fetzer (eds.): *The Place of Probability in Science*. Open Court.

Hausman, D. M. and J. Woodward: 2004, 'Modularity and the Causal Markov Condition: A Restatement'. *British Journal for the Philosophy of Science* **55**, 147–61.

Heckerman, D.: 1998, 'A Tutorial on Learning with Bayesian Networks'. In: M. Jordan (ed.): *Learning in Graphical Models*. Cambridge, MA: MIT Press, pp. 301–54.

Henrion, M.: 1988, 'Propagating Uncertainty in Bayesian Networks by Logic Sampling'. In: J. Lemmer and L. Kanal (eds.): *Uncertainty in Artificial Intelligence, 2.* Amsterdam: North-Holland, pp. 149–63.

Hiddleston, E.: 2004, 'Causal Powers'. *British Journal for the Philosophy of Science.* forthcoming.

Hitchcock, C. R.: 2001a, 'The Intransitivity of Causation Revealed in Equations and Graphs'. *Journal of Philosophy* **98**, 273–99.

Hitchcock, C. R.: 2001b, 'A Tale of Two Effects'. *The Philosophical Review* **110**, 361–96.

Hitchcock, C. R.: 2004a, 'Probabilistic Causation'. In: R. Haenni and S. Hartmann (eds.): *Causality, Uncertainty and Ignorance: Lecture Notes.* University of Konstanz.

Hitchcock, C. R.: 2004b, 'Routes, Processes and Chance-lowering Causes'. In: P. Dowe and Noordhof (eds.): *Cause and Chance.* Routledge, pp. 138–51.

Hofer-Szabo, G. and M. Redei: 2004, 'Reichenbachian Common Cause Systems'. *International Journal of Theoretical Physics.* forthcoming.

Holland, J., K. Holyoak, R. Nisbett, and P. Thagard: 1986, *Induction.* MIT Press.

Humphreys, P.: 1985, 'Why Propensities Cannot Be Probabilities'. *Philosophical Review* **94**, 557–70.

Irzik, G.: 1996, 'Can Causes Be Reduced to Correlations?'. *British Journal for the Philosophy of Science* **47**, 249–70.

Kitcher, P.: 1989, 'Explanatory Unification and the Causal Structure of the World'. In: P. Kitcher and W. C. Salmon (eds.): *Minnesota Studies in the Philosophy of Science*, Vol. XIII. Univ of Minnesota, pp. 410–505.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick: 2004, 'Varieties of Causal Intervention'. In: *Pacific Rim International Conference on AI'04.* pp. 322–31.

Korb, K. B. and A. E. Nicholson: 2004, *Bayesian Artificial Intelligence.* Boca Raton, FL: CRC/Chapman and Hall.

Lewis, D.: 1973, 'Causation'. *Journal of Philosophy* **70**, 556–67.

Lewis, D.: 1986, *Philosophical Papers, Volume II.* Oxford Univ.

Lewis, D.: 2000, 'Causation as Influence'. *Journal of Philosophy* **97**, 182–97.

Lewis, D.: 2004, 'Void and Object'. In: J. Collins, N. Hall, and L. A. Paul (eds.): *Causation and Counterfactuals.* MIT Press, pp. 277–90.

Menzies, P.: 2002, 'Causal models, token causation and processes'. *Philosophy of Science.* Forthcoming supplementary volume for PSA 2002.

Menzies, P.: 2004, 'Difference Making in Context'. In: J. Collins, N. Hall, and L. Paul (eds.): *Counterfactuals and Causation.* MIT Press, pp. 139–80.

Murphy, K. P.: 2001, 'Active Learning of Causal Bayes Net Structure'. Technical report, University of California at Berkeley. http://www.ai.mit.edu/ murphyk/papers.html.

Neapolitan, R. E.: 2003, 'Stochastic Causality'. In: *International Conference on Cognitive Science, Sydney, Australia.*

Neapolitan, R. E.: 2004, *Learning Bayesian Networks.* Prentice Hall.

Pearl, J.: 2000, *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press.

Popper, K. R.: 1959, 'The Propensity Interpretation of Probability'. *The British Journal for the Philosophy of Science* **10**, 25–42.

Reichenbach, H.: 1956, *The Direction of Time.* University of California Press.

Reynolds, C. W.: 1987, 'Flocks, Herds and Schools: A Distributed Behavioral Model'. *Computer Graphics* pp. 25–34.

Rosen, D.: 1978, 'In Defense of a Probabilistic Theory of Causality'. *Philosophy of Science* **45**, 604–13.

Salmon, W. C.: 1980, 'Probabilistic Causality'. *Pacific Philosophical Quarterly* **61**, 50–74.

Salmon, W. C.: 1984, *Scientific Explanation and the Causal Structure of the World.* Princeton Univ.

Schaffer, J.: 2000, 'Trumping Prevention'. *Journal of Philosophy* **97**, 165–81.

Sober, E.: 1988, 'The Principle of the Common Cause'. In: J. Fetzer (ed.): *Probability and Causality.* Kluwer, pp. 211–28.

Spirtes, P., C. Glymour, and R. Scheines: 2000, *Causation, Prediction and Search.* MIT Press, second edition.

Steel, D.: 2004, 'Biological Redundancy and the Faithfulness Condition'. In: *Causality, Uncertainty and Ignorance: Third International Summer School.* Univ of Konstanz, Germany.

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

Tong, S. and D. Koller: 2001, 'Active Learning for Structure in Bayesian Networks'. In: B. Nebel (ed.): *17th International Conference on Artificial Intelligence.* pp. 863–9.

Twardy, C. R. and K. B. Korb: 2004, 'A Criterion of Probabilistic Causality'. *Philosophy of Science* **71**, 241–62.

van Fraassen, B.: 1982, 'The Charybdis of Realism: Epistemological Implications of Bell's Inequality'. *Synthese* **52**, 25–38.

Verma, T. S. and J. Pearl: 1991, 'Equivalence and Synthesis of Causal Models'. In: S. D'Ambrosio and Bonissone (eds.): *6th Conference on Uncertainty in AI.* pp. 255–68.

von Mises, R.: 1957, *Probability, Statistics, and Truth.* London: George Allen & Unwin, second English edition.

Williamson, J.: 2001, 'Foundations for Bayesian Networks'. In: J. Corfield and J. Williamson (eds.): *Foundations of Bayesianism*. Kluwer, pp. 75–115.

Woodward, J. and C. R. Hitchcock: 2003, 'Explanatory Generalizations, Part I'. *Nous* **37**, 1–24.

Wright, S.: 1934, 'The Method of Path Coefficients'. *Annals of Mathematical Statistics* **5**(3), 161–215.