Interventions

Frederick Eberhardt
Carnegie Mellon University
fde@cmu.edu

Abstract

This note, which was written chiefly for the benefit of psychologists participating in the Causal Learning Collaborative,[1] lays out a general framework for interventions. I provide a minimal definition of an intervention and two extensions of this definition, one of which corresponds to the notion of randomization (or Pearl's notion of "surgical" intervention), while the other is closely related to the idea of instrumental variables (and Korb's notion of "dependent" interventions). The framework provides the foundation for an analysis of discovery by experiment with different types of interventions. I also consider how different types of interventions interact with background assumptions to provide tools for the discovery of causal structure. I include at the end a discussion of three philosophical issues germane to interventions: the semantics of policy variables, the assumption of exogeneity, and interventions and free will.

1 Introduction

Experimental interventions and manipulations are the standard for discovering causal relations or confirming causal hypotheses, and causal hypotheses in turn make claims about the results of interventions, even if only of very hypothetical interventions. But experiments can be carried out in many ways, with many different kinds of interventions, producing different kinds of information. This paper aims to provide a framework for classifying interventions that is useful for understanding the information available from various experimental arrangements.

[1]The Causal Learning Collaborative Initiative funded by the James S. Mc Donnell Foundation and led by Alison Gopnik at UC Berkeley brings together psychologists, cognitive scientists and philosophers from the University of Washington (Seattle), UC Berkeley, Caltech, University of Michigan (Ann Arbor), Carnegie Mellon University and MIT to study causal learning in humans. The author is supported by a fellowship from this collaborative.

2 Background on Interventions as Tools of Discovery

Before I give a formal account of interventions I will outline some of the historical developments on interventions as a discovery tool. For this preliminary discussion I will take an intervention to loosely mean some external cause of a (subset of) variable(s), i.e. the intervention has a causal influence on some variable under investigation, but is uncaused by any such variable. Although I will not stick to this view, I will now – for intuition's sake – suppose that there is some form of control over the form the intervention may take. The basic idea is that the distribution of values of variables subject to an intervention is changed from its unmanipulated, passive observational distribution by an "external" influence.

Galileo's astronomical observations and experiments with inclined planes illustrate the contrast between passive observational and experimental discovery. In the case of the astronomical observations Galileo was able to record the positions of some of Jupiter's moons and noted that their periodic appearance and disappearance was consistent with an orbit around Jupiter. Galileo was able to develop hypotheses about the orbits and derive testable predictions but was limited to observing and making inferences about a system he had no control over.

In contrast, in his experiments with inclined planes Galileo was able to carefully control size, weight and initial velocity for each object on the plane, as well as the length and inclination angle of the plane. By fixing all the other variables to particular values he could test whether a change in weight resulted in a change in the acceleration of the object on the inclined plane (as suggested by Aristotle). In modern terminology we would say that Galileo fixed or clamped all but one potential cause variable, varied the one remaining potential cause variable and measured the difference in the outcome variable. While the strategy of this procedure is evident, Galileo does not give an explicit account of the methodological role or the advantages of the interventions (clamping or varying variables) to the aim of discovery. The implicit argument is that changes in the outcome could only have arisen due to changes in the causal influence of the varied variable, since all other variables are held constant. While Galileo's experiments are surely not the first instantiation of such a method, they are the earliest we have good records for.

First accounts of a methodology of causal discovery appear in Bacon [1]. Bacon suggested that in order to find the cause of a particular phenomenon one should construct two lists, one of positive and one of negative instances. The list of positive instances should be ordered by increasing degree of the occurrence of the phenomenon. The cause of the phenomenon is then the set of properties present in all the positive instances and absent in all the negative instances and the intensity of the properties increases according to the ordering in the list of positive instances. Bacon does not distinguish between observing the instances and bringing particular circumstances about artificially, so it seems as if his

method restricts itself to passive observational discovery.

Much later Mill develops a very similar methodology for causal discovery with his experimental methods of agreement and difference [11]. Mill is aware of the distinction in implementation of passive observation vs. intervention but is ambiguous about the difference in the epistemic access the two approaches provide.[2] Mill puts forth five canons:

First Canon (Method of agreement): If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all instances agree is the cause (or effect) of the given phenomenon.

Second Canon (Method of difference): If an instance in which the phenomenon under investigation occurs and an instance in which it does not occur have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

Third Canon: If two or more instances in which the phenomenon occurs have only one circumstance in common, while two or more instances in which it does not occur have nothing in common save the absence of that circumstance, the circumstance in which alone the two sets of instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

Third Canon (Method of residues): Subduct from any phenomenon such part as is known by previous inductions to be the effect of certain antecedents, and the residue of the phenomenon is the effect of the remaining antecedents.

Fifth Canon (Method of concomitant variations): Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.

Mill indicates that the first two Canons essentially embody Bacon's methodology,[3], however he takes the method of difference to embody more explicitly the controlled experimental design that Galileo used with the inclined planes,

[2] "For the purpose of varying the circumstances, we may have recourse (according to a distinction commonly made) either to observation or to experiment; we may either find an instance in nature suited to out purposes or, by an artificial arrangement of circumstances, make one. [...] There is, in short, no difference in kind, no logical distinction, between the two processes of investigation. There are, however, practical distinctions to which it is of considerable importance to advert." ([11], p. 211) But contrast this quote with: "But if we cannot artificially produce the phenomenon A, the conclusion that it is the cause of a remains subject to very considerable doubt. [...] Unfortunately, it is hardly ever possible to ascertain all the antecedents unless the phenomenon is one which we can produce artificially." ([11], p. 213)

[3] [11], p. 216

although he does not refer to Galileo in particular.[4] However, while Mill speaks of artificial experiments, there is no explicit discussion of whether the other potential cause variables should be *clamped* (i.e. held fixed) at some particular value or whether one should aim through careful experimental design to increase the likelihood of obtaining corresponding samples where the other variables *incidentally happen* to have the same values. This distinction is now known as the difference between *statistically conditioning* on a variable as opposed to *clamping* the variable and it is doubtful whether Mill was aware of the difference.

If Mill could assume that the experimenter is dealing with a causally sufficient set of variables (i.e. there are no unmeasured common causes) and if he could ensure that his matching of samples does not amount to conditioning on a common effect of two variables, then the epistemological difference between clamping and statistically conditioning disappears anyway, since both can be used to isolate the causal connection between one potential cause-effect pair. However, using conditioning only may still pose a significant data collection problem.[5] It does not seem plausible that Mill was aware of these aspects. In a discussion of the limitations of his methodology Mill hints at the problem of causally insufficient sets of variables.[6] But he does not discuss the methodological value of the assumption of causal sufficiency and provides no principled method to ensure either that causal sufficiency is satisfied or that conditioning on common effects is avoided, or what one ought to do if either assumption fails or is not known to be satisfied.

In 1935 R.A. Fisher lays out in detail the methodological aspects of discovery using randomized trials in The Design of Experiments.[8]. Fisher had realized that if one randomized the values of the purported cause variable, the *treatment*, one could break any correlation due to latent common causes of the treatment and outcome variables, thereby removing spurious correlations. This insight has a further consequence not discussed by Fisher that will become relevant to the more general problem of causal discovery: Such a randomized intervention can distinguish causal direction. If it is known that A causes B or B causes A, but not which, a randomization of one of the variables, say A, can distinguish the two cases. In the first case, it would make the two variables independent, whereas they would appear dependent in the second. Fisher focuses on uniform distributions over the values of the treatment variable, but the result applies to any distribution that is placed over the values of the treatment variable as long as it makes the intervened variable independent of its normal causes. To Fisher, the most important aspect of this "intervention distribution" on the treatment variables is that it provides a known reference distribution that together with an

[4] "Of these methods, that of difference is more particularly a method of artificial experiment, while that of agreement is more especially the resource employed where experimentation is impossible." [11], p. 216

[5] Some constellations of variable values might be very rare if one does not force them by clamping.

[6] "In other cases, when we intend to try an experiment, we do not reckon it enough that there be no circumstance in the case the presence of which is unknown to us. We require, also, that none of the circumstances which we do know shall have effects susceptible of being confounded with those of the agents whose properties we wish to study." [11], p. 251

assumption about the functional form of the distribution of the outcome variable allowed the estimation of statistical parameters representing the strength of the causal influence of the treatment on the outcome.

Fisher's insight led to a vast development of experimental designs involving randomized trials. Fisher's basic idea of one treatment and one outcome variable is extended to sets of treatment and outcome variables and provides results on the optimal assignment of values to the treatment variables that would enable an efficient discovery of the dependencies. These experimental designs are known as Latin Squares, Graeco-Latin Squares and Factor experiments. All these designs preserve the bipartite separation of the variables into a set of treatment and outcome variables and do not – at least not in any principled manner – address cases that do not fit such a framework. His methods for searching for causal relations-which Fisher takes to be the very center of scientific inquiry– is limited to a very narrow set of structures. The aim of the experiment is to determine whether or not the treatment variable has an effect on the outcome variable, and if so, how strong it is. Causal dependencies within the set of treatment or effect variables are not considered, or considered only in an ad hoc way (e.g. Rubin [13, 14, 15]).

On the basis of these developments – and despite their limitations – randomized trials have become the golden standard to test the efficacy of new treatments in medical research. This may be due to the fact that a bipartite separation of the set of variables is supported in this field (e.g. due to time ordering information), but one can certainly imagine many cases even in this domain where such restrictions are unwarranted (e.g. when there are several causal pathways between the treatment and outcome variables).

The generalization of Fisher's idea was not developed in statistics but rather in a branch of computer science and philosophy: theories of interventions in graphical models. In contrast to the structures assumed to underlie the models in statistics, the only structural assumption made for theories of intervention on causal Bayes nets is acyclicity - and even this restriction has been lifted in some cases. The framework allows for the representation of interventions (as in the case of randomized experiments) and the computation of their effects.

Pearl [12] developed a theory of interventions called the do-calculus, which was inspired by computational features of a simpler and more general theory provided by Spirtes, Glymour & Scheines (SGS) [16]. Both theories capture Fisher's insight of an intervention that makes the intervened variable independent of its normal causes. Randomized trials and Latin Square designs can be modeled easily, but the framework is less restrictive the possible underlying structures and allows for different types of interventions. The detailed representation of causal systems in the Bayes net framework in general and the representation of interventions in particular led to a new approach to the notion of causality in the philosophical community. Woodward and Hitchcock [17, 9] have argued in a related vein that *interventions* are really the key to an account of causality in general. Pearl does not give any metaphysical account of causation over and above the technical machinery of what it is to stand in a cause-effect relation. Spirtes, Glymour & Scheines simply take a cause to be a primitive

notion. Woodward [17], on the other hand, appears to take an intervention as the more fundamental notion. He specifies the following two conditions, the first he considers sufficient for causation and the second necessary:

(SC) *If (i) there is a possible intervention that changes the value of X such that (ii) carrying out this intervention (and no other interventions) will change the value of Y, or the probability distribution of Y, then X causes Y.*

(NC) *If X causes Y then (i) there is a possible intervention that changes the value of X such that (ii) if this intervention (and no other interventions) were carried out, the value of Y (or the probability of some value of Y) would change.*

The difficulty of such conditions is (a) to determine the appropriate notion of possibility – which Woodward discusses at length – and (b) what is required by the claims "the intervention (on X) *will* change Y" and "if the intervention were carried out the value of Y *would* change". There has to be something akin to an implicit *ceteris paribus* clause as provided in the following definition:

Definition 2.1 (Direct Cause) A is a cause of B if there exists an assignment α of values to all other variables, such that if those variables are *clamped* at α, there exists a change of values in A which will result in a change of values in B.

However, in such a reconstruction of Woodwards account we have arguably both an intervention on A, namely "the change of values in A", and an intervention on all other variables except A and B in the form of clamping.[7] As Woodward has argued ([17, 18]), it is not clear how we can make sense of interventions without the notion of cause. But if we describe interventions in terms of causes and causes in terms of interventions, then we had better make sure that we are dealing with a virtuous circle and not a vicious one.

This brief overview suggests that the role of interventions is a relatively recent addition to the ancient discussion of causation. Without doubt interventions have been used for causal discovery all along, but formal accounts and the recognition of the specific qualities of interventions are relatively new (by philosophical standards). In addition, the insights on this matter have come from a variety of fields: from first scientific experiments (Galileo), first formalizations of experimental methods in philosophy (Bacon, Mill), randomized trials in statistics (Fisher), general representation of causal models in form of causal Bayes nets in computer science and philosophy (Pearl; Sprites, Glymour & Scheines) and philosophical accounts of causality in terms of interventions (Woodward).

[7] One could argue that "the change of values in A" does not require a separate intervention, as long as (i) such a change is passively observed while all other variables are clamped, and (ii) there are no latent common causes of A and B.

3 Interventions

In this section I will give a more precise account of interventions. Defining an intervention within the causal Bayes net framework is relatively straight forward, but there are several philosophical and technical issues that arise for any real world intervention that is supposed to satisfy the conditions of a formal intervention as described here. I will first present the formal definitions, but will attempt to address some of the philosophical concerns in a section at the end of this chapter.

3.1 Formal Definitions

The causal Bayes Net framework allows for the representation of several different types of interventions. Different interventions place different restrictions on how they can be applied and constrain what can be learned from a system they are applied to. To distinguish interventions from normal causal variables, I will require that interventions are *exogenous*.

Definition 3.1 (exogenous) A variable X is *exogenous* to a set of variables $\mathbf{V}$ if $X \notin V$ and there does not exist a variable $Y \in \mathbf{V}$ such that Y is a cause of X.

The following then provides a minimal definition of an intervention in the causal Bayes net framework:

Definition 3.2 (Intervention) Given a set of measured variables V, an intervention I on a subset $S \subseteq V$ satisfies the following criteria:

1. $I \notin V$ is a variable with two states $(1/0)$[8] representing that the intervention is either active or inactive,
2. I is a direct cause of each variable $X \in S$,
3. I is exogenous to V.
4. When $I = 0$, the passive observational distribution over V obtains, i.e.

$$\begin{aligned} P(V|I=0) &= P(V) \\ &= \prod_{V_i \in V} P(V_i|pa(V_i)) \\ &= P(S|pa(S)) \prod_{V_i \in V \setminus S} P(V_i|pa(V_i)) \end{aligned}$$

[8]The number of states of the intervention variable may be increased if one wants to represent different forms of intervention, i.e. different manipulated distributions over the intervened variable. In this case one would not just have one state to represent that the intervention is occurring, but several "on"-states that represent different ways of performing the intervention. Essential to an intervention is that it has at least one active state ($I = 1$) and can be switched off to be completely ineffectual ($I = 0$). I return to this issue in the discussion on instrumental variables and the semantics of policy variables.

5. When $I = 1$, the conditional distribution over S is manipulated, i.e.

$$P(V|I = 1) = P(S|pa(S), I = 1) \prod_{V_i \in V \setminus S} P(V_i|pa(V_i))$$

where

$$P(S|pa(S), I = 1) = \prod_{X \in S} P^*(X|pa(X))$$

and for each $X \in S$ we have

$$P^*(X|pa(X)) \neq P(X|pa(X), I = 0)$$

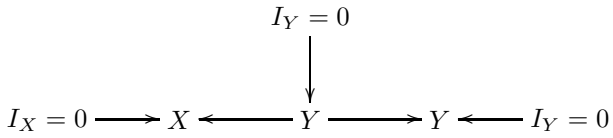
An intervention variable is represented in a causal Bayes net as an additional variable with direct arrows into each variable in S.

3.1.1 Example

Suppose the true graph with its probability distribution is:

$$X \longleftarrow Y \longrightarrow Z \qquad P(X, Y, Z) = P(Y)P(X|Y)P(Z|Y)$$

This representation is identical to the graph that includes the intervention variables set to 0, i.e.



However, if the intervention variables are 0 then we do not include them for simplicity of representation. There could be many different types of intervention variables connected to many different variables, which would be redundant and clutter the representation if they are not used. If we intervene on X, i.e. set $I_X = 1$, then we have

$$I_X = 1 \longrightarrow X \longleftarrow Y \longrightarrow Y$$

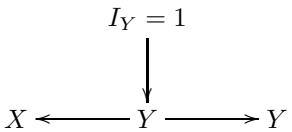
$$P(X, Y, Z|I_X = 1) = P(Y)P(X|Y, I_X = 1)P(Z|Y)$$

where

$$P(X|Y, I_X = 1) \neq P(X|Y) = P(X|Y, I_X = 0)$$

Below I specify how these factors have to differ for the different types of interventions, but for the general definition I only require that they differ.

Note, that if we intervened on Y, i.e. set $I_Y = 1$, then we would get the following:



$$P(X, Y, Z|I_Y = 1) = P(Y|I_Y = 1)P(X|Y)P(Z|Y)$$

where

$$P(Y|I_Y = 1) \neq P(Y) = P(Y|I_Y = 0)$$

3.1.2 Discussion: Intervention

The above definition should be understood as an absolutely minimal definition of an intervention. In most cases, as will be described in more detail below, additional restrictions are placed on interventions. But it is important to notice that these additional restrictions are by no means necessary. In particular this minimal definition does not specify the following features:

1. An intervention can affect more than one variable in V, i.e. S need not be a singleton set. If $|S| > 1$ we say the intervention is *confounding*.[9]

2. When $I = 1$, there are (virtually[10]) no restrictions on the probability distribution I can assign to the variables in S.

3. There may be *latent* (unmeasured) causes of an intervention.[11]

4. When $I = 1$, it need not make the variables in S independent of their causes (although this will be true for "structural" interventions).

5. The intervention variable need not have a well-defined marginal distribution over its values. If it does not have a distribution over its values, we will refer to such a variable as a *policy variable* as it then represents a decision point. Of course, in any sample, there will be a (marginal) sample distribution for the intervention variable, but there is no commitment to the existence of a marginal population distribution for intervention variables.[12]

Given the above minimal definition of interventions and their representation in causal Bayes nets we can now proceed to consider more specific types of interventions.

3.2 Structural Intervention

Interventions which make the intervened variable independent of its other causes are sometimes referred to as *randomizations* (following Fisher), *surgical* interventions (following Pearl), *ideal* interventions (following SGS) or *independent*

[9] See section on confounding interventions below.

[10] The only restriction is that the distribution has to be such that I is a cause of each variable in S, which means that for each $X \in S$, $P(X|pa(X), I = 1) \neq P(X|pa(X), I = 0)$.

[11] See the discussion of exogeneity below.

[12] See the discussion on the semantics of intervention variables below.

interventions (following Korb). I will refer to them as *structural* interventions, because they manipulate the causal structure among the variables.[13]

Definition 3.3 (Structural Intervention) Given a set of measured variables V, a *structural* intervention I_s on a subset $S \subseteq V$ is an intervention on S that satisfies the following additional constraint:

1. When $I_s = 1$, I_s makes every variable in S independent of its causes (breaks the edges that are incident on the variables in S). I_s *determines* the distribution of S, that is, in the factored joint distribution $P(V)$, the term $P(S|pa(S))$ is replaced with the term $P(S|I_s = 1)$, all other terms are unchanged.

The definition of a structural intervention implies that the causal structure (as opposed to just the parameterization) is manipulated, since any causal influence on the intervened variable (other than from the intervention) is destroyed. The causal structure after the intervention is referred to as the *post-manipulation graph*.

Definition 3.4 (Post Manipulation Graph) Given a graph G and a set of interventions I, the post-manipulation graph is the graph where all the edges incident on any intervened variable are removed.

The structural change goes along with a change in the probability distribution over the variables given by the manipulation theorem for structural interventions:

Theorem 3.5 (Manipulation Theorem for Structural Interventions) [14] *Let $G = \{V, E\}$ be a directed acyclic graph and let $\mathbf{I}$ be the set of variables in V that are subject to a structural intervention. Then G_{unman} is the unmanipulated graph corresponding to the unmanipulated distribution $P_{unman}(V)$ and G_{man} is the manipulated graph, in which for each variable $X \in \mathbf{I}$ the edges incident on X are removed and an intervention variable $I_{s(X)} \rightarrow X$ is added. A variable $X \in V$ is in $man(\mathbf{I})$ if it is subject to an intervention, i.e. if it is a direct child of an intervention variable $I_{s(X)}$. Then*

$$P_{unman(\mathbf{I})}(V) = \prod_{X \in V} P_{unman(\mathbf{I})}(X|pa(G_{unman}, X))$$

[13] It is not my love for the proliferation of terminology that makes me suggest a new term, in fact I am generally in full support of following clear existing terminology. However, "randomizations" is too vague, since the parametric interventions I introduce below are in some sense just dependent randomizations. "Surgical" is a great term, but leads to confusion if one wants to write within the medical domain of experimental design. "Ideal" is too vague, since there are too many things one might want to be ideal about interventions, and misleading, since it suggests some kind of optimality. "Independent" is also good, but I find it is easily confused with what I describe here as exogeneity or with non-confounding interventions.

[14] This is taken directly from [16].

$$P_{man(\mathbf{I})}(V) = \prod_{X \in man(\mathbf{I})} P_{man(\mathbf{I})}(X | I_{s(X)} = 1) \times \prod_{X \in V \setminus man(\mathbf{I})} P_{unman(\mathbf{I})}(X | pa(G_{unman}, X))$$

for all values of V for which each of the conditional distributions is defined.

3.2.1 Discussion: Structural Intervention

This definition of structural interventions is still very general, since it underdetermines the following points:

1. I_s can affect more than one variable. However, generally I_s is assumed to affect only one variable.

2. There are (virtually) no restrictions on the probability distribution I_s assigns to the variables in S as long as the distribution for each $X \in S$ makes X independent of its causes. Generally, a non-zero probability for each assignment of values to variables $X \in S$ is desirable in order to ensure that it is possible to discover every causal effect (e.g. in the case of interactive causes). In most experiments involving randomizations where the treatment variables have a finite number of values, the manipulated distribution of each value of each $X \in S$ is uniform, if X is infinite-valued, Gaussian distributions are common.

Fisher-type randomized trials, as they are found in medical research, are the best known example of structural interventions. Participants are assigned randomly to the treatment or control group. Generally, the assignment is done by some mechanism essentially representing a coin flip: The probability of being assigned to the treatment or control group is (usually) equal, and the coin flip is assumed to be causally independent of any features of the study participant that may be relevant to the study outcome.

3.2.2 Clamping

The literature on experimental design often suggests that one way to discover whether X is a direct cause of Y is to hold all other variables in V fixed at particular values v (to clamp the variables), and then randomize X and see whether Y covaries with X. If it does for some set of values v assigned to $V \setminus \{X, Y\}$, then X is a direct cause of Y.[15]

[15] Clamping a variable Z to a particular value z_1 is of course to be distinguished from conditioning on $Z = z_1$. In the case of clamping, an intervention is required that sets and holds Z at z_1 during the experiment independently of what values the other variables assume. In the case of conditioning, Z can assume many different values in the course of the experiment. However, if we find that $X \not\perp\!\!\!\perp Y | Z = z_1$, then this is not an indication of a direct cause between X and Y, but may be due to the following causal structure: $X \rightarrow Z \leftarrow Y$.

Clamping can be modeled as a degenerate form of a structural intervention. It makes the intervened variable independent of its parents in the graph, but the manipulated distribution assigns all its probability to one value of the intervened variable.

The problem with such a degenerate intervention is with interactive causes: If X only has an effect on Y when Z is in a particular state $Z = z_1$, then it is of no use to clamp $Z = z_2$. In search for causal structure we do not know a priori which causes are interactive and which are not.

3.3 Parametric Interventions

Structural interventions are not the only possible type of intervention. It is not necessary that an intervention makes a variable independent of its causes, it just needs to have an influence on the conditional distribution. This weaker form of an intervention is captured in the notion of a *parametric* intervention, also sometimes referred to as a *partial*, *conditional*, *soft* or *dependent* intervention.

Definition 3.6 (Parametric Intervention) Given a set of measured variables V, a *parametric* intervention I_p on a subset $S \subseteq V$ is an intervention on S that satisfies the following constraint:

1. When $I_p = 1$, I_p does not make the variables in S independent of their causes in V (it does not break any edges that are incident on variables in S).[16] In the factored joint distribution $P(V)$, the term $P(S|pa(S))$ is replaced with the term $P^*(S|pa(X), I_p = 1)$, where

$$P^*(S|pa(X), I_p = 1) \neq P(S|pa(X), I_p = 0).$$

   Otherwise all terms remain unchanged.

Although a parametric intervention does not imply any structural changes among the variables in V and the post-manipulation graph is only changed by the addition of the intervention variables, its influence is evident in the manipulated probability distribution.

Theorem 3.7 (Manipulation Theorem for Parametric Interventions) *Let $G = \{V, E\}$ be a directed acyclic graph and let $\mathbf{I}$ be the set of variables in V that are subject to a parametric intervention. Then G_{unman} is the unmanipulated graph corresponding to the unmanipulated distribution $P_{unman}(V)$ and G_{man} is the manipulated graph, in which for each variable $X \in \mathbf{I}$ an intervention variable $I_{p(X)}$ is added with $I_{p(X)} \rightarrow X$. A variable $X \in V$ is in man($\mathbf{I}$) if it is subject to an intervention, i.e. if it is a direct child of an intervention variable $I_{p(X)}$. Then*

$$P_{unman}(V) = \prod_{X \in V} P_{unman}(X|pa(G_{unman}, X))$$

[16] Note, that I restrict parametric interventions to those types of interventions that do not break any structure but instead *only* influence parameters.

$$P_{man}(V) = \prod_{X \in \mathbf{I}} P_{man}(X|pa(G_{unman}, X), I_{p(X)} = 1) \times \prod_{X \in V \setminus \mathbf{I}} P_{unman}(X|pa(G_{unman}, X))$$

for all values of V for which each of the conditional distributions is defined.

3.3.1 Discussion: Parametric Intervention

Again, the definition of a parametric intervention is not very restrictive:

1. I_p can affect more than one variable. However, generally I_p is assumed to affect only one variable.

2. There are (virtually) no restrictions on the probability distribution that I_p assigns to the variables in S as long as I_p is a direct cause of each variable in S and does not make S independent of its causes.

Since I_p does not make the variables in S independent of their causes (parents in the graph) I_p is not a structural intervention. Instead, I_p changes (and increases the number of) the parameters in the conditional distribution of the intervened variable on its parents.

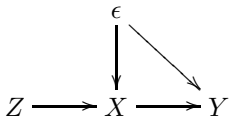
A simple example of a parametric intervention is an intervention on the income of participants in an experiment: Rather than setting their income according to an independent probability distribution, thereby determining it completely, a parametric intervention increases their income by, say, $1,000. This would have the effect that people with high incomes would still have high incomes, determined largely by the original causes for their high income, but we would have changed the conditional probability distribution, due to the influence of the intervention. It should be noted that it is not necessary for the parametric intervention to consist of adding a constant to the variable value. It is possible to perform a parametric intervention on binary variables as well – all that is required is that there is a change in the conditional probability distribution from the passive observational case such that:

$$P(X|pa(X), I_{p(X)} = 0) = P(X|pa(X)) \neq P(X|pa(X), I_{p(X)} = 1)$$

3.3.2 Parametric Interventions and Instrumental Variables

The technique of using parametric interventions in causal discovery is closely related to the theory of instrumental variables in economics. Suppose there are two variables, X and Y with $Y = \beta X + \epsilon$, where ϵ is an error term. The problem for the estimation of β arises when X is correlated with ϵ, the error. In such a case a consistent estimator of β can be found if there is a variable Z (called an instrument) that is correlated with X, but independent of ϵ and correlated with Y only through X, i.e. $Z \perp\!\!\!\perp Y | \{X, \epsilon\}$.

Graphically, this can be represented by assuming that ϵ is a latent common cause of X and Y, as shown in the following figure:



This structure mirrors the set-up for parametric interventions: If Z were an intervention variable of X, then we would have the same independence relations: The instrument is independent of ϵ and correlated with Y only through X, which is the same as requiring that the intervention is exogenous (and uncaused) with respect to the set $V = \{\epsilon, X, Y\}$. The independence relations implied by this particular constellation of variables make both instrumental variables and parametric interventions a powerful discovery tool even for causally insufficient sets of variables.

The difference between the two is mainly in the semantics: Instrumental variables are generally taken to be real variables, corresponding to some causally relevant feature in the real world, whereas intervention or policy variables are a feature of the model. It is generally assumed that there is a well-defined marginal population distribution over an instrumental variable, whereas intervention variables represent decision points that need not have such a distribution. Furthermore, there is generally no *requirement* for instrumental variables to have an "off"-state for which they have no effect on the set of variables under investigation. Unlike intervention variables, all of their states can have an active influence on the set of variables.[17]

4 Contrast of Structural and Parametric Interventions

Structural and parametric interventions are the two extremes on a continuum of weaker to harder interventions. The structural intervention makes the intervened variable independent of its causes whereas a parametric intervention is only an intervention on the parameterization of the causal model. There can be all sorts of other interventions that have a weaker or stronger effect on the structure or parameterization, making the intervened variable independent of more or fewer of its causes. The distinction of the two extreme forms of interventions as I have presented it here can be found in work by Korb [10] and the need for a weaker version than just the structural intervention is described with various examples by Campbell [2]. Korb also discusses the possibility of mixed interventions. In that case the the manipulated distribution is a mixture of two

[17] One may argue that this last point is an artefact of my stringent definiton of interventions, but if one wants to achieve a reasonably clean separation of interventions and ordinary causal variables, I think general intuitions require the possibility of a passively observed system distinct from a manipulated one.

manipulated distributions, one structurally manipulated and one parametrically manipulated. As he notes, these mixtures can be represented by manipulated distributions that are somewhere between structurally and parametrically manipulated ones. It shows that there is a wide variety of additional modelling assumptions one can make about the particular nature of the manipulated distribution. The effect (and problems) with regard to causal discovery in light of the different types of interventions are discussed in detail in my thesis.

The interventions I present here are all designed for static models, i.e. they do not work without adjustment for time series models or dynamic Bayes nets. Developing a full account of interventions (structural and parametric) for time series is one of my long term goals with this work. The difficulty is to account for how fast and for how long the effect of interventions percolates through a dynamic system. Further, one has to distinguish between an intervention at one time instance and a continuously occurring intervention. Furthermore, it is not guaranteed that the effect of an intervention will fade away, since one might have chaotic effects in a dynamic system.[18]

4.1 Basic Implications for Discovery

[*A far more elaborate treatment of the problem of search and discovery for causal structure is given in my thesis and various papers I have published [3, 4, 5, 6, 7]. However, the key difference for discovery algorithms with regard to different types of interventions hinges on the points I make here.*]

With regard to discovery of causal structure the efficacy of interventions depends on how much structure one can recover from a single experiment that may involve several simultaneous interventions.

A structural intervention destroys all edges incident on the intervened variable. For example, suppose that the true causal graph is $X \leftarrow Y$, then the figures below show the unmanipulated ($I_s = 0$) graph on the left and the manipulated (post-manipulation) graph on the right.

$$I_s = 0 \quad X \leftarrow Y \qquad\qquad I_s = 1 \rightarrow X \quad Y$$

[18] Here is a short list of problems that a full account of interventions on time series needs to consider:

1. What is the nature of an intervention on a time series?
2. Is there a single intervention at one time tick or is the intervention repeated at every time tick?
3. In order to draw inferences from data, does the data sampling have to be synchronized with the interventions?
4. Does the dynamic system return to an equilibrium state after an intervention? How is this ensured?
5. If the dynamic system is non-linear, its development might be sensitive to initial conditions and hence predictions of interventions may be impossible. How does the model accommodate this?

The edge from Y to X is removed when $I_s = 1$ since this cannot be recovered from sample data.

In the case of a parametric intervention, we do not destroy any edges (see below) because their existence can be detected and their direction determined from the data we obtain in an experiment.

$$I_p = 0 \quad X \leftarrow Y \qquad\qquad I_p = 1 \rightarrow X \leftarrow Y$$

The key is the unshielded collider[19] a parametric intervention creates between I_p, X and Y when $I_p = 1$. This will need some clarification, since this aspect is crucial to any efficient discovery procedure.

If our data contains samples *both* for when a variable X is subject to a structural intervention *and* when it is passively observed, then - using the sample distribution of I_s - we can find that even in the case of *structural* interventions, I_s, X and Y form an unshielded collider. The adjacency is obtained from the subsample where $I_s = 0$, while the direction is determined when $I_s = 1$.

Such a sample constitutes a mixture of populations, one manipulated and one unmanipulated, as we may find it in a randomized trial with a control group that is not subject to any intervention. However, in studies where *each* condition in the randomized trial involves a structural intervention[20] (e.g. comparative medical studies) we do not have such a passively observed control group that would capture the unmanipulated structure: While the intervention conditions are different (with regard to the manipulated distribution they impose), they both destroy causal connections incident on the intervened variable and consequently prevent the discovery of causal structure from the sample in cases as the above. Furthermore, if we perform multiple simultaneous structural interventions (e.g. structurally intervening on both X and Y in the above graph) and only obtain data where for any sample either *all* intervention variables are on or *all* intervention variables are off, then again we cannot recover the causal structure in cases as the above. We find the adjancency, but cannot direct it.

In contrast, even if we do not have samples from the passive observational distribution ($I_s = 0$), we can discover the causal structure if the intervention is *parametric* – as long as the sample contains two different active intervention states. Similarly, in the case of multiple simultaneous parametric interventions the causal structure can be found even when we only obtain data where for each sample either all intervention variables are on or all intervention variables are off.

The key is that since parametric interventions do not destroy structure, they can be combined independently and performed simultaneously without interfering, while this is not the case for structural interventions.

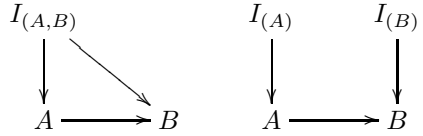
[19] Three variables X, Y and Z form a *collider* if Y is a common effect of X and Z, i.e. $X \rightarrow Y \leftarrow Z$. The collider is *unshielded* if X is not a cause of Z and Z is not a cause of X. Unshielded colliders can be discovered in the data, since they have the unique structural feature that implies that $X \perp\!\!\!\perp Z$ and $X \not\perp\!\!\!\perp Z|Y$.

[20] In this case the "on"-state ($I_s = 1$) of the intervention variable would have to be augmented to different on-states.

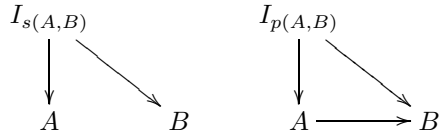
4.2 Confounding Interventions

The description of interventions suggests that one generally assumes that each intervention affects one variable only. But as the definitions show, there is no necessity for this assumption. We can model a structural or parametric intervention that intervenes on a set of variables in a correlated manner - we call such an intervention a *confounding* intervention, since it manipulates several variables like a common cause. If the confounding is not intended by the experimenter, such an intervention is also sometimes referred to as "fat-hand", since it affects variables in a way similar to someone who does not have sufficiently slim fingers to just manipulate one variable.

These confounding interventions, applicable to both the structural and parametric case, have to be distinguished from multiple simultaneous, but *independent* interventions. A confounding intervention has one policy variable that is a common cause of several variables in V and hence the variables will appear correlated when subject to an intervention even when they are not causally connected in the graph over just V (left, in figure below). In contrast, multiple simultaneous interventions have an individual policy variable for each variable in V that is subject to an independent intervention (right, in figure below).



Graph surgery rules and the manipulation theorems apply in the same way for confounding interventions as they do for individual interventions – depending on whether it is a structural or parametric intervention. Concretely, if we have a graphical structure where $A \to B$ and we perform a structural confounding intervention $I_{s(A,B)}$ on A and B, then the post-manipulation graph is shown below on the left: the edge $A \to B$ is destroyed by the structural intervention on B. For a parametric confounding intervention on the same causal structure we obtain the graph on the right.



5 Philosophical Issues germane Interventions

5.1 Semantics: Policy Variables

In the causal Bayes net framework the intervention variable I is represented as an additional variable, referred to as a policy variable. Formally, the only difference between a policy variable and an ordinary Bayes net variable is that

a policy variable need not have a well-defined distribution over its values. It contributes to or determines – depending on the type of intervention – the distribution over the intervened variable, but there need not be a distribution over its states. It represents a decision point of whether or not to intervene, or to perform a particular type of intervention if there are several options (in which case the policy variable has several "on"-states). In that sense one can speak of *setting* the policy variable. The decision as such, is not required or assumed to have a distribution, although its outcome in the sample will obviously have a sample distribution. The policy variable itself does not correspond to a real causally relevant feature of the world – at least not unless one wants to *model* the decision as a full causal variable that may be influenced in its own right.

Consequently, the particular nature of a policy variable depends on the reference frame in which we consider the causal discovery problem. As long as the intervention is exogenous, we do not have to refer to a distribution over the values of the policy variable. However, if one wants to enlarge the view then one might not want to say that the intervention is *uncaused* per se. Instead, in a larger reference frame, say "God's causal Bayes net", there might be factors which could be said to cause an intervention and then a distribution over the values of the policy variable would be appropriate and meaningful. The decision point represented by the policy variable would become (in the model) a regular causal variable. In this sense the policy variable marks the frame of reference for the causal system we are considering and is consequently relative. It can, but need not be seen, as a point where agency is required.[21]

5.2 Exogeneity

The assumption of exogeneity captures the notion that an intervention brings in something external to the system under investigation. If an intervention is caused by some variable in the system then it is unclear in what sense we want to call it an intervention at all. But exogeneity, as I have defined it here, is a rather weak assumption. It does not assume in principle that the intervention is uncaused, which might lead to some requirement that an intervention requires agency or free will (see next section). This is not the case. Exogeneity only requires that no variable in the system under consideration, i.e. no variable in V, causes the intervention. Whether or not something is exogenous depends on the set of variables we are looking at. We might consider a machine to be performing an intervention on a set of test tubes if it adds a particular chemical to them. However, if we look at the larger apparatus, we might want to say that this intervention was caused by some, say, timing mechanism. And, for that matter, the timing mechanism might also have other causal effects in the system or on the test tubes. There is no harm in assuming that there are unmeasured common causes of the intervention and variables in V. But we do want to contrast this case from a situation where the state of the test tubes, say, the color of their content, triggers some sensor which activates the addition

[21] See section on free will.

of the chemical. In this case, adding the chemical would not be an intervention since it is caused by a variable under investigation. This is exactly the case that the assumption of exogeneity rules out. That is to say, one and the same mechanism may be an intervention in some systems while it is not in others. But that question is independent of whether there is a cause of the intervention or not. It is, of course, true that *if* the intervention is free from any latent common causes, then discovery procedures are simpler and possibly more powerful, since certain sources of correlation between variables can be ruled out. But the point is that the model can easily be adapted to take into account such cases.

As I hinted earlier, the requirement for an *uncaused* intervention is more a matter of fixing the reference frame than one that is a necessary requirement for causal discovery; exogeneity is what matters for interventions.

5.3 Interventions and Free Will

The preceeding discussion of exogeneity should indicate that I am doubtful whether notions such as agency or free will are useful to understanding interventions or causation in general. I do not even think that a freely willed action is a paradigmatic case of an intervention. That is not to say, that it is not an intervention or cannot be one, but rather that I am less clear on what a freely willed action or the notion of agency might be than I am on the notion of intervention. I take questions of intervention and causality to be more fundamental and primary to an understanding of free will and agency, not vice versa.

An intervention does not have to be performed by an agent with free will or the ability to be able to make decisions and it does not have to be the start of a causal chain (in the grand scheme of things). Instead, I take an intervention to be a particular type of cause that satisfies the exogeneity assumption relative to a system of variables whose causal structure we want to investigate. In that sense, I can have a machine that is completely deterministically pre-programmed to perform an intervention on a set of variables, as long as the variables subject to the intervention have no causal influence on the machine. Similarly for a human, in order to perform an intervention we do not have to wait for an answer in the debate on free will – or at least not for an answer of the type that assures us that our free will constitutes the start of a causal chain. Such an answer would be sufficient, but entirely unnecessary, since what matters is whether or not our decision to perform an intervention is *exogenous to the set of variables we are intervening on*, which is a weaker notion than to say it is *undetermined*, let alone *uncaused*. I therefore consider the whole question of whether or not humans or rats or anyone else can perform interventions uninformative with respect to any question of free will. If anything, it is indicative of a representation – conscious or not – of causal separation and independence of the intervention from the intervened set of variables. If we really wanted to reduce the notion of free will down to something where the ability to perform interventions is relevant, then free will would have to have something to do with exogeneity, as I defined it above. But then free will would have to be assigned to a variety of other processes that we would generally not assign free will to just because other

systems can perform interventions.

I do take the ability to perform interventions as a necessary – though certainly not sufficient – condition for free will. But this is a completely uninteresting statement: For some entity not to be able to perform any intervention in any setting implies that this entity is a causal effect of every set of variables that it might possibly intervene on. Such an entity would have to be at the bottom of every causal chain, it would be completely passive (at least in the acyclic causal systems I consider here), and surely we would not think that there would be a case for it to be classified as having free will. I have tried here to give a definition of an intervention and free will is not required for such a definition. Free will and interventions are separate issues.

References

[1] F. Bacon. *Novum Organum*. Parry & MacMillan, 1620, 1854.

[2] J. Campbell. An interventionist approach to causation in psychology. In A. Gopnik and L. Schulz, editors, *Causal Learning: Psychology, Philosophy and Computation*. Oxford University Press, 2006.

[3] F. Eberhardt. Error rates for strategies using sequences of experiments to discover the causal structure. *North Eastern Student Colloquium on Artificial Intelligence (NESCAI)*, 2006.

[4] F. Eberhardt. Sufficient condition for pooling data from different distributions. *Symposium on Philosophy, History, and Methodology of ERROR*, 2006.

[5] F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In F. Bacchus and T. Jaakkola, editors, *Proceedings of the 21 st Conference on Uncertainty and Artificial Intelligence*, pages 178–184. AUAI Press, Corvallis, Oregon, 2005.

[6] F. Eberhardt, C. Glymour, and R. Scheines. N-1 experiments suffice to determine the causal relations among n variables. In D. E. Holmes and L. C. Jain, editors, *Innovations in Machine Learning*, volume 194 of *Theory and Applications Series: Studies in Fuzziness and Soft Computing*. Springer-Verlag, 2006.

[7] F. Eberhardt and R. Scheines. Interventions and causal inference. *20th biennial meeting of the Philosophy of Science Association*, 2006.

[8] R. Fisher. *The design of experiments*. Hafner, 1935.

[9] C. Hitchcock. On the importance of causal taxonomy. In A. Gopnik and L. Schulz, editors, *Causal Learning: Psychology, Philosophy and Computation*. Oxford University Press, 2005.

[10] K. B. Korb, L. R. Hope, A. E. Nicholson, and K. Axnick. Varieties of causal intervention. In C. Zhang, H. W. Guesgen, and W. K. Yeap, editors, *Proceedings of the 8 th Pacific Rim International Conferences on Artificial Intelligence*. Springer, 2004.

[11] J. S. Mill. *Philosophy of scientific method*. Hafner, 1843, 1950.

[12] J. Pearl. *Causality*. Oxford University Press, 2000.

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

[14] D. Rubin. Assignment to treatment group on the basis of covariate. *Journal of Educational Statistics*, 2:1–26, 1977.

[15] D. Rubin. Bayesian inference for causal effects: The role of randomizations. *Annals of Statistics*, 6:34–58, 1978.

[16] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2 edition, 2000.

[17] J. Woodward. *Making Things Happen*. Oxford University Press, 2003.

[18] J. Woodward. Interventionist theories of causation in psychological perspective. 2005.