# Varieties of Causal Intervention

Kevin B. Korb, Lucas R. Hope, Ann E. Nicholson, and
Karl Axnick

School of Computer Science
and Software Engineering
Monash University
Clayton, Victoria 3800
Australia
{korb,lhope,annn}@csse.monash.edu.au

**Abstract.** The use of Bayesian networks for modeling causal systems has achieved widespread recognition with Judea Pearl's *Causality* (2000). There, Pearl developed a "do-calculus" for reasoning about the effects of deterministic causal interventions on a system. Here we discuss some of the different kinds of intervention that arise when indeterminstic interventions are allowed, generalizing Pearl's account. We also point out the danger of the naive use of Bayesian networks for causal reasoning, which can lead to the mis-estimation of causal effects. We illustrate these ideas with a graphical user interface we have developed for causal modeling.

## 1 INTRODUCTION

Little progress has been made in understanding the nature of causality in the last 2500 years, after Aristotle made the first serious foray. David Hume made some negative observations about what causality is not — pointing out, for example, that causal relations are not directly observable. What causality may actually be remains a perplexing problem, but progress has been made in relating it to other concepts whose understanding appears to be more accessible. The influential, and first, text on learning Bayesian networks from data, *Causation, Prediction and Search* (1993) [12], notably eschewed any attempt to define the central concept, focusing instead on the relation between (undefined) causal structure and probabilistic structure. More recently, Judea Pearl has used Bayesian networks to make progress in understanding philosophical problems about causal concepts, giving accounts of counterfactual reasoning [10], experimental methods [9] and token causality [6, 7]. James Woodward, among others, is using Pearl's account of causal intervention to improve upon an old philosophical tradition, attempting to make sense of causality in terms of manipulation [14]. This convergence of artificial intelligence (automated Bayesian networks) and philosophy is, we think, wholly to the good and promises to be fruitful for both sides of the collaboration.

The philosophical use of Bayesian networks largely depends upon a causal interpretation of the arc structure and the probabilistic interpretation of causality, stemming from the work of Patrick Suppes [13] and Wesley Salmon [11]. Although there is some dispute within the AI community about the merits of the causal interpretation, most of this seems to be fueled by the observation

that any Bayesian network can be reordered back-to-front and still represent the very same probability distribution, using Chickering's arc reversal rule [3]. What that observation ignores is that any such reordering can only lead from simple to complex networks when they begin from a perfect map — that is, one whose arcs are both necessary and sufficient for identifying a probabilistic dependency in the system being modeled.[1] A causal interpretation of the Bayesian network implies that we can use the network for causal reasoning and not just probabilistic reasoning. And this further implies the ability to use such models to reason hypothetically about the consequences of interventions. Thus, the difference between the "statistically equivalent" models $Cancer \leftarrow Gene \rightarrow Smoking$ and $Smoking \rightarrow Gene \rightarrow Cancer$ may be determininable by experimentally setting the value of *Smoking*, though not by any observation of the three variables.

Judea Pearl has notably discussed causal interventions and their modeling with Bayesian networks in *Causality* [10]. There, he favors representing an intervention on a variable $C$ by arc-cutting: by setting $C$ to a desired value and cutting all arcs from its parents. He formalizes this approach in his "do-calculus." An alternative method is to introduce a new node $I_c$ as an additional parent of $C$, where setting (or observing) $I_c$ to be TRUE models an intervention. Once the alterations to the network are applied for either method of modeling intervention, ordinary Bayesian network propagation rules can be used. The arc-cutting method is in many ways simpler, but we suggest the simplicity comes at a price: foregoing the possibility of modeling many situations realistically.

In this paper we describe extensions to these techniques for modeling interventions with Bayesian networks and especially (re)introducing indeterminism into that modeling. We hope this will contribute to the collaboration of AI and philosophy of science, as well as open up the wider practical application of Bayesian networks. After a brief defence of indeterminism, we proceed by defining the concept of intervention and presenting a classification system for different kinds of intervention. We then discuss the concept of the effectiveness of an intervention, and finally we describe our programmatic representation of interventions and a GUI for managing them in a Bayesian network tool.

## 2 INDETERMINISTIC CAUSAL MODELS

One curiosity of the collaboration between AI and philosophy of science thus far is a widespread agreement that, at bottom, these networks are deterministic, despite that fact that they are explicitly probabilistic models. Pearl, for one, is adamant that a deterministic conception of causality is required and for three reasons [10, pp. 26-7]:[2]

1. Determinism is intuitive.

[1] Granted, this claim has not yet exactly been proved in the literature. Indeed, it is demonstrably false in cases of measure zero, that is, cases where parameters in the network must be given an *exact* value for the network to be a perfect map. However, outside of measure zero cases, the relation between network minimality and causality is very clear empirically and, we believe, susceptible to compelling arguments. Presenting these, however, would take us beyond the scope of this paper.

[2] For a philosopher voicing the same opinion see, for example, [2].

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. However, we note in passing that there is a growing consensus amongst philosophers of science that such intuitions are insufficient reason for dismissing the probabilistic analysis of causality, which is explicitly indeterministic. Pearl has been collaborating with Joseph Halpern in developing an important account of causal explanation [6, 7]; we hope, however, that an indeterministic account of causal explanation is not actually impossible, since we are developing one.

As for Pearl's last point, 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: $X \rightarrow Z \leftarrow Y$. 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 prediction 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 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.[3]

[3] To convert any deterministic system into an indeterministic system, simply remove the error terms. If there are none, the system is surely correctly described as deterministic, but that is no bar to *representing* it with an indeterministic system having only extreme probability parameters.

## 3 OBSERVATION VERSUS INTERVENTION

Unfortunately, while the causal interpretation of Bayesian networks is becoming more widely accepted, the distinction between causal reasoning and observational reasoning remains for many obscure. This is particularly true in application areas where the use of regression models, rather than Bayesian networks, is the norm, since regression models (in ordinary usage) simply lack the capability of modeling interventions.

We illustrate the difference between intervention and observation with a simple example. Figure 1 presents a three-variable causal model of coronary heart disease (CHD) risk, which is loosely based upon models of the Framingham heart disease data (e.g., [1, 4]). As is normal, each arrow represents a direct and unerasable causal connection between variables.[4] Two contributing factors for CHD are shown: hypertension (HT; elevated blood pressure) at age 40 and HT at age 50. The higher the blood pressure, the greater the chance of CHD, both directly and indirectly. That is, hypertension at 40 directly causes heart disease (in the terms available in this simplified network of three variables!), but also indirectly through HT at 50. In this simplified model, the direct connection between HT at 40 and CHD between 50 and 60 represents all those implicit causal processes leading to heart disease which are *not* reflected in the later HT.

Figure 2(a) shows the results of observing no HT at age 50. The probability of CHD has decreased from a baseline of 0.052 to 0.026, as expected. But what if we intervene (say, with a medication) to lower blood pressure as in Figure 2(b)? The probability is reduced by a lesser amount to 0.033. By intervening on HT at 50 we have cut the indirect causal path between HT at 40 and CHD, but we have not cut the direct causal path. That is, there are still implicit causal processes leading from HT at 40 to CHD which the proposed intervention leaves intact. Observations of low HT at 50 will in general reflect a lower activation of those implicit processes, whereas an intervention will not. In short, it is better to have low blood pressure at 50 *naturally* than to achieve that by artificial means—and this causal model reflects these facts.

A real-world example of people getting this wrong is in the widespread use of regression models in public health. To assess the expected value of intervention on blood pressure at age 40, for example, regression models of the Framingham data have been used [1, 4]. If those models had exactly the same structure as ours, then (aside from being overly simplistic) there would be no actual problem, since HT at 40 being a root node there is no arc-cutting needed. However, the models actually used incorporate a reasonable number of additional variables, including parents of HT at 40, such as history of smoking, cholesterol levels, etc. By simply *observing* a hypothetical low blood pressure level and computing expected values, these models are being used for something they are incapable of representing.[5] The mis-estimation of effects may well be causing bad public policy decisions.

[4] Unerasable means that, no matter what other variables within the network may be observed, there is some joint observational state in which the parent variable can alter the conditional probability of the child variable. In case this condition does

**Fig. 2.** The hypertension causal model where HT at age 50 is (a) observed as low (b) set to low.

## 4 DEFINING AN INTERVENTION

In ordinary usage, an intervention represents an influence on some causal system which is extraneous to that system. What kind of influence we consider is not constrained. It may interact with the existing complex of causal processes in the system in arbitrary ways. For example, a poison may induce death in some animal, but it may also interact with an anti-toxin so that it does not. Or again, the action of the poison may be probabilistic, either depending on unknown factors or by being genuinely indeterministic. Also, an intervention may impact on multiple factors (variables) in the system simultaneously or be targeted to exactly one such variable. In the extant literature of both philosophy and computer science there seems to have been an implicit agreement only to consider the very simplest of cases. In that literature, interventions are deterministic, always achieving their intended effect; and their intended effect is always to put exactly one variable into exactly one state. As a consequence, interventions never interact with any other causes of the targeted variable, rather their operation renders the effect of those other parents null. While such a simple model of interaction may be useful in untangling some of the mysteries of causation (e.g., it may have been useful in guiding intuitions in Halpern and Pearl's study of token causation, [6, 7]), it clearly will not do for a general analysis. Nor will it

---

not hold we have an unfaithful model, in the terminology of [12]. We will not be considering such models here.

[5] In order to be *capable* of representing interventions we require a graphical representation in which the parental effects upon an intervened-upon variable can be cut (or altered). This minimally requires moving from ordinary regression models to path models or structural equation models, and treating these in the ways suggested in this paper.

do for most practical cases. Medical interventions, for example, often fail (patients refuse to stop smoking), often interact with other causal factors (which explains why pharmacists require substantial training before licensing), often impact on multiple variables (organs) and often, even when successful, fail to put any variable into exactly one state (indeterminism!). Hence, we now provide a more general definition of intervention (retaining, however, reference to a single target variable in the system; this is a simplifying assumption which can easily be discharged).

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

1. *$I_c$ is introduced with the intention of changing $C$.*
2. *$I_c$ is exogenous in $M'$.*
3. *$I_c$ directly causes (is a parent of) $C$.*

We take it that interventions are *actions* and, therefore, intentional. In particular, there will be some intended *target distribution* for the variable $C$, which we write $P^*(C)$. $I_c$ itself will just be a binary variable, reflecting whether an intervention on $C$ is attempted or not. However, this definition does not restrict $I_c$'s interaction with $C$'s other parents, leaving open whether the target distribution is actually achieved by the intervention. Also, the definition does allow variables other than $C$ to be directly caused by $I_c$; hence, anticipated or unanticipated side-effects are allowed.

## 5 CATEGORIES OF INTERVENTION

We now develop this broader concept of intervention by providing a classification of the different kinds of intervention we have alluded to above. We do this using two "dimensions" along which interventions may vary. The result of the intervention is the adoption by the targeted variable of a new probability distribution over its states (even when a single such state is forced by the intervention, when the new probability distribution is degenerate), whether or not this achieved distribution is also the target distribution. To be sure, the new distribution will be identical to the original distribution when the intervention is not attempted or is entirely ineffectual. This special case can be represented

$$P_{M'}(C|\pi_c, \neg I_c) = P_M(C|\pi_c) \tag{1}$$

where $\pi_c$ is the set of the original parents of $C$.

**Dimensions of Intervention**

1. The degree of *dependency* of the effect upon the existing parents.
   (a) An entirely independent intervention leads to an achieved distribution which is a function only of the new distribution aimed for by the intervention. Thus, for an independent intervention, we have

$$P_{M'}(C|\pi_c, I_c) = P^*(C) \tag{2}$$

(b) A dependent intervention leads to an achieved distribution which is a function of both the target distribution and the state of the variable's other parents.

An independent intervention on $C$ simply cuts it off from its parents. Dependent interventions depend for their effect, in part, on the pre-existing parents of the target variable. The dependency across the parents, including the new $I_c$, may be of any variety: linear, noisy-or, or any kind of complex, non-linear interaction. These are precisely the kinds of dependency that Bayesian networks model already, so it is no extension of the semantics of Bayesian networks to incorporate them. Rather, it is something of a mystery that prior work on intervention has ignored them.

2. Deterministic versus stochastic interventions.
   (a) A deterministic intervention aims to leave the target variable in one particular state — i.e., the target distribution is extreme.
   (b) A stochastic intervention aims to leave the target variable with a new distribution with positive probability over two or more states.

A deterministic intervention is by intention simple. Say, get Fred to stop smoking. By factoring in the other dimension, allowing for other variables still to influence the target variable, however, we can end up with quite complex models. Thus, it might take considerable complexity to reflect the interaction of a doctor's warning with peer-group pressure.

The stochastic case is yet more complex. For example, in a social science study we may wish to employ stratified sampling in order to force a target variable, say age, to take a uniform distribution. That is an independent, stochastic intervention. If, unhappily, our selection into experimental and control groups is not truly random, it may be that this selection is related to age. And this relation may induce any kind of actual distribution over the targeted age variable.

Any non-extreme actual distribution will be subject to changes under Bayesian updating, of course, whether it is for a targeted variable or not. For example, a crooked Blackjack dealer who can manipulate the next card dealt with some high probability, may intervene to set the next deal to be an Ace with probability 0.95. If the card is later revealed to be an Ace, then obviously that probability will revised to 1.0.

Most interventions discussed in the literature are independent, deterministic interventions, setting $C$ to some one specific state, regardless of the state of $C$'s other parents. We can call this sort of intervention Pearlian, since it is the kind of intervention described by Pearl's "do-calculus" [10]. This simplest kind of intervention can be represented in a causal model simply by cutting all parent arcs into $C$ and setting $C$ to the desired value.

## 6 MODELING EFFECTIVENESS

There is another "dimension" along which interventions can be measured or ranked: their effectiveness. Many attempted interventions have only some probability, say $r$, of taking effect — for example, the already mentioned fact that doctors do not command universal obedience in their lifestyle recommendations.

Now, even if such an intervention is of the type that when successful will put its target variable into a unique state, the attempt to intervene will not thereby cut-off the target variable from its parents; it is not Pearlian. The achieved distribution will, in fact, be a mixture of the target distribution and the original distribution, with the mixing factor being the probability $r$ of the intervention succeeding.

Classifying or ranking interventions in terms of their effectiveness is often important. However, we have not put this scale on an equal footing with the other two dimensions of intervention, simply because it is conceptually derivative. That is, any degree of effectiveness $r$ can be represented by mixing together the original with the target distribution with the factor $r$. In case the intended intervention is otherwise independent of the original parents, we can use the equation:

$$P_{M'}(C|\pi_c, I_c) = r \times P^*(C) + (1 - r) \times P_M(C|\pi_c) \tag{3}$$

This being a function of all the parents of $C$, it is a subspecies of dependent interventions.

In practical modeling terms, to represent such interventions we maintain two Bayesian networks: one with a fully effective intervention and one with no intervention. (Note that the first may still be representing a dependent intervention, e.g., one which interacts with the other parents.) There are then two distinct ways to use this mixture model: we can do ordinary Bayesian net propagation, combining the two at the end with the weighting factor to produce new posterior distributions or expected-value computations; or, if we are doing stochastic sampling, we can flip a coin with bias $r$ to determine which of the two models to sample from.

## 7 REPRESENTING INTERVENTIONS

Any Bayesian network tool can be used to implement interventions just by generating the augmented model manually, as in Section 4.[6] However, manual edits are awkward and time consuming, and they fail to highlight the intended causal semantics. Hence, we have developed a program, the *Causal Reckoner*, which runs as a front-end to the BN tool Netica [8] [7].

The *Causal Reckoner* makes Pearlian interventions as easy as observing a node and implements more sophisticated interventions via a pop-up, and easy to use, GUI. The mixture modeling representation of effectiveness (§6) is implemented via a slider bar, and the target distribution is set by gauges. The full scope of possible interventions is not yet implemented (e.g., causally interactive interventions), as this requires arbitrary replacement of a node's CPT.

Our program provides better visualization and intervention features than any other we have seen. Indeed, *Genie* [5] is the only program with similar capabilities that we know of; it has the feature of 'controlling' nodes to perform Pearlian interventions. Our visualization for basic interventions is shown in Figure 2(b) in

[6] Alternatively, decision nodes can be used to model Pearlian interventions, since their use implies the arc-cutting of such interventions. However, that is an abuse of the semantics of decision nodes which we don't encourage.

[7] The software can be downloaded from: http://www.datamining.monash.edu.au/cgi-bin/cgiwrap/mdmc/run-cvstrac.cgi/causal/wiki.

**Fig. 4.** (a) The hypertension causal model where a stochastic medical intervention has been made. In (b) an observation has also been entered.

Section 3. The node is shaded and a hand icon (for "manipulation") is displayed. We don't show the intervention node, simplifying and saving screen space.

When visualizing less than fully effective interventions, it is useful to report extra information. Figure 3(a) shows a 90% effective intervention intended to set low blood pressure at age 50. The target distribution is shown to the right of the node's actual distribution, which is a mixture of the original and target distributions. In the hypertension example, the intervention can be interpreted as a drug which fails in its effect 10% of the time. A drug with a weaker effect is shown in Figure 3(b).

Even a fully effective intervention can result in an actual distribution that deviates from the target distribution. This can happen when the intervention is stochastic, since other observational evidence also must be incorporated. Figure 4(a) shows the hypertension example given a fully effective stochastic intervention. Take a drug that sets the chance of low blood pressure to 95%, irrespective of other causal influences. This particular drug reduces the chances of CHD from 0.052 to 0.038. But what if the patient gets CHD anyway? Figure 4(b) shows that under this scenario, it is less likely that the drug *actually* helped with hypertension, since people with hypertension are more susceptible to CHD than others.

In short, the *Causal Reckoner* provides a GUI for mixing observations and interventions seamlessly. We can take existing networks in any domain and investigate various intervention policies quickly, without the trouble of creating new nodes and manually rewriting arbitrarily large CPTs.

## 8 CONCLUSION

Recent research exploring the causal interpretation of Bayesian networks has been very fruitful. However, the theory needs to find its way into practical application. For that purpose, tools such as the *Causal Reckoner* are needed to make

it easy to model causal interventions and reason about their consequences and more difficult to make blunders, such as substituting an observational value for an intervention value.

In addition to these virtues of our work, we believe the nearly universal tendency to focus on deterministic models and deterministic interventions, while in part motivated by a healthy preference for the simple, either dismisses whole regions of potentially important applications or else invites new blunders in oversimplifying them. By taking seriously the indeterminism of the probabilistic relations in Bayesian networks, we have readily found a variety of intervention models that Pearlian interveners have yet to consider, including partially effective interventions, stochastic interventions and causally interactive interventions. Furthermore, it is clear that a great many real systems exhibit just these features.

**Acknowledgements** We thank Charles Twardy for helpful discussions.

## References

[1] Keaven M. Anderson, Patricia M. Odell, Peter W.F. Wilson, and William B. Kannel. Cardiovascular disease risk profiles. *American Heart Journal*, 121:293–298, 1991.

[2] Nancy Cartwright. What is wrong with Bayes nets? *The Monist*, 84:242–264, 2001.

[3] D. Chickering. A transformational characterization of equivalent Bayesian network structures. In D. Poole P. Besnard and S. Hanks, editors, *Proc of the 11th Conference on Uncertainty in AI*, pages 87–98, San Fransisco, CA, 1995. Morgan Kaufmann.

[4] R.B. D'Agostino, M.W. Russell, and D.M. Huse. Primary and subsequent coronary risk appraisal: new results from the framingham study. *American Heart Journal*, 139:272–81, 2000.

[5] Marek J. Druzdzel. SMILE: Structural modeling, inference, and learning engine and GeNIe: A development environment for graphical decision-theoretic models. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI–99)*, pages 902–903, Orlando, FL, July 18–22 1999.

[6] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach — Part I: Causes. In J. Breese and D. Koller, editors, *Uncertainty in AI*, pages 194–202, 2001.

[7] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach — Part II: Explanation. In *IJCAI '01*, 2001.

[8] Norsys. Netica. http://www.norsys.com, 2000.

[9] J. Pearl. Statistics, causality, and graphs. In A. Gammerman, editor, *Causal Models and Intelligent Data Management*, pages 3–16. Springer, Berlin, 1999.

[10] J. Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, Cambridge, UK, 2000.

[11] Wesley Salmon. Probabilistic causality. *Pacific Phil Qtly*, 61:50–74, 1980.

[12] P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction and search.* Springer-Verlag, New York, 1993.

[13] Patrick Suppes. *A Probabilistic Theory of Causality.* Amsterdam, 1970.

[14] James Woodward. Causation and manipulability. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Fall 2001.