Causality:
a decision theoretic foundation*

Pablo Schenone†
Arizona State University

This version: September 6, 2019.
First version: September 1, 2017

Abstract

We propose a decision theoretic model akin to that of Savage [19] that is useful for defining causal effects. Within this framework, we define what it means for a decision maker (DM) to act as if the relation between two variables is causal. Next, we provide axioms on preferences and show that these axioms are equivalent to the existence of a (unique) directed acyclic graph (DAG) that represents the DM's preferences. The notion of representation has two components: the graph factorizes the conditional independence properties of the DM's subjective beliefs and arrows point from cause to effect. Finally, we explore the connection between our representation and models used in the statistical causality literature (for example, Pearl [16]).

KEYWORDS: causality, decision theory, subjective expected utility, axioms, representation theorem, intervention preferences, Bayesian graphs
JEL CLASSIFICATION: D80, D81

*I wish to thank David Ahn, Arjada Bardhi, Jeff Ely, Simone Galperti, Bart Lipman, and Marciano Siniscalchi for insightful discussions on the paper.
†Department of Economics, W.P. Carey School of Business, Arizona State University, Tempe, AZ. E-mail: pablo.schenone@asu.edu. All remaining errors are, of course, my own.

1 Introduction

Consider a statistician (say, Alex) who investigates the relation between intellectual ability, education level, and lifetime earnings of a particular citizen (say, Mr. Kane). As a good statistician, Alex is able to choose between the following options. A safe bet that pays \$0 for sure or the risky bet defined below.

- If Mr. Kane has a college degree and earns more than \$ 100K a year, Alex gets \$1
- If Mr. Kane has a college degree and earns less than \$ 100K a year, Alex gets -\$1
- If Mr. Kane does not have a college degree, Alex gets \$0.

For concreteness, suppose Alex chooses the risky option. Her behavior reveals that, *conditional on obtaining a college degree*, Alex believes that it is more likely that Mr. Kane earns more than \$100K a year than it is that he earns less than \$100K a year. Now, assume Alex is presented with the same choice but "college degree" is replaced with "high school degree"; moreover, assume that Alex now prefers receiving \$0 for sure. Her behavior reveals that, *conditional on obtaining a high school degree*, Alex believes that it is more likely that Mr. Kane earns less than \$100K a year than it is that he earns more than \$100K a year. Alex's behavior reveals that she believes Mr. Kane's education level and lifetime earnings are qualitatively positively *correlated*: she accepts a \$1 gamble that Mr. Kane is making more than \$100K a year conditional on *observing* that Mr. Kane obtained a college degree but not conditional on *observing* that Mr. Kane obtained only a high school degree. Finally, if Alex is probabilistically sophisticated, then we can represent her beliefs with a joint probability distribution over all relevant variables. In particular, this probability distribution is such that education and lifetime earnings are positively correlated.

Alex is now approached by a benevolent politician who wants to improve his constituents' lifetime earnings. Since Alex believes that education and earnings are positively correlated, this politician expects that a policy that forces everyone to obtain a college degree would be useful to improve lifetime earnings. However,

Alex rejects that conclusion. While she believes Mr. Kane's education level and lifetime earnings are positively correlated, she is of the opinion that policies that change the population's education levels while keeping all other things equal are useless for affecting lifetime earnings. Alex believes that high education levels are associated with high intellectual ability, that high intellectual ability is associated with higher lifetime earnings, and that this is the only channel through which education levels and lifetime earnings are related. Thus, a policy that improves education levels but leaves intellectual ability unchanged is useless to improve lifetime earnings.

The apparent tension between Alex's belief that education and earnings are positively correlated, while maintaining a position that policies that affect only education are useless to affect lifetime earnings, is rationalized by the adage "correlation is not causation". In this context, *causation* has a specific meaning: a variable *subjectively causes* another variable if, holding all other variables constant, policy interventions on the first variable affect Alex's beliefs about the second. That Alex believes education policies are useless to affect lifetime earnings (holding fixed intellectual ability) means that she believes education levels do not cause lifetime earnings.

The above definition of causal effect is entirely subjective. As such, this definition is not about objective truths or uncovering the laws of nature. However, this definition captures exactly how causality is understood in economics. In economics, causal relations are correlations that, *in the analyst's subjective opinion*, are valid grounds for making policy recommendations. While disagreements exist with regards to *how* one arrives at the conclusion that an observed correlation is sufficient grounds for making policy recommendations, the definition of causation as the bridge between correlation and policy recommendation is undisputed. This dichotomy — when are two variables correlated versus when is one variable a useful policy tool to affect the other — is the foundation of our definition of causal effect. By identifying a unique numerical representation of this definition, our paper provides a foundation for selecting models with which empirical researchers can estimate causal effects.

The purpose of axiomatic exercises like Savage's [19] is to provide a link between some numerical model and the way a rational decision maker (henceforth, DM) approaches the issue of interest (in this case, causality). The goal is to guarantee that the numerical model treats the object of study the way a rational DM would. For empirical research, the role of the DM is played by the researcher's econometric model (which, presumably, wants to behave rationally), and the role of the DM's beliefs is played by the probability laws the researcher feeds into the numerical model. The subjectivity in the definition of causation reflects that researchers need to make assumptions about the causal structure of the world, and these assumptions carry on the the researcher's econometric model (i.e. the DM in our paper). This paper provides a theoretical foundation for selecting amongst models of causality by proposing normative axioms for how the analyst's model should treat uncertainty.

This paper is structured in three steps. First, we propose a decision problem similar to Savage's: there is a set of states, a set of acts mapping states into monetary amounts, and a DM who chooses among acts. The DM makes choices as if picking the best alternative according to a preference relation. This language is sufficient to talk about the subjective correlation structure in the DM's beliefs. However, to discuss causal effects, we also need language to talk about preferences over intervention policies that affect the states. Therefore, we extend the language in the Savage model to accommodate for the possibility of choosing policies that affect the states. Section 3 describes the model, and Section 4 formally defines causality. Second, we propose a set of axioms that capture -in a normative sense- how a rational DM should treat uncertainty and causality. Section 5 presents the axioms. Finally, conduct a standard decision theoretic analysis: we propose a numerical representation of the DM's beliefs (see section 6) and show that our axioms hold if, and only if, we can numerically represent the DM's beliefs. Section 7 presents our main theorems.

As the reader may anticipate, the statistics, computer science, and economics literature addressing causal effects is extensive. The related literature is discussed in Section 8, and we delay a discussion of it until after we present our results because our results depend on a series of definitions and terms related to various

literatures. Hence, we do not yet have the language to meaningfully discuss the related work.

2 How Alex fits in the grand scheme.

As a preamble to the formal model, this section uses a simple example to illustrate how our representation contributes to the modeling of causal effects. As such, it may be skipped without loss of continuity. The first two observations relate to graphical methods in general; the last two pertain to the specifics of our representation.

Let us expand the example in the introduction to include an "occupation" variable. The model thus contains four variables: Ability (A), Education level (E), Lifetime earnings (L) and occupation (O). Alex the statistician wants to estimate the causal effect of education on earnings.

All models of causal analysis will result in the same type of conclusion: the causal effect of E on L is obtained by looking at the joint distribution of E and L after suitably conditioning on (or *controlling for*) possible confounders. The essence of the exercise is to decide exactly which variables must be conditioned on. Of course, the answer to this question is a function of the assumptions Alex makes about the causal structure of the world. In carrying out this exercise Alex wants to transmit what assumptions she makes in the simplest, most transparent way possible. She also wants to decide what variables need to be conditioned on using the most tractable method she can.

Graphical methods allow Alex to lay out her assumptions in a succinct and crisp manner. To do this, Alex draws a graph where arrows point from *cause* to effect. For concreteness, say Alex draws the graph in Figure 1. Amongst other things, this graph claims that ability is a joint cause of occupation and earnings, but is neither cause nor consequence of education level; this assumption might be controversial but it is clearly and transparently stated. As a by-product, Alex is also transmitting all her assumptions about conditional independence. Indeed, there is a one-to-one correspondence between arrows in a graph and statements about conditional independents (see Section6 for details). In particular, Alex is assuming that E and A are the only independent variables, and all other variables are

statistically dependent on each other. Again, this could be controversial, but it is clearly and succinctly stated. Finally, by omitting any other variable, Alex is explicitly making the assumption that only these variables matter in her analysis. With a single picture Alex transmits *all* the assumptions she is making: how are the variables causally related to each other, which variables are statistically independent of each other, what variables matter for her analysis, and (by exclusion from the graph) what variables do not matter for her analysis. While Alex could have written down an equivalent set of potential outcome equations, together with a set of weak and strong ignorability assumptions (see Rosenmbaum-Rubin [18], for example), she manages to convey the exact same information simply by drawing Figure 1. This economy of language afforded by graphical methods becomes increasingly important as the number of variables grows.

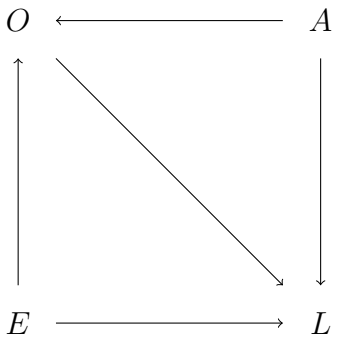


Figure 1: A crisp and succinct description of all causal assumptions and all assumptions on conditional independence; $\mu(E, O, A, L) = \mu(E)\mu(A)\mu(O|A, L)\mu(L|A, E, O)$.

Graphical methods also allow Alex to quantify causal effects in a tractable manner. In this case the causal effect of E on L is obtained by looking at how the expression $\mu(L|E)$ moves with E. Despite the possible confounders, no additional conditioning is required. Pearl [16] provides a simple way to check this by looking at the paths that connect E to L. Checking which variables to condition on amounts to checking two simple properties of the paths that connect the variables of interest. In this case, conditioning on a variable with head-to-head arrows (such as O) introduces spurious correlation, thus O should not be controlled for (again, see the original Pearl paper or section 7.1 for the technical details). Importantly, simple algorithms exist that take a causal graph as an input and immediately out-

put which variables must be controlled for (DAGitty, for instance, is a web-browser based algorithm for doing this). While the same expression for the causal effect of E on L is obtained from a potential outcomes model, understanding *all* implications of the model –both causal and in terms of correlations– becomes quickly intractable. Graphical methods offer a naturally tractable way to obtain the result. As before, this tractability is increasingly important as the number of variables grow large, as it tends to do in economics.

These two points illustrate that graphical methods are not a substitute to classical methods (like instrumental variables or potential outcomes) but a powerful complement. For example, [10] shows how causal instrumental variable analysis can be understood within the language of Bayesian graphs. Graphical methods allow us to encode similar assumptions, and obtain similar results, but they do so via a succinct, transparent, and tractable language –these things become increasingly important as the number of variables increases.

However, our representation offers two further insights into the analysis of causation that are absent from the traditional work on graphical methods. When Alex constructs the graph with which to analyze the problem, her graph must serve two roles at once. First, her graph must represent her causal model: arrows point from *cause* (whichever which way Alex defines this word) to effect. Second, her graph represents the assumptions Alex makes about correlation structures. Demanding that a graph satisfies these two conditions at once amounts to a joint assumption: an assumption about how Alex defines causality and an assumption about how causality and correlations interact. What are the exact conditions on both the definition of causality and the way causal effects interact with correlations that permit representation via a single graph? The Bayesian Graphs literature is silent about this question, which we answer here. Given our formal definition of causal effect (Definition 2) Theorem 1 states the exact set of conditions that are necessary and sufficient for the existence of a graph that simultaneously represents a causal model and a correlation structure.

A further insight of our model is to highlight a lax use of the word *causality*. In the context of figure 1, we claimed that the causal effect of E on L is captured by

$\mu(E|L)$. Our model highlights two subtleties about this statement. First, there are three types of "causal effect" encoded in figure 1. The direct effect encoded by the arrow $E \to L$, the indirect effect encoded by the path $E \to O \to L$, and the total effect derived from the existence of these two paths simultaneously. Our definition of causality makes explicit the difference between these three effects. In particular, it makes explicit that the expression $\mu(E|L)$ only captures the total causal effect of E on L, while the direct effect $E \to L$ is captured via a separate formula (see section 7.1 for details). Second, the tools used to derive this formula in Pearl [16] depend on a formalism called a do-probability (see section 7.1 for details). Theorem 2 shows that this formalism is valid only when we add an extra axiom relative to those in Theorem 1. Therefore, the notion of causal effect quantified by Pearl [16] is strictly stronger than what is encoded by the arrows in a graph. This is a point that is not made explicit in the literature on Bayesian graphs but that is made explicit on our paper.

3 Model and notation

3.1 General notation

The following useful notation is used throughout this paper. The set $\mathcal{N} = \{1, ..., N\}$ is a set of indexes. For each $\mathcal{J} \subset \mathcal{N}$, let $\{X_j : j \in \mathcal{J}\}$ be a family of sets indexed by $\mathcal{J}$. We denote by $X_{\mathcal{J}} = \Pi_{j \in \mathcal{J}} X_j$ the Cartesian products of the family and by $x_{\mathcal{J}} = (x_j)_{j \in \mathcal{J}}$ a canonical element in $X_{\mathcal{J}}$. Moreover, all complements are taken with respect to $\mathcal{N}$: if $\mathcal{J} \subset \mathcal{N}$, then $\mathcal{J}^{\complement} \equiv \mathcal{N} \backslash \mathcal{J}$. Finally, if $\mathcal{J} \subset \mathcal{N}$ and $E \subset X_{\mathcal{J}}$, then $\mathbb{1}_E : X_{\mathcal{J}} \to \{0, 1\}$ denotes the indicator function that event E has occurred; that is, $\mathbb{1}_E(x_{\mathcal{J}}) = 1 \Leftrightarrow x_{\mathcal{J}} \in E$.

The following notation refers to the graph theoretic component of the model. A directed graph is a pair (V, E) such that V is a (finite) set of nodes and $E \subset V \times V$ is the set of edges. If two nodes, i and j, satisfy that $(i, j) \in E$, we simplify the notation by writing $i \to j$. Moreover, the set of *parents* for a node $v \in V$ is the set $Pa(v) = \{v' \in V : (v', v) \in E\}$. A node $v \in V$ is a descendant of a node $v' \in V$ whenever a directed path exists from v' to v. Formally, if a sequence $(v_1, ..., v_T) \in V^T$ exists such that $v_1 = v'$, v_t is a parent of v_{t+1} for each $t \in \{1, ..., T-1\}$, and $v_T = v$. Likewise, v' is an ancestor of v whenever v is a descendant of v'. A directed

graph is a DAG if, and only if, for all $v \in V$, v is not a descendant of v. We denote by $D(v)$ the set of descendants of v and by $ND(v)$ the set of non-descendants.

3.2 Model description

Our DM faces a variant of the standard Savage problem. The state space is $S = \Pi_{i=1}^{N} X_i$, where each X_i is finite. We make this assumption for technical simplicity because causality is orthogonal to whether state spaces are finite or infinite. We let $\mathcal{N} = \{1, ..., N\}$, and we call each $i \in \mathcal{N}$ a *variable*. Set $\mathcal{A} = \mathbb{R}^S$ is the set of Savage acts, and a DM has preferences $\succ$ over $\mathcal{A}$.

However, our problem differs from Savage's since we incorporate policies that affect the states. This added language allows us to distinguish correlations from other types of relations among variables. A set of *intervention policies* is a set $\mathcal{P} = \Pi_{i=1}^{N}(X_i \cup \{\varnothing\})$. The interpretation is as follows. Let a policy $p \in \mathcal{P}$ be such that $p_i = \varnothing$ for some $i \in \mathcal{N}$. Then, this policy leaves variable i unaffected; that is, i is determined as it would have been in a standard Savage world. However, if for some $j \in \mathcal{N}$, we have $p_j = x_j \in X_j$, then policy j forces variable j to take the value x_j; that is, the value of variable j is not determined as it would have been in a Savage problem but is chosen by the DM. Therefore, each policy implies a collection of interventions of the state space. Our model is one where the DM first chooses a policy from the set of all policies, and then chooses a Savage act from the set of acts defined over the non-intervened variables.

We now define the primitive choice domain for our DM. Let $p \in \mathcal{P}$ be any policy, and let $\mathcal{N}(p) = \{i \in \mathcal{N} : p_i = \varnothing\}$. That is, $\mathcal{N}(p)$ are the variables that p leaves unaffected. Furthermore, let $\mathcal{A}(p) \equiv \mathbb{R}^{X_{\mathcal{N}(p)}}$ be the set of acts defined over the variables that p leaves unaffected. Then, the primitive domain of choice for the DM is the set $\{(p, a) : p \in \mathcal{P}, a \in \mathcal{A}(p)\}$. That is, our DM's problem is to select an intervention policy and a Savage act over the non-intervened variables. We endow this DM with a preference relation $\bar{\succ}$ on $\{(p, a) : p \in \mathcal{P}, a \in \mathcal{A}(p)\}$.

Given $\bar{\succ}$, each p induces an *intervention preference* on $\mathcal{A}(p)$: for each $p \in \mathcal{P}$ and each $f, g \in \mathcal{A}(p)$, we say $f \succ_p g$ if, and only if, $(p, f) \bar{\succ} (p, g)$. Since our axioms are focused on the DM's intervention preferences, it is convenient to express intervention preferences explicitly in terms of the values at which the variables are

intervened. For each policy $p \in \mathcal{P}$, if $p_{\mathcal{N}(p)^\complement} = x_{\mathcal{N}(p)^\complement}$, we use $\succ_{x_{\mathcal{N}(p)^\complement}}$ to denote $\succ_{p_{\mathcal{N}(p)^\complement}}$. The special case where $p = (\varnothing, ..., \varnothing)$, so that no variables are intervened, corresponds to the DM's preferences in a standard Savage world. For such a p, we use $\succ_{(\varnothing,...,\varnothing)} = \succ$ for notational simplicity.

From intervention preferences we obtain *intervention beliefs*. For each $p \in \mathcal{P}$, we say that $\succ_p$ has a *belief representation* if there is a probability distribution μ_p on $X^{\mathcal{N}(p)}$ such that for all $E, F \subset X^{\mathcal{N}(p)}$, $\mu_p(E) > \mu_p(F)$ if, and only if, $\mathbb{1}_E \succ_p \mathbb{1}_F$. When such a representation exists we say μ_p is an *intervention belief*.

Intervention preferences (resp. beliefs) look like Savage conditional preferences (resp. beliefs) but have important differences. Savage conditional preferences capture betting behavior conditional on the DM observing that a certain event was realized, whereas intervention preferences (beliefs) capture betting behavior after a controlled intervention of the relevant variables. To illustrate the difference, consider Example 1 below. Conditional preferences (resp. beliefs) are statements about item [1.], whereas intervention preferences (resp. beliefs) are statements about item [2.]. These are clearly different statements that do not imply one another. Therefore, we need language to distinguish these two distinct decision problems and intervention policies provide such language.

Example 1. *Let acts f and g over lifetime earnings be defined as follows. Act f pays* \$1 *if lifetime earnings are greater than* \$100K *per year and* $-$\$1 *otherwise. Act g is the opposite: it pays* $-$\$1 *if lifetime earnings are greater than* \$100K *per year and* \$1 *otherwise. Consider the following statements:*

1. *"Having observed that Mr. Kane earned a college degree (of his own free will and ability), Alex prefers f to g."*
2. *"Having forced Mr. Kane to obtain a college degree (regardless of his desire or ability to do so) Alex prefers f to g".*

Because this paper is concerned with understanding what a rational agent's approach to causality is, the role of axioms is exclusively normative. Whether actual humans adhere to these axioms is orthogonal to this paper. While the counter-factual based setup presented above may seem hard to test in a laboratory

with actual human subjects, this is not the objective of the exercise at hand. Since the DM in our paper is an analyst's econometric model of the world, the only question that matters is whether the analyst finds the axioms normatively appealing or not. Moreover, econometric models (as opposed to human subjects in a laboratory) are naturally built to handle counter-factual analysis of the sort presented above.

4 Definition of causal effect

In this section we introduce the definition of causal effect, which formalizes the intuitive definition given in Section 1.

We begin by introducing the definition of intervention independence. Consider a set of variables $\mathcal{K}$ and two variables $i, j \notin \mathcal{K}$. Informally, i is $\mathcal{K}$-independent of j if, after eliminating the possibility that i and j are related through variables in $\mathcal{K}$, the choice of acts over i is insensitive to interventions of j. Formally, we say that i is $\mathcal{K}$-independent of j if the following holds: $(\forall x_{\mathcal{K}} \in X_{\mathcal{K}})$, $(\forall x_j, x'_j \in X_j)$, and $(\forall f, g \in \mathbb{R}^{X_i})$,

$$\begin{aligned} f \succ_{x_j, x_{\mathcal{K}}} g \quad &\Leftrightarrow \quad f \succ_{x'_j, x_{\mathcal{K}}} g, \\ f \succ_{x_j, x_{\mathcal{K}}} g \quad &\Leftrightarrow \quad f \succ_{x_{\mathcal{K}}} g.^{1} \end{aligned}$$

The first line indicates that having intervened $\mathcal{K}$ at value $x_{\mathcal{K}}$, intervening j at different values does not affect the DM's choice of act in $\mathbb{R}^{X_i}$. The second line indicates that having intervened $\mathcal{K}$, the ability to intervene j at all, regardless of the values at which it is intervened, does not affect the DM's choice of act in $\mathbb{R}^{X_i}$. Note that the second of these conditions implies the first. Indeed, if the second condition holds, then we have that for all x_j, x'_j,

$$f \succ_{x_j, x_{\mathcal{K}}} g \Leftrightarrow f \succ_{x_{\mathcal{K}}} g \Leftrightarrow f \succ_{x'_j, x_{\mathcal{K}}} g,$$

so the first equation also holds. This motivates the formal definition of intervention independence.

Definition 1. *For all $i, j \in \mathcal{N}$ and $\mathcal{K}$ such that $i, j \notin \mathcal{K}$, we say variable i is*

$\mathcal{K}$-independent *of variable* j *if for all* $f, g \in \mathbb{R}^{X_i}$,

$$f \succ_{x_j, x_{\mathcal{K}}} g \quad \Leftrightarrow \quad f \succ_{x_{\mathcal{K}}} g,$$

To illustrate Definition 1, consider a DM who believes *A*bility has a direct impact on *E*ducation and that Education has a direct impact on *L*ifetime earnings but that Ability has no direct impact on Lifetime earnings. This is depicted in Figure 2 below. If $a, a' \in A$ are two ability levels and f, $g \in \mathbb{R}^L$ are two acts on lifetime earnings, we might have the DM behave as follows: $f \succ_a g$ and $g \succ_{a'} f$. This reversal indicates that A and E are not $\{\varnothing\}$-independent, which is intuitive: interventions of A affect beliefs about E, and beliefs about E affect beliefs about L. However, this is an effect of A on L that is mediated through E. As such, we don't want to use this as basis to claim that A *causes* L. The correct way to capture the causal effect of A on L is to look at intervention preferences $\succ_{(a,e)}$ as a function of a, for each fixed $e \in E$. In other words, we want to ask if A and L are $\{E\}$-independent. This motivates the formal definition of causal effect.

$$A \longrightarrow E \longrightarrow L$$

Figure 2: Variable A has no direct causal effect on L, but non-ceteris paribus interventions of A affect L through E.

Definition 2. *For all* $i, j \in \mathcal{N}$, *we say variable* j causes *variable* i *if* i *is not* $\{i, j\}^{\complement}$*-independent of* j.
Let $Ca(i) = \{j \in \mathcal{N} : j \text{ causes } i\}$ *denote the* causal set of i.
Finally, we say j *is an indirect cause of* i *if there is a sequence* $j_0, ..., j_T$ *such that, for all* $t \in \{0, ..., T-1\}$, j_t *causes* j_{t+1}, $j_0 = j$ *and* $j_T = i$.

Finally, if a variable i is such that $Ca(i) = \varnothing$, we say i is an *exogenous primitive*; otherwise, we say it is an endogenous variable. Indeed, when a DM forms a causal model of the world, the set of primitives of such model is precisely the set of variables that are not caused by any other variable in the model. Exogenous primitives are relevant in our discussion of Axiom 1.

We conclude this section by defining the causal graph associated with a pref-

erence, $\bar{\succ}$. Causal graphs are an integral part of our representation, which is introduced in Section 6. Given $\bar{\succ}$, draw a graph by letting the set of nodes be the set of variables and the set of arrows be defined by the causal sets, that is, by letting $j \to i \Leftrightarrow j \in Ca(i)$. This graph is well defined because $Ca(i)$ is well defined for each $i \in \mathcal{N}$. We denote such a graph as $G(\bar{\succ})$.

Definition 3. *Let $\bar{\succ}$ be a preference and $\{Ca(i) : i \in \mathcal{N}\}$ be the collection of causal sets derived from $\bar{\succ}$. Define $G(\bar{\succ}) = (V, E)$ by setting $V = \mathcal{N}$ and $E = \{(j, i) : j \in Ca(i)\}$.*

5 Axioms

Our axioms are normative statements about how the DM should treat uncertainty as a function of the DM's causal model. Hence, our axioms tackle variations of the following question: given the DM's causal graph as per Definition 3, what *normative* restrictions should we impose on the DM's intervention beliefs?

As such, the axioms are about conditional independence properties of the various $\succ_p$ preferences. Since the act notation for conditional independence is somewhat heavy, we use the following simplifying notation.

Definition 4. *Let $i \in \mathcal{N}$ and let $\mathcal{J}, \mathcal{K}, \mathcal{H} \subset \mathcal{N}$ be disjoint sets such that $i \notin \mathcal{J} \cup \mathcal{K} \cup \mathcal{H}$. We say that i is independent of $\mathcal{J}$ conditional on $\mathcal{K}$ after intervening $\mathcal{H}$ if the following is true for all $x_{(\mathcal{J}\cup\mathcal{K}\cup\mathcal{H})} \in X^{(\mathcal{J}\cup\mathcal{K}\cup\mathcal{H})}$ and all $f, g \in \mathbb{R}^{X_i}$:*

$$\mathbb{1}_{x_\mathcal{K}} f \succ_{x_\mathcal{H}} \mathbb{1}_{x_\mathcal{K}} g \Leftrightarrow \mathbb{1}_{x_\mathcal{K}} \mathbb{1}_{x_\mathcal{J}} f \succ_{x_\mathcal{H}} \mathbb{1}_{x_\mathcal{K}} \mathbb{1}_{x_\mathcal{J}} g. \tag{1}$$

When the above holds, we write

$$i \perp_\mathcal{H} \mathcal{J} | \mathcal{K}. \tag{2}$$

In the case $\mathcal{J}$ is a singleton, $\mathcal{J} = \{j\}$, we simply write

$$i \perp_\mathcal{H} j | \mathcal{K}. \tag{3}$$

In terms of behavior, conditional independence says the following: i and j are independent if a DM would never pay for information about j when their task is to predict i. Imagine a DM intervened variables $\mathcal{H}$ to a specific level. For instance, the DM carried out a controlled experiment, or this could simply be a thought experiment. Imagine further that the DM observed a specific realization of variables in $\mathcal{K}$. In this context, if the DM had to choose choice between f and g he would have to compare $\mathbb{1}_{x_{\mathcal{K}}} f$ with $\mathbb{1}_{x_{\mathcal{K}}} g$ using preferences $\succ_{x_{\mathcal{H}}}$. To aid the DM's decision, someone offers to reveal the DM the value of the variables in $\mathcal{J}$ at a fee of $\varepsilon > 0$. Is there an ε small enough that the DM would purchase this information revelation? If the DM bought this information about $\mathcal{J}$ then his problem becomes to compare $\mathbb{1}_{x_{\mathcal{K}}} \mathbb{1}_{x_{\mathcal{J}}} f$ with $\mathbb{1}_{x_{\mathcal{K}}} \mathbb{1}_{x_{\mathcal{J}}} g$ using preferences $\succ_{x_{\mathcal{H}}}$. Since the axiom says that his choice under both situations is the same, the information is useless. Thus, the DM would not accept any price $\varepsilon > 0$.

We start with Assumption 1, which defines all relevant aspects of the DM's probabilistic model. To state Assumption 1 we recall the definition of a subjective expected utility preference. We say that a preference $\succ_p$ is a (monotone) subjective expected utility preference if there exists a unique probability distribution, $\mu_p \in \Delta(X^{\mathcal{N}(p)})$, and a (monotone increasing) function $u_p : \mathbb{R} \to \mathbb{R}$ such that for all acts $f, g \in \mathbb{R}^{X^{\mathcal{N}(p)}}$ condition 4 below holds. There are many axiomatizations of monotone expected utility preference that fit the framework of our model, such as Gul [5] , Fishburn [2], and Theorem 3 in Karni [12], amongst others. We let the reader pick their favorite axiomatization.

$$f \succ_p g \Leftrightarrow \sum_{x_{\mathcal{N}(p)} \in X_{\mathcal{N}}(p)} u_p(f(x_{\mathcal{N}(p)})) \mu_p(x_{\mathcal{N}(p)}) > \sum_{x_{\mathcal{N}(p)} \in X_{\mathcal{N}}(p)} u_p(g(x_{\mathcal{N}(p)})) \mu_p(x_{\mathcal{N}(p)}). \quad (4)$$

Assumption 1. *For each $\mathcal{J} \subset \mathcal{N}$, the following are true.*

i- For each $p \in \mathcal{P}$, the preferences $\succ_p$ are monotone subjective expected utility preferences.

ii- The states space is complete: $(\forall i, j \in \mathcal{N})$, $(\forall x_{\mathcal{N}\setminus\{i\}} \in X_{\mathcal{N}\setminus\{i\}})$, and $(\forall f, g \in \mathbb{R}^{X_i})$, if $j \in Ca(i)$, then $f \succ_{x_{\mathcal{N}\setminus\{i\}}} g \Leftrightarrow \mathbb{1}_{x_j} f \succ_{x_{\mathcal{N}\setminus\{i,j\}}} \mathbb{1}_{x_j} g$.

iii- *There are no null states: for all* $x \in X$, $\mathbb{1}_x \succ \mathbb{1}_X 0$.

iv- *Policies do not affect preferences:* $(\forall x, y \in \mathbb{R})$, $(\forall p, p' \in \mathcal{P})$, $\mathbb{1}_{X_{\mathcal{N}(p)}} x \succ_p \mathbb{1}_{X_{\mathcal{N}(p)}} y \Leftrightarrow \mathbb{1}_{X_{\mathcal{N}(p')}} x \succ_{p'} \mathbb{1}_{X_{\mathcal{N}(p')}} y$

Assumption 1 captures the basics of the DM's beliefs. As such it is orthogonal to issues of causation, hence the reason we don't refer to it as an *axiom* per se. Below, we examine each of the restrictions Assumption 1.

As we already mentioned, many axiomatizations exist that will deliver item $[i.]$, each with it's own advantages and disadvantages. We let the reader decide what their favorite axiomatization of monotone expected utility is. The importance of item $[i.]$ is that all intervention preferences are probabilistically sophisticated, so intervention beliefs are always well defined. Item $[iii.]$ states that getting paid \$1 if realization $x \in X$ occurs is strictly preferred to getting \$0 for sure, thus guaranteeing that all states receive positive probability. Item $[iv.]$ rules out the possibility that policies have a direct impact on the Bernoulli utility indexes, thus making $u_p = u'_p$ for all $p, p' \in \mathcal{P}$.[2]

Item $[ii.]$ in Assumption 1 says that the state space is complete: given two variables (say, i and j), the state space includes all variables that could mediate effects between i and j. Once all variables $k \neq i, j$ are intervened, either i causes j, j causes i, or the i and j are independent. If j causes i, since all possible confounding variables have been intervened, observing that $x_j \in X_J$ was realized or intervening variable j to value x_j should lead to the same preference over $\mathbb{R}^{X_i}$. Violations of this axiom are reasonable only if the state space is missing some potential confounding variables. In line with Savage [19], we assume that the state space is complete.

That the state space is complete has the following implication for the representation of preferences. Pick two variables (say, i an j) and assume all other variables are intervened at a level $x_{\{i,j\}^\complement}$. Let $\mu_{x_{\{i,j\}^\complement}}$ be the belief representation of $\succ_{x_{\{i,j\}^\complement}}$. In this 2-variable environment, correlation and causation should coincide.

[2]This assumption is not strictly needed but it simplifies notation in some proofs. Since causality is orthogonal to whether Bernoulli utilities are constant in $\mathcal{P}$, we feel comfortable keeping this assumption.

Therefore, if i causes j, conditioning on j or intervening j should lead to the same posteriors about i. Namely,

$$\mu_{x_{\{i,j\}^\complement}}(x_i|x_j) = \mu_{x_{\{i,j\}^\complement},x_j}(x_i). \tag{5}$$

Importantly, this relation is not symmetric. If j does not cause i then the symmetric expression $\mu_{x_{\{i,j\}^\complement}}(x_j|x_i) = \mu_{x_{\{i,j\}^\complement},x_i}(x_j)$ is false. The left hand side is a non-constant function of x_i whereas the right hand side is constant in x_i. Thus, under complete state spaces, equation 5 identifies the direction of causality. We will use this observation in Section 6 when defining when a graph represents a preference .

That the state space is complete does *not* imply the DM must know what all the relevant variables are. For instance, assume Alex in the introduction is worried that the interaction between Ability, Education and Lifetime earnings might be affected by some other variable. Concretely, she thinks some other variable might influence education levels: she does not know what this variable is, but she believes it exists. For concreteness, denote this variable as "Unknown but possibly exiting variable". Assumption 1 says that in her state space she should include such a variable. Therefore her state space should not be $A \times E \times L$ but rather $A \times E \times L \times U$, where U stands for "Unknown but possibly existing variable". In short, Assumption 1 does allow the econometrician to add variables that act as proxies for unknown shocks to the system. Indeed, modeling a potential unknown confounder as exogenous noise shocks is a common way to proceed in empirical studies.

Assumption 1 states that the DM is probabilistically sophisticated but is silent about the statistical properties of causal sets. Without further axioms to discipline how the causal sets behave, we cannot guarantee that these sets will have any properties that we normatively associate with causation. Axioms 1 through 4 provide such discipline.

Axiom 1. *For all $i \in \mathcal{N}$, i is not an indirect cause of i.*

Axiom 1 is equivalent to the following statement: for each set of variables $\mathcal{I} \subset \mathcal{N}$, there exists $i \in \mathcal{I}$ such that $Ca(i) \cap \mathcal{I} = \varnothing$. That is, if the DM is asked to ex-

plain the relation between variables in $\mathcal{I}$ and only those in $\mathcal{I}$, the DM has an explanation that involves at least one exogenous primitive relative to $\mathcal{I}$. Models without primitives describe identities rather than relations among logically independent variables. Therefore, Axiom 1 states that the DM's state space includes only logically independent variables.

A potential critique of this axiom is that certain systems are inherently cyclical. For instance, the relation between the speed of a car, the distance traveled by the car, and the time traveled by the car is inherently circular: any two determine the third. The problem with this system is that speed is not *caused* by distance and time traveled; rather, speed is *defined* in terms of distance and time traveled. Therefore, the model includes variables that are not logically independent of one another. The correct model to analyze this situation is one in which the only variables are time and distance traveled by the car, as these variables are the only logically independent variables. In this sense, the assumption that no causal cycles exist is sensible.

A related critique of Axiom 1 is that it precludes the DM from viewing the world as a system of recursive structural equations. As such, Axiom 1 could be seen as precluding the DM from reasoning in terms of equilibrium equations (see, for example, the critique in Heckman and Pinto [9]). This assessment stems from interpreting functional relations as causal relations. However, the equations in a model (in particular, equilibrium equations) are succinct descriptions of the specific values that the variables may obtain; they say nothing of *how* those values are achieved. As such, causality and equilibrium equations are orthogonal issues.

To make the above discussion concrete, consider a general equilibrium model with aggregate demand curve D and aggregate supply curve S. Equilibrium is defined as follows: (p^*, q^*) constitute an equilibrium if $D(p^*) = q^*$ and $S(p^*) = q^*$. Note that this is a definition; as such, *equilibrium* price and *equilibrium* quantity are not logically independent. These equations describe the values one should expect for prices and quantities but are silent regarding the mechanism that generated them. This silence motivates the equilibrium convergence literature. For example, a tatônnement convergence process is compatible with the general equi-

librium equations without invoking feedback loops: a DM posits that prices in period t cause quantities in period t (via consumer/producer optimization) and that quantities in period t cause prices in period $t+1$ (through a process that increases/decreases the price in response to excess demand/supply). That the system stabilizes at a point where $p_t = p_{t+1} = p^*$ and $q_t = q_{t+1} = q^*$ is orthogonal to the issue of causation. In short, one should not mistake functional equations, which simply describe relations between variables, for causal statements.

Axiom 2. $(\forall i \in \mathcal{N})$ $(\forall j \in Ca(i))$, $(\forall \mathcal{H}, \mathcal{K} \subset \mathcal{N} \backslash \{i, j\})$ *that are disjoint,*

$$i \not\perp_{\mathcal{K}} j | \mathcal{H}$$

Axiom 2 captures the following normative property about causation: the causes of a variable (say, i) are the most proximal sources of information about i. Similar to conditional independence, if j is the most proximal source of information about i then i and j should never be independent of each other. That is, if a DM had to predict the value of i, and j is a cause of i, there should be an $\varepsilon > 0$ small enough that the DM would pay ε in exchange for information on the value of j. Thus, axiom 2 captures that causes are the most proximal source of information by stating that causes are never independent from their consequences. As a final remark, notice that Axiom 2 is symmetric in the following sense: the only fundamental sources of information about i are are causes of i, and those variables that are directly caused by i.

Axiom 3. $(\forall i, j \in \mathcal{N})$, $(\forall \mathcal{K} \subset \mathcal{N} \backslash \{i, j\})$, *if* $i \notin Ca(j)$ *and* $j \notin Ca(i)$ *then*

$$i \perp_K j | (Ca(i) \cup Ca(j) \backslash \mathcal{K})$$

While axiom 2 describes the conditional independence properties of variables that are directly related to each other, axiom 3 analyzes the independence properties of variables that are not directly related to each other. Axiom 3 states that two variables that do not cause each other are independent of each other once we condition on their causes.

To understand Axiom 3's normative appeal, consider the DAG in figure 3 below, where arrows point from cause to effect. If a DM had to predict the value of i and he knew the realizations x_b and x_c, should the DM pay for information about the realization of j? Our axiom says the DM shouldn't pay for this information, which is quite sensible: once the DM knows the values of b and c, he knows all there is to know about the relation between i and j. Thus, extra information on j is useless to predict the value of i.

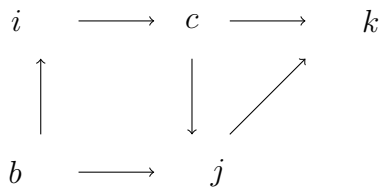


Figure 3: i and j are independent conditional on their respective causes.

A more general analysis of Axiom 3 proceeds in three steps. First, it is normatively appealing to say that i and j are not independent. Because i causes c and c causes j, it stands to reason that any information we have about i will (via c) provide information about j. Likewise, b provides another link between i and j: since b is a common cause of both, then any information we have about i should allow us to make inferences about b and, in turn, inferences about j. Second, because neither i causes j nor vice-versa, any information i provides about j will be mediated by some variable. Third, the mediating variable will either be a cause of i, a cause of j, or both. Indeed, if i provided information about j that is not mediated by any cause of j then i is providing information about j that is more proximal than the information contained by any cause of j. Therefore, i should itself be a cause of j, which it is not. Putting these three observations together implies that, if we condition on both the causes of i and of j then i and j should be conditionally independent.

However, there are cases where the conclusion of axiom 3 is normatively too weak, and a stronger conclusion would be more appealing. To see this, consider figure 4: if we want to understand when is i independent from the causes of j (in this case, c and b) we could simply apply axioms 2 and 3. By axiom 2 i is

never independent from b. By axiom 3, i and c are independent conditional on the cause of i and the causes of c (b and d in this example). This would result in the conclusion that i is independent of c conditional on both b and d. However, this is an overly weak conclusion: conditional on b, d is playing no role in the relationship between i and c.

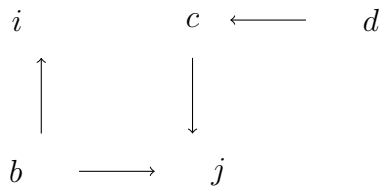


Figure 4: Axiom 3 implies $i \perp c|\{b,d\}$, which normatively is too strong a conclusion. Axiom 4 strengthens this conclusion to $i \perp c|\{b\}$ which is normatively more appealing.

The above discussion motivates axiom 4.

Axiom 4. $(\forall i \in \mathcal{N})$, $(\forall j \in ND(i))$ *and* $(\forall \mathcal{K} \subset \mathcal{N} \backslash \{i,j\})$,

$$i \perp_{\mathcal{J}} (Ca(j) \backslash \mathcal{K}) | (Ca(i) \backslash \mathcal{K})$$

indent Axiom 4 states that if i is not an indirect cause of j, then i is independent of the causes of j once we condition on the causes of i. First, assume i is an indirect cause of j. If i is an indirect cause of j, then i is an indirect cause of the causes of j. Therefore, it is irrational to impose that i be independent of the causes of j when we condition on the causes of i alone. For this reason axiom 4 only restricts behavior when $j \in ND(i)$. Suppose then that i is not an indirect cause of j. Then i is not an indirect cause of the causes of j. Thus, the causes of the causes of j never provide fundamental information about i. Then, any relation between i and the causes of j will eventually be mediated by the causes of i. This is because the causes of i are the most proximal sources of information about i. This is the essence of Axiom 4, which is illustrated in figure 4.

While Axioms 1 through 4 are our basic axioms, Axiom 5 is a supplementary axiom that is relevant for Theorem 2. We present it here in the interest of keeping all axioms, and their corresponding discussions, contained within a single section.

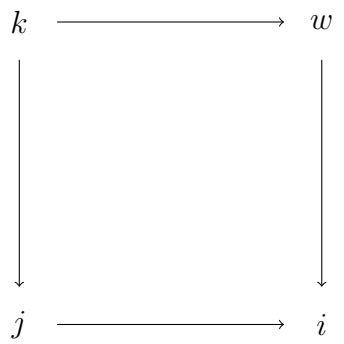


Figure 5: Observing or intervening j makes the DM update differently about k. This difference in updating may affect the DM's beliefs about i.

Axiom 5. $(\forall i \in \mathcal{N})$, $(\forall \mathcal{J} \subset \mathcal{N} \backslash \{i\})$ $(\forall f, g \in \mathbb{R}^{X_i})$, $(\forall x_{Ca(i) \cup \mathcal{J}} \in X_{Ca(i) \cup \mathcal{J}})$,

$$\mathbb{1}_{\{x_{Ca(i)}\}} f \succ \mathbb{1}_{\{x_{Ca(i)}\}} g \Leftrightarrow \mathbb{1}_{x_{Ca(i) \backslash \mathcal{J}}} f \succ_{x_{\mathcal{J}}} \mathbb{1}_{x_{Ca(i) \backslash \mathcal{J}}} g.$$

Axiom 5 states that two particular decision problems are equivalent. Given a variable i and acts $f, g \in \mathbb{R}^{X_i}$, the first problem is to choose f or g when their payments are contingent on the causes of i obtaining a particular value, $x_{Ca(i)}$. In the second decision problem, the DM intervenes a subset of causes of i, (say, $\mathcal{J} \subset Ca(i)$) to the value $x_{\mathcal{J}}$, and the payments of f and g are now contingent on the values of the non-intervened causes, $x_{\mathcal{J}^c}$, being realized. From a numerical standpoint, both these situations result in the same value for the causes of i (namely, $x_{Ca(i)}$); the difference is *how* those values are achieved. In the first problem it is simply by selecting a standard Savage conditional act, whilst in the second problem it is by a combination of interventions and Savage conditional acts acts. Because Axiom 5 imposes that these two problems are treated identically, Axiom 5 implies that the only aspect of interventions that matters is the value the intervention sets for the variable. In other words, intervening a variable does not change the DMs structural view of the world.

We use Figure 5 below to illustrate the normative appeal of Axiom 5.

First, we explain why Axiom 5 involves sets that are weakly larger than $Ca(i)$. Suppose a DM has to choose between two acts over i (say, $f, g \in \mathbb{R}^{X_i}$) whose

payments are contingent on j taking value x_j. That is, the DM has to choose between $\mathbb{1}_{x_j} f$ and $\mathbb{1}_{x_j} g$. Note that $\{j\}$ is a strict subset of $Ca(i)$. Observing that j took the value x_j gives the DM information about the value of k; in turn, this information about k gives the DM information about w which, ultimately, gives the DM information about i. Thus, observing that j took the value x_j is informative about i in two ways: directly, because $j \in Ca(i)$, and indirectly, via k and w. If the DM intervenes j at value x_j, he receives the same direct information about i but loses the indirect information mediated via k and w. Thus, the DM could say that $\mathbb{1}_{x_j} f \succ \mathbb{1}_{x_j} g$ but $g \succ_{x_j} f$. Clearly, observing x_j or intervening variable j to value x_j are different problems in terms of the DMs updating.

Now consider the situation above but where the payments of f and g involve *all* causes i, j *and* w. That is, for some x_j and x_w, the DM chooses between $\mathbb{1}_{x_j,x_w} f$ and $\mathbb{1}_{x_j,x_w} g$. For concreteness, suppose that $\mathbb{1}_{x_j,x_w} f \succ \mathbb{1}_{x_j,x_w} g$. If the DM intervened j to the value x_j and then had to choose between $\mathbb{1}_{x_w} f$ and $\mathbb{1}_{x_w} g$, would the DM lose any information? Put differently: if, at a cost $\varepsilon > 0$ the DM could intervene the value of j to x_j, rather than simply conditioning his choice on value x_j being realized, is there a $\varepsilon > 0$ small enough that the DM would pay for this option? In both problems, w is observed to take the value x_w; therefore, any information j could indirectly provide about i through w is still directly captured in the observation of x_w. Thus, intervening j entails no information loss relative to simply observing that j took the value x_j. Thus, the DM has the same information in both problems and should thus treat the problems equivalently. This result is precisely what Axiom 5 requires.

Both of the above discussions treated $\mathcal{J} \subset Ca(i)$, but to complete our discussion of Axiom 5 we must allow that $\mathcal{J}$ contains non-causes of i. Axiom 5 states that once we know the value of all the causes of i, intervening variables that are not causes of i is uninformative about i. In Figure 5, if an act's payments are contingent on x_w and x_j, then intervening the value of k to some x_k is uninformative about i.

6 Representation

In this section we define the representation we seek for $\bar{\succ}$. Since $\bar{\succ}$ will ultimately be associated with a collection of probability distributions, we proceed in two steps.

First define what it means for a DAG to represent a single probability distribution. Then, we generalize to a family of probability distributions. For a reminder of our graph theoretic notation see Section 3.1.

Lauritzen et al. [14] provide a definition for when a DAG represents a probability distribution, say $\mu \in \Delta(\Pi_{i\in\mathcal{N}} X_i)$. The objective of such a definition is to represent graphically the conditional independence structure of μ. Let $\mu \in \Delta(\Pi_{i\in\mathcal{N}} X_i)$, and let $G = (\{1, ..., N\}, E)$ be a DAG. The chain rule implies the following

$$(\forall x \in \Pi_{i\in\mathcal{N}} X_i),\ \ \mu(x) = \Pi_{i=1}^{N} \mu(x_i | ND(i)). \tag{6}$$

Now consider the DAG in figure 6 below.

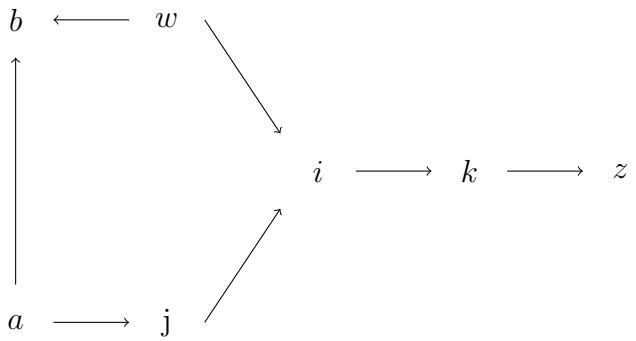


Figure 6: A DAG representing the distribution
$\mu(a, b, w, j, i, k, z) = \mu(a)\mu(w)\mu(b|w, a)\mu(j|a)\mu(i|w, j)\mu(k|i)\mu(z|k)$.

In a DAG as the one above, an arrow between two nodes represents that the two nodes are never statistically independent. In this way, arrows encode which variables provide *fundamental information* about other variables, in the sense that the information transmitted by the source is not contained in any other variable. For instance, the DAG in figure 6 conveys that w and j contain fundamental information about i and thus i is never independent from $\{w, j\}$. Likewise, i is never independent of it's direct descendant, k. Now, consider a variable that is an ancestor of i; for example, a. Clearly, a and i are not independent: a provides fundamental information about j which provides fundamental information about i. However, any information a has about i is implicitly encoded in $j \in Pa(i)$. Indeed, if a carried fundamental information about i, there should be an arrow

$a \to i$, but such arrow is absent. Likewise, b provides information about i: b is informative about $\{a, w\}$, both of which are informative about i. However, any information b has about i is encoded in $\{j, w\}$. What this implies is that, once we condition on the parents of i (in this case, $\{w, j\}$), all non-descendants of i are conditionally independent of i. Therefore, the terms $\mu(x_i|ND(i))$ in equation 6 simplify to $\mu(x_i|Pa(i))$. This observation motivates definition 5 below.

Definition 5. *Let $\mu \in \Delta(\Pi_{i\in\mathcal{N}} X_i)$. A DAG $(\{1, ..., N\}, E)$ represents μ if, and only if, the following hold:*
$(\forall x \in \Pi_{i\in\mathcal{N}} X_i)$,

$$\begin{aligned} \mu(x) &= \Pi_{i=1}^{N} \mu(x_i|Pa(i)) \\ (\forall(\mathcal{T}_i)_{i\in\mathcal{N}})(\mathcal{T}_i \subset Pa(i)),\ \textit{if}\ \mu(x) &= \Pi_{i=1}^{N} \mu(x_i|\mathcal{T}_i) \Rightarrow (\forall i \in \mathcal{N}), \mathcal{T}_i = Pa(i) \end{aligned}$$

Definition 5 makes two statements. First, a DAG represents a probability distribution if, and only if, the DAG summarizes the conditional independence properties of μ in the sense discussed previously. Second, the set of parents is the smallest set that allows for such a decomposition. Indeed, consider a set of nodes $V = \{a, b, c\}$ and a probability distribution $\mu(x_a, x_b, x_c) = \mu(x_a)\mu(x_b)\mu(x_c)$. Since all variables are statistically independent, both DAGs in Figure 7 represent this μ. Indeed, both $\mu(x_a, x_b, x_c) = \mu(x_a)\mu(x_b|x_a)\mu(x_c|x_a)$ and $\mu(x_a, x_b, x_c) = \mu(x_a)\mu(x_b)\mu(x_c)$ are true statements. However, the first representation includes irrelevant arrows: the minimality requirement prevents this.

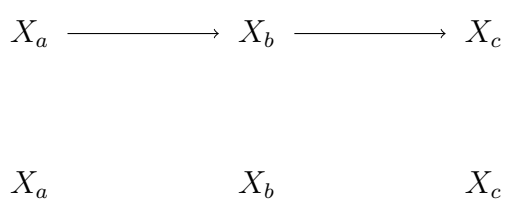


Figure 7: Both DAGs above represent the same probability distribution, $\mu(x_a, x_b, x_c) = \mu(x_a)\mu(x_b)\mu(x_c)$, but the top one includes irrelevant arrows.

Using definition 5 we can define when a graph represents a standard Savage preference. Suppose that $\succ$ was the DM's Savage preference defined on $\mathbb{R}^X$. Under

Assumption 1 there is a well defined belief representation of $\succ$, $\mu \in \Delta(X)$. We can then say that a graph G represents $\succ$ if G represents μ, in the sense of definition 5.

However, definition 5 is not enough to define when a graph represents a preference $\bar{\succ}$. Indeed, a preference $\bar{\succ}$ is associated with the collection of induced Savage preferences $\{\succ_p : p \in \mathcal{P}\}$. As such, $\bar{\succ}$ is associated to a family of beliefs, rather than a single belief, as in Savage's model. Thus, to define when a DAG represents preferences $\bar{\succ}$, we first define what it means for a DAG to represent a collection of probability distributions rather than a single probability distribution.

To define when a DAG represents preferences $\bar{\succ}$, we first define the *truncation* of a DAG. Let $G = (V, E)$ be a DAG, and let $W \subsetneq V$. The W-truncated DAG, G_W, is the DAG obtained by eliminating all nodes in W, together with their incoming and outgoing arrows. Formally, $G_W = (V \backslash W, E \cap W^{\complement} \times W^{\complement})$. This DAG is a useful representation of intervention beliefs. After variables in W are intervened, they no longer form part of the DM's statistical model; they are now deterministic objects that are statistically uninformative about the value of their parents. Thus, we exclude these variables from the corresponding DAG. For example, if Alex observes that Mr. Kane obtained a college degree, then his education level is no longer random, but Alex can still make inference about Mr. Kane's intellectual ability. Thus, education remains a legitimate element of Alex's statistical model. However, if Mr. Kane's education is intervened to "college degree", then his eduction level is no longer random and, furthermore, is uninformative about his ability level. Thus, we exclude education level from the DM's post-intervention model.

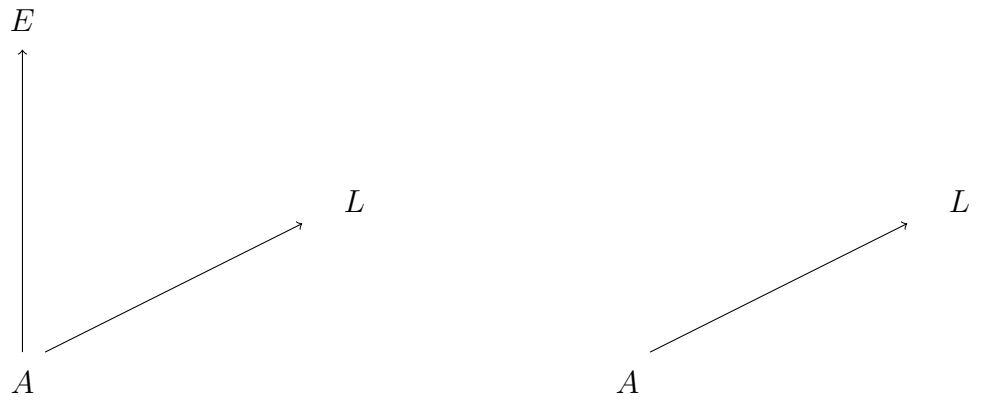


Figure 8: Right: the full econometric model; knowing E is informative about L because knowing E is informative of it's cause, A.
Left: once E is intervened it is no longer part of the econometric DAG since E is uninformative about it's causes.

We can now define when a DAG, G, represents a preference $\bar{\succ}$. We then say a graph represents a preference if the appropriately truncated subgraph represents the corresponding intervention preference and the arrows are consistent with the direction of causality. This is formally presented in definition 6 below.

Definition 6. *Let $G = (\mathcal{N}, E)$ be a DAG and $\bar{\succ}$ be a DM's preference. Assume that for each $\mathcal{T} \subset \mathcal{N}$ and each $x_{\mathcal{T}} \in X_{\mathcal{T}}$, $\succ_{x_{\mathcal{T}}}$ has a well defined belief representation; let $\mu_{x_{\mathcal{T}}}$ be the corresponding belief representation. We say that G* represents *$\bar{\succ}$ if the following are true for each $\mathcal{T} \subset \mathcal{N}$ and each $x \in X$:*

i $G_{\mathcal{T}}$ represents $\mu_{x_{\mathcal{T}}}$,

ii If $(i,j) \in E$ then $\mu_{x_{\{i,j\}^{\complement}}}(x_j|x_i) = \mu_{x_{\{j\}^{\complement}}}(x_j)$.

Notice that nothing in this section is related to causality. Indeed, the statement that a graph represents a probability distribution is purely a statement about statistical independence. As such, the representation of a probability by a DAG is a statement about correlation, not causation. At this point in the exposition, DAGs used to represent probability distributions and DAGs used to represent causal statements are completely unrelated. It is precisely the job of Theorems 1 and 2 to show the exact conditions under which a DAG can simultaneously represent the DMs beliefs as well as the DMs causal model.

7 Results

Our first theorem is Theorem 1, stated below.

Theorem 1. *Let $\bar{\succ}$ satisfy Assumption 1. The following are equivalent:*

i Axioms 1 through 4 hold,

ii $(\exists G)$ *such that G is a DAG and represents $\bar{\succ}$ in the sense of Definition 6.*

Furthermore, if G represents $\bar{\succ}$, then $G = G(\bar{\succ})$.

As mentioned in section 2, the literature of Bayesian graphs assumes that causal DAGs fulfill a dual role: they represent both causal assumptions and assumptions on conditional independence. This is clearly a joint assumption about how the analyst defines causality, and how the analyst's definition of causality interacts with statements of conditional independence. Theorem 1 states the exact conditions under which a DAG can fulfill this dual role.

In particular, the uniqueness result implies that Definition 2 is the only definition of causality that satisfies our axioms. Suppose a researcher has a definition of causality in mind (say, C) so that statements of the form "i causes j according to criterion C" are well defined. If C satisfies our axioms then C can be represented via a DAG, G, such that $i \to j$ holds if, and only if, "i causes j according to criterion C". The uniqueness claim in Theorem 1 says that $G = G(\bar{\succ})$. Therefore, $i \to j$ also holds if, and only if, i causes j in the sense of definition 2. Thus, C must coincide with Definition 2.

Theorem 1 also provides a foundation for unifying and structuring our understanding of causation. The theorem states that any formal discussion of causality (as understood by Definition 2) begins with two items: a collection of probability laws, $\{\mu_p \in \Delta(X_{\mathcal{N}(p)}) : p \in \mathcal{P}\}$, and a DAG, G, that represents those laws. Models that include these components can legitimately be called models of "causation" regardless of any other details the model might include. However, models that cannot be phrased in terms of intervention beliefs and their representing DAG are not models of causality (again, as understood by Definition 2). In short, researchers that find our axioms normatively appealing, and that agree that definition 2 is a

sensible definition of causal effect, are encouraged to use DAG-based models for conducting causal inference. Researchers who find our axioms normatively unappealing, or disagree with definition 2 as a sensible definition of causal effect, are encourage to stay away from DAG-based models. In this way, Theorem 1 provides a foundation for selecting among models with which to empirically study causal effects. As usual, whether the axioms are normatively appealing or not is something each reader has to decide for themselves.

7.1 Identification of intervention beliefs

In this section, we consider the following question. Let $\mu \in \Delta(X)$ be the DM's beliefs elicited from his Savage preference and μ_p be the DM's beliefs elicited from an intervention preference $\succ_p$. When can we express μ_p as a function of μ? Proposition 1 and Theorem 2 in this section answer this question.

Answering the question above is useful to make the model applicable to empirical research. When μ_p is expressed in terms of μ (henceforth, when μ_p is *identified*), any information that allows a DM to update his Savage beliefs, μ, also allows the DM to update his intervention beliefs, μ_p. If an analyst had access to a perfectly controlled setting, the analyst could estimate each μ_p directly and models of causal inference would be unnecessary. However, most empirical work in economics is observational, in the sense that direct policy interventions are unavailable to the researcher. Proposition 1 and Theorem 2 bridge the gap between intervention beliefs –what the econometrician *wants* to estimate– and standard conditional probabilities –what the econometrician *can* estimate.

When added to Axioms 1 through 4, Axiom 5 yields a model in which different intervention beliefs, μ_p, can be expressed in terms of μ. In what follows, we remind the reader of Axiom 5 and illustrate Theorem 2 by means of two simple examples. Then, we state and discuss the general form of Theorem 2.

Axiom 5. $(\forall i \in \mathcal{N})$, $(\forall \mathcal{J} \subset \{i\}^{\complement})$ $(\forall f, g \in \mathbb{R}^{X_i})$, $(\forall x_{Ca(i)\cup\mathcal{J}} \in X_{Ca(i)\cup\mathcal{J}})$,

$$\mathbb{1}_{\{x_{Ca(i)}\}} f \succ \mathbb{1}_{\{x_{Ca(i)}\}} g \Leftrightarrow \mathbb{1}_{x_{Ca(i)\setminus\mathcal{J}}} f \succ_{x_{\mathcal{J}}} \mathbb{1}_{x_{Ca(i)\setminus\mathcal{J}}} g.$$

Example 2. *Blake is an econometrician who believes that ability causes both*

education and lifetime earnings and also that education causes lifetime earnings. This model is graphically depicted in Figure 9. To understand the direct effect of education on lifetime earnings, Blake has to understand how $\mu_{a,e}(\cdot)$ changes with $e \in E$, for each fixed $a \in A$. However, Blake cannot access a controlled environment, so Blake has no data on $\mu_{(a,e)}$. However, under Axiom 5 data from controlled environments is unnecessary. When $\mathcal{J} = \{A, E\}$, Axiom 5 implies $\mu_{(a,e)}(\cdot) = \mu(\cdot|a, e)$. Thus, the direct causal effect of education on lifetime earnings is calculated by computing how $\mu(\cdot|a, e)$ varies with e for each value of a. Note that $\mu(\cdot|a, e)$ is a standard conditional probability, and data on this quantity can be estimated with access to observational datasets. Blake can therefore use data from outside a controlled environment to form his intervention beliefs.

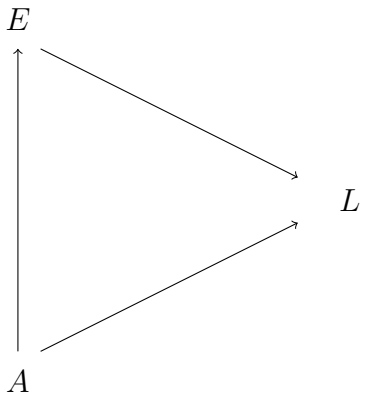


Figure 9: Causal effects are identified: $\mu_{(a,e)(l)=\mu(l|a,e)}$.

Example 3. *Charlie is a colleague of Blake. However, Charlie believes that people are not born with intrinsic ability. By the contrary, it is education that causes ability, and this ability is the sole cause of lifetime earnings. Charlie's causal DAG is depicted in Figure 10. Charlie is interested in studying the* indirect *effect that education policies have on lifetime earnings, which can be done by applying Axiom 5 twice. First, set $\mathcal{J} = \{E\}$, $i = A$ to obtain $\mu_a(e) = \mu(e|a)$ for each $(a, e) \in A \times E$. Second, set $\mathcal{J} = \{E\}$, $i = L$ to obtain $\mu_e(l|a) = \mu(l|a)$ for each*

$(e, a, l) \in E \times A \times L$. *Finally, we obtain the following derivation.*

$$\begin{aligned}\mu_e(l) &= \sum_a \mu_e(l, a) \\ &= \sum_a \mu_e(l|a)\mu_e(a) \\ &= \sum_a \mu(l|a)\mu(a|e).\end{aligned}$$

Thus, calculating the indirect effects of E and L requires computing $\mu(l|a)$ and $\mu(a|e)$, both of which can be computed with data from observational studies. Even if access to a controlled environment is unavailable, the identification of μ_e implies that such data is unnecessary.

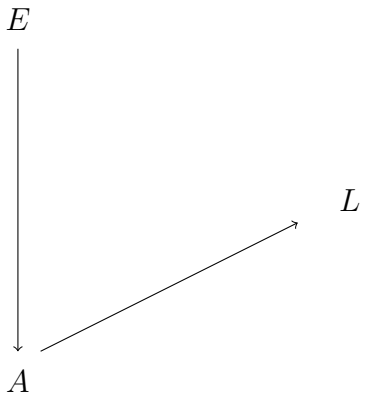


Figure 10: Indirect causal effect of E on L is identified: $\mu_e(l) = \sum_a \mu(l|a)\mu(a|e)$.

The examples highlight two simple cases in which intervention beliefs are identified. First, if j is a cause of i, then the direct causal effect that j has on i is identified via the formula $\mu_{x_{\{i,j\}^\complement}, x_j}(x_i) = \mu(x_i|x_j, x_{Ca(i)\setminus\{j\}})$. Thus, one can obtain the direct causal effect of j on i by conditioning on all causes of i and analyzing how that conditional probability varies with x_j. Similarly, if j causes k, k causes i, and this is the only connection between j and i, the indirect causal effect of j on i is calculated by following the chain rule: $\mu_{x_j}(x_i) = \sum_{x_k} \mu(x_i|x_k)\mu(x_k|x_j)$. However, other intervention beliefs may also be identified. The rest of this section is devoted to understanding the exact conditions under which intervention beliefs

are identified.

Given a family of intervention beliefs and a DAG that represents these beliefs, what is the set of all intervention beliefs that are identified, and how are they identified? Answering this question requires two definitions : we need to define specific truncations of a DAG and we need the definition of a blocked path. We provide these definitions and then formally state the result. Appendix B discusses the intuition behind why we need these definitions.

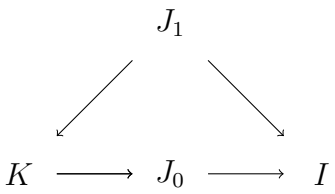
Definition 7. *Given G and three disjoint sets of variables $\mathcal{I}, \mathcal{J}, \mathcal{K} \subset \mathcal{N}$, the truncated DAGs $G_{\mathcal{I}^{in}}$, $G_{\mathcal{J}^{out}}$, and $G_{\mathcal{I}^{in},\mathcal{J}(\mathcal{K})^{out}}$ are defined as follows:*

1 $G_{\mathcal{I}^{in}}$ is obtained from G by eliminating all arrows pointing to nodes in $\mathcal{I}$,

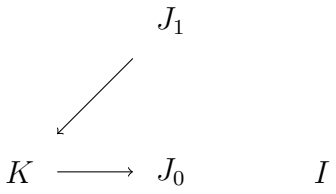
2 $G_{\mathcal{I}^{in},\mathcal{J}^{out}}$ is obtained from G by eliminating all arrows emerging from nodes in $\mathcal{J}$ and all arrows pointing to nodes in $\mathcal{I}$,

3 $G_{\mathcal{I}^{in},\mathcal{J}(\mathcal{K})^{in}}$ is obtained by eliminating all arrows pointing to nodes in $\mathcal{J}(\mathcal{K})$ and $\mathcal{I}$, where $\mathcal{J}(\mathcal{K})$ is the set of $\mathcal{J}$ nodes that are not ancestors of any $\mathcal{K}$ nodes in $G_{\mathcal{I}^{in}}$.

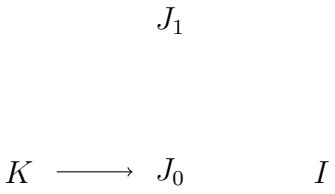
The following figures show the base DAG, G, and its corresponding truncations. In all cases, $\mathcal{J} = \{J_0, J_1\}$, $\mathcal{I} = \{I\}$, $\mathcal{K} = \{K\}$.



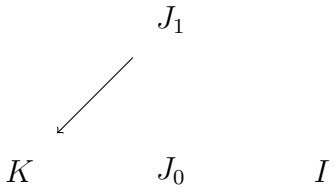
(a) The base DAG, G.



(b) DAG $G_{\mathcal{I}^{in}}$ obtained by eliminating all arrows into I.



(c) DAG $G_{\mathcal{I}^{in},\mathcal{J}^{out}}$ obtained by: (i) eliminating arrows into I, and (ii) eliminating all arrows emerging from J_0 and J_1.



(d) DAG $G_{\mathcal{I}^{in},\mathcal{J}(\mathcal{K})^{in}}$ obtained by: (i) eliminating all arrows into $\mathcal{I}$, and (ii) eliminating all arrows into J_0 since J_0 is the only $\mathcal{J}$ node that is not an ancestor of a $\mathcal{K}$ node.

Figure 11: Different truncations of a *DAG*.

For the following definition, suppose Q is an undirected path between two nodes, *i.e.* a collection of nodes, regardless of directionality, and that q is a node on Q. For example, Figure 11 shows an undirected path $Q = (J_1, I, J_0, K)$ from J_1 to K. We say that Q has *converging arrows at* q if there exist nodes q_0 and q_1 that are adjacent to q in Q such that $q_0 \to q \leftarrow q_1$. For example, path $Q = (J_1, I, J_0, K)$ has converging arrows at I. We say that Q does not have converging arrows at q if for all nodes q_0 and q_1 that are adjacent to q in Q, either $q_0 \to q \to q_1$ or $q_0 \leftarrow q \to q_1$ holds. For example, $Q = (J_1, I, J_0, K)$ does not have converging arrows at J_0.

Definition 8. *Let $\mathcal{I}, \mathcal{J}, \mathcal{K}$ be three disjoint sets of variables, and let Q be an undirected path between a node in $\mathcal{I}$ and a node in $\mathcal{J}$. We say* $\mathcal{K}$ blocks Q *if there exists a node q on Q such that one of the following conditions holds:*

- *Q has converging arrows at q, and neither q nor any of its descendants is in $\mathcal{K}$,*

- *Q does not have converging arrows at q and $q \in \mathcal{K}$.*

Below, we state the two *rules of causal calculus* and Proposition 1. Versions of these rules are well known in causal statistics (see Pearl [16] and Huang and Valtorta [11]), and we comment on the connections in the next section.

Rule 1. *(Exchanging intervention and observation.) Let $\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ be disjoint sets of variables. If $\mathcal{I}_1 \cup \mathcal{I}_3$ block all paths from $\mathcal{I}_0$ to $\mathcal{I}_2$ in graph $G_{\mathcal{I}_1^{in}, \mathcal{I}_2^{out}}$, then*

$$\mu_{x_{\mathcal{I}_1}, x_{\mathcal{I}_2}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}) \quad = \quad \mu_{x_{\mathcal{I}_1}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_2}, x_{\mathcal{I}_3}). \tag{7}$$

Rule 2. *(Eliminating interventions.) Let $\mathcal{I}_0, \mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ be disjoint sets of variables. If $\mathcal{I}_1 \cup \mathcal{I}_3$ block all paths from $\mathcal{I}_0$ to $\mathcal{I}_2$ in graph $G_{\mathcal{I}_1^{in}, \mathcal{I}_2(\mathcal{I}_3)^{in}}$, then*

$$\mu_{x_{\mathcal{I}_1}, x_{\mathcal{I}_2}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}) \quad = \quad \mu_{x_{\mathcal{I}_1}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}). \tag{8}$$

Proposition 1. *Let $\succsim$ satisfy Axioms 1 through 4, let G represent $\succsim$, and let $\{\mu_p : p \in \mathcal{P}\}$ be the DM's intervention beliefs. Then, the following statements are equivalent.*

- *$\succsim$ satisfies Axiom 5.*
- *Rules 1 and 2. Furthermore, if μ_p is identified for some $p \in \mathcal{P}$, then the identification is obtained by iterative application of these two rules.*

With Proposition 1, we can refer to Example 3 and obtain the identification result by applying Rules 1 and 2. In Rule 2, set $\mathcal{I}_0 = \{L\}$, $\mathcal{I}_1 = \varnothing$, $\mathcal{I}_2 = \{E\}$, and $\mathcal{I}_3 = \{A\}$. The corresponding truncated DAG is G itself. In G, A blocks the unique path from E to L since no converging arrows exist at A. Thus, $\mu_e(l|a) = \mu(l|a)$. Likewise, in Rule 1, set $\mathcal{I}_0 = \{A\}$, $\mathcal{I}_1 = \varnothing$, $\mathcal{I}_2 = \{E\}$, and $\mathcal{I}_3 = \varnothing$. In the truncated graph that results, E is isolated from all other variables, so any path from E to A is blocked; thus, $\mu_e(a) = \mu(a|e)$. These two conclusions yield the identification of $\mu_e(l) = \sum_a \mu(l|a)\mu(a|e)$.

7.2 Markov representations and do-probabilities

Proposition 1 is obtained purely from adding Axiom 5 to the list of Axioms imposed on $\bar{\succ}$. As such, Rules 1 and 2 depend only on the Axioms. However, Pearl [16] and Huang and Valtorta [11] obtain similar results using a formalism called *do-probability*. In this section we define what do-probabilities, and we explore the connections between do-probabilities and intervention beliefs. We also explore the contribution of our results to the do-probability framework.

Definition 9. *Let $\mu \in \Delta(X)$. For each $i \in \mathcal{N}$, let μ_i be the marginal over X_i. For each $i \in \mathcal{N}$, let ε_i be a random variable with range $\mathcal{E}_i$, let G be the DAG defined by a family of sets of parents $(Pa(i))_{i\in\mathcal{N}}$, and let h_i be a function $h_i : X_{Pa(i)} \times \mathcal{E}_i \to X_i$. Let ϕ be the joint distribution of the vector $(\varepsilon_1, ..., \varepsilon_N)$. A* Markov representation *of μ is a tuple $(G, (h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$ that satisfies the following:*

- *$(\forall i, j)$, ε_i is independent of ε_j,*
- *μ can be recovered implicitly as a solution to the following system of equations:*

$$\mu_i(x_i) = \phi(\{\varepsilon : h_i(x_{Pa(i)}, \varepsilon_i) = x_i\}), \quad (i \in \{1, ..., N\}). \tag{9}$$

Markov representations are used in statistical causality to numerically represent causal effects (see Pearl [16]). The interpretation is as follows. Each variable i is a deterministic function of a set of variables, $Pa(i)$, and idiosyncratic noise, ε_i. Each h_i is interpreted as a random production function for variable i, with $Pa(i)$ as the set of inputs and ε_i as the random component. The causal effect of a variable j on i is (loosely speaking) calculated by observing how $h_i(\cdot)$ changes as we change the value of variable j. For a more precise statement, we need the definition of do-probability, which we take from Pearl [16]. See examples 4 and 5 in Appendix C for a concrete illustration of how to calculate do-probabilities and how they differ from standard conditional probabilities.

Definition 10. *Let $\mu \in \Delta(X)$ be a probability distribution, and let $((h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$ be a Markov representation of μ. Given two disjoint sets of variables, $\mathcal{I}$ and $\mathcal{J}$, the do-probability $\mu(x_\mathcal{I}|do(x_\mathcal{J}))$ is calculated as follows:*

1 *For all* $j \in \mathcal{J}$*, eliminate from system (9) in Definition 9 all the formulas* $\mu_j(x_j) = \phi(\{\varepsilon : h_j(x_{Pa(j)}, \varepsilon_i) = x_j\}$.

2 *For each* $i \notin \mathcal{J}$ *and for each* $j \in Pa(i) \cap \mathcal{J}$*, input value* x_j *into the corresponding formula in system (9) of Definition 9.*

3 *Calculate the probability of realization* $x_{\mathcal{I}}$ *in the model resulting from applying steps* 1 *and* 2 *above.*

While do-probabilities are commonly referred to as the causal effect of one variable on another, it is important to be cautious with the language. Do-probabilities reflect the effect that an intervention on a set of variables has on the whole system of equations; that is, do-probabilities capture both the direct and indirect effects of interventions. For example, consider the DAG in Figure 12. This DAG states that there is no direct causal effect of A on C; however, $Pr(x_C|do(x_A)) = Pr(x_C|x_A)$, which is a non-trivial function of x_A. Indeed, intervening A has an effect on B, which in turn, affects C. In this example, $Pr(x_C|do(x_A))$ captures this indirect effect. In line with our definition of causal effect, the causal effect of A on C is given by how $Pr(x_C|do(x_A, x_B))$ changes with x_A. In this case, $Pr(X_c|do(x_A, x_B))$ is a constant function of x_A, which is consistent with A having no direct causal impact on x_C.

Figure 12: A has no direct causal effect on C, but $pr(x_C|do(x_A))$ is a non-trivial function of x_A.

Having defined Markov representations and do-probabilities, we can now state Theorem 2.

Theorem 2. *Let* $\succsim$ *satisfy Assumption 1, and let* $(\mu_{x_{\mathcal{I}}})_{\mathcal{I} \subset \mathcal{N}}$ *be the subjective beliefs elicited from* $\succsim$*. The following statements are equivalent:*

- *Axioms 1, 2, and 5 hold,*
- *There exists a Markov representation of* μ*,* $(G, (h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$*, such that*

– $(\forall \mathcal{J} \subset \mathcal{N})$, $(\forall x_{\mathcal{J}} \in X_{\mathcal{J}})$; $\mu_{x_{\mathcal{J}}} = \mu(\cdot|do(x_{\mathcal{J}})) \in \Delta(X_{\mathcal{J}^{\complement}})$,

– G *represents* $\bar{\succ}$.

Furthermore, if G *represents* $\bar{\succ}$, *then* $G = G(\bar{\succ})$.

The crucial contribution of Theorem 2 is that it clarifies the role of do-probabilities in the understanding of causal effects. Do-probabilities are often presented as the definition of a causal effect. As Pearl writes in [17]: " *the definition of a "cause" is clear and crisp; variable* X *is a probabilistic-cause of variable* Y *if* $P(y|do(x)) \neq P(y)$ *for some values* x *and* y." Theorem 1 states that one can legitimately represent causal effects based on interventions via a DAG which, nonetheless, is incompatible with any system of do-probabilities. The causal DAG will be compatible with a set of do-probabilities only when adding Axiom 5 to the list of basic axioms. This result is analogous to the exercise conducted by Machina-Schmeidler [15]: just as expected utility and probabilistic sophistication can be behaviorally separated, we show that the graph theoretic aspects of Pearl-like models can be separated from the do-probability formalism. The substantive assumptions about causality are conveyed by the DAG, while do-probabilities represent an additional assumption about when interventions and simple observations can be used interchangeably. In short, the notion of causality represented by a do-probability is strictly stronger that the notion of causality represented by a DAG.

Theorem 2 further clarifies that Axiom 5 is the fundamental property that links do-probabilities with intervention beliefs. When defining Markov representations, the functions $h(\cdot)$ are not indexed by whether their arguments have been observed or intervened. The functions $h(\cdot)$ only care about the numerical values of their arguments and not the method through which these numerical values are obtained. This is an implicit assumption of the Pearl model and it is delivered by Axiom 5.

Jointly, Proposition 1 and Theorem 2 imply that Pearl's rules of causal calculus serve as an axiomatization of do-probability. Indeed, Huang and Valtorta [11] show that, in a do-probability model, Rules 1 and 2 summarize all obtainable identification results. To the best of our knowledge, whether other probabilistic models are consistent with the aforementioned result is unknown. We show that when Rules 1 and 2 summarize all obtainable identification results, Axiom 5 must

hold so that intervention beliefs are do-probabilities. Therefore, the rules of causal calculus are a complete description of all obtainable identification results if, and only if, the intervention probabilities are do-probabilities.

As a final remark on Theorem 2, notice that Definition 9 implicitly requires that the Markov representation that defines do-probabilities has a unique solution. While this characteristic has sometimes been pointed to as a limitation of the theory (see Halpern [6]), under Axiom 5, this result is without loss of generality.

8 Literature Review

In economic theory, the work most closely related to ours is a series of papers by Spiegler ([20], [21], [22]). The main difference is the focus of the papers. Spiegler's work does not provide a definition of the term "causal effect", except that it can be represented via a DAG that satisfies two properties. First, the DAG factorizes the correlation structure in the DM's beliefs; second, the arrows in the DAG are interpreted as pointing from cause to effect. Given these assumptions, Spiegler asks what types of mistake a DM with a misspecified causal model might make. In our paper, we first define what a causal relation is and then seek to understand which axioms on behavior allow us to represent causal effects in the language of DAGs that factorize the DM's beliefs. The uniqueness claim in Theorem 1 provides the point of contact between both papers. Under our definition of causality, a DAG can simultaneously factorize the DM's beliefs while retaining a causal interpretation only if Axioms 1 through 4 hold. Furthermore, under the axioms, a graph G both represents a DM's correlation structure and is interpreted causally (in the sense that arrows point from cause to effect) only is the definition of causal effect is as in Definition 2.

In decision theory, Karni ([12], [13]) explores models where a DM can affect the states that are realized. In those papers, the primitive objects are a set of actions and a set of consequences. States of nature are defined as a mappings from actions a DM might take to consequences that arise from those actions; that the mapping $Action \to Outcome$ is stochastic reflects that states are stochastically realized. A DM can affect the states that occur by making an appropriate choice of action. This idea is similar to our idea of a policy intervention, since a policy p can be

seen as an action the DM takes that affects the realization of states. Indeed, we can set Karni's set of action to our set $\mathcal{P}$, Karni's outcomes to realizations of our state space, X, and a Karni state is a mapping $s : \mathcal{P} \to X$. The main difference arises in that we impose –objectively– a consistency condition: if a policy p intervenes variable j to value x_j, a state s cannot map this policy to a realization $x'_j \neq x_j$. Karni has a version of this condition but it is imposed subjectively in the preferences. Moreover, the focus of Karni's paper is not to use these ideas to talk about causal effects, or understand what types of models reflect normative definitions of causality. Rather, Karni focuses on obtaining subjective expected utility representations of his preferences. For this reason, while a strong formal connection exists, the substance of the research agenda is different.

The statistics and computer science literature includes research that uses graphical methods to represent the conditional independence structure of any given joint probability law (see Dawid [1], Geiger et al. [4], Lauritzen et al. [14]). Specifically, Dawid [1] and Geiger et al. [4] show that, given a probability distribution over a set of variables, $p(\cdot)$, and given a graph G that represents p, the *D-separation* criterion for graphs (see Definitions 8 and 11) summarizes the independence structure of p. Our proofs rely on the one-to-one correspondence between variables that satisfy the D-separation criterion and variables that are conditionally independent. This is the main point of contact between our paper and that body of work. Lauritzen et al. [14] provide alternative graphical tests for D-separation which may be used to obtain alternative proofs for our results.

In causal statistics, the most closely related papers are those in the Bayesian networks literature (see Spirtes [23], Pearl [16], and follow-up work). Two main points of contact between that literature and our paper exist. First, the statistical causality literature offers no formal definition of the term "causal relation", and the exact meaning of this phrase is left to the researcher's common sense. As Pearl states "*The first step in this analysis is to construct a causal diagram such as the one given in Fig. [1] (sic.), which represents the investigator's understanding of the major causal influences among measurable quantities in the domain*" and later " *The purpose of the paper is not to validate or repudiate such domain-specific assumptions but, rather, to test whether a given set of assumptions is sufficient for*

quantifying causal effects from non-experimental data, for example, estimating the total effect of fumigants on yields". Second, the numerical value of the causal effect of one variable on another (say, Education on Lifetime earnings) is given by the *do-probability* formalism. As Pearl writes in [17]: " *the definition of a "cause" is clear and crisp; variable X is a probabilistic-cause of variable Y if $P(y|do(x)) \neq P(y)$ for some values x and y.*" By contrast, we show that, under Axioms 1 through 5, there exists a unique definition of causal effect that is both representable via a DAG and consistent with an interventionist perspective of causality. Thus, we show that causal models based on causal diagrams implicitly impose a specific definition of causality. Moreover, Axioms 1 through 5 neither imply, nor are implied by, a representation of causality in terms of do-probabilities. Contrary to Pearl's quote, do-probabilities neither define nor are defined by the definition of causality embodied by the causal diagram. Theorem 2 shows that under Axioms 1 through 5, causal effects are representable via a DAG that is compatible with the do-probability formulas. This makes explicit the fundamental restrictions imposed by using do-probabilities to numerically quantify causal effects.

In terms of axiomatic definitions for causal effects, Galles and Pearl [3], Halpern [6], and Halpern and Pearl ([7], [8]) provide an alternative approach. Specifically, Halpern [6] expands on Galles and Pearl [3] and axiomatizes a more general model. Rather than a decision theoretic approach, Halpern [6] axiomatizes causal effects through a syntactic logic approach; that is, rather than using a DM's preferences over a suitably defined choice domain as a primitive, Halpern's axiomatization is in terms of the syntactic structure of a base language. The main results show that different axioms on the languages considered axiomatize various classes of causal models. Those papers axiomatize not only the basic Pearl [16] model, which is the model we axiomatize here, but also more general models that cannot be captured in our framework. However, the primitives in those models are not directly associated with objects that economists use to reason about causality. In particular, whether the Pearl model is a suitable model for causal analysis in economics is unclear from the axiomatization. By providing an axiomatic foundation of the same model based on the choice of Savage acts and policy interventions, we show that the Pearl model is indeed a suitable choice for reasoning about causality in economics.

References

[1] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

[2] P. C. Fishburn. Preference-based definitions of subjective probability. *The Annals of Mathematical Statistics*, 38(6):1605–1617, 1967.

[3] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

[4] D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[5] F. Gul. Savage's theorem with a finite number of states. *Journal of Economic Theory*, 1992.

[6] J. Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

[7] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 56(4):843–887, 2005.

[8] J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 56(4):889–911, 2005.

[9] J. Heckman and R. Pinto. Causal analysis after haavelmo. *Econometric Theory*, 31(1):115–151, 2015.

[10] M. A. Hernan and J. M. Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

[11] Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

[12] E. Karni. Subjective expected utility theory without states of the world. *Journal of Mathematical Economics*, 42(3):325–342, 2006.

[13] E. Karni. States of nature and the nature of states. *Economics & Philosophy*, 33(1):73–90, 2017.

[14] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

[15] M. J. Machina and D. Schmeidler. A more robust definition of subjective probability. *Econometrica: Journal of the Econometric Society*, pages 745–780, 1992.

[16] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[17] J. Pearl. Bayesianism and causality, or, why i am only a half-bayesian. In *Foundations of bayesianism*, pages 19–36. Springer, 2001.

[18] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[19] L. J. Savage. *The foundations of statistics*. Courier Corporation, 1972.

[20] R. Spiegler. Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics*, 131(3):1243–1290, 2016.

[21] R. Spiegler. Data monkeys: A procedural model of extrapolation from partial statistics. *The Review of Economic Studies*, 84(4):1818–1841, 2017.

[22] R. Spiegler. Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association*, 2018.

[23] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

A Proofs

Proposition 2. *Let $\bar{\succ} = (\succ_{\mathcal{J}})_{\mathcal{J}\subset\mathcal{N}}$ be a DM's preferences, and let $G(\bar{\succ}) = (\mathcal{N}, E)$ be the directed graph defined by setting $Pa(i) = Ca(i)$ for each $i \in \mathcal{I}$. If $\bar{\succ}$ satisfies Assumption 1, then the following are true:*

- *If $G = (\mathcal{N}, F)$ is a directed graph that represents $\bar{\succ}$, then $(j, i) \in F \Rightarrow j \in$*

$Ca(i)$.

- *If $G = (\mathcal{N}, F)$ is a directed graph that represents $\bar{\succ}$, then $j \in Ca(i) \Rightarrow (j,i) \in F$ or $i \in Ca(j)$.*

Proof. Let $\bar{\succ}$ be as in the statement of the proposition, $G(\bar{\succ})$ be the directed graph defined by setting $Pa(i) = Ca(i)$ for each $i \in \mathcal{N}$, and $G = (\mathcal{N}, F)$ be any other directed graph that represents $\bar{\succ}$. For each $\mathcal{I} \subset \mathcal{N}$ and each realization $x_{\mathcal{I}} \in X_{\mathcal{I}}$, let $\mu_{x_{\mathcal{I}}} \in \Delta(X_{\mathcal{I}^{\complement}})$ represent beliefs obtained from $\succ_{x_{\mathcal{I}}}$.
We first show $j \in Ca(i) \Rightarrow (j,i) \in F$ or $i \in Ca(j)$. If $j \in Ca(i)$ then the function $T : X_j \to \mathbb{R}$ defined as $T(x_j) = \mu_{x_{\{i,j\}^{\complement}}, x_j}(x_i)$ is not constant in x_j. Also, by Assumption 1', $\mu_{x_{\{i,j\}^{\complement}}}(x_i|x_j) = T(x_j)$. Thus, i and j are not independent after intervening $\{i,j\}^{\complement}$. Because G represents $\bar{\succ}$ then $G_{\{i,j\}^{\complement}}$ represents $\succ_{\{i,j\}^{\complement}}$. Thus, either $(i,j) \in F$ or $(j,i) \in F$ (if not, $G_{\{i,j\}^{\complement}}$ would treat i and j as independent, which is a contradiction). If $(j,i) \in F$ the proof concludes. Therefore, let $(j,i) \notin F$ so that $(i,j) \in F$. Because G represents $\bar{\succ}$ this means that $\mu_{x_{\{i,j\}^{\complement}}}(x_j|x_i) = \mu_{x_{\{j\}^{\complement}}}(x_j)$. By definition, the above equation says $i \in Ca(j)$, as desired.
We now show $(j,i) \in F \Rightarrow j \in Ca(i)$. First, note that for all $x \in X$, $\mu_{x_{\{i,j\}^{\complement}}}(x_i, x_j) = \mu_{x_{\{i,j\}^{\complement}}}(x_j)\mu_{x_{\{i,j\}^{\complement}}}(x_i|x_j)$. Because G represents $\bar{\succ}$, $(j,i) \in F$ and the minimality condition in Definition 5, jointly imply that i and j are not independent after intervening $\{i,j\}^{\complement}$. That is, $\mu_{x_{\{i,j\}^{\complement}}}(x_i|x_j)$ is not constant in x_j. Moreover, because G represents $\bar{\succ}$ and $(j,i) \in F$, we get that $\mu_{x_{\{i,j\}^{\complement}}}(x_i|x_j) = \mu_{x_{\{i,j\}^{\complement}, x_j}}(x_i)$. Therefore, there is a value of $x_{\{j\}^{\complement}}$ for which $T(x_j) = \mu_{x_{\{i,j\}^{\complement}, x_j}}(x_i)$ is not constant in x_j. Therefore, $j \in Ca(i)$. □

Remark 1. *Without axiom 2, any representing graph must include the causal links in the sense of Definition 2 (*i.e., *$(j,i) \in F \Rightarrow j \in Ca(i)$) but F could omit some arrows. However, only arrows involved in 2-cycles are omitted.*

Before proving Theorem 1, we need two Lemmas. Let i be a variable and $\mathcal{I}, \mathcal{J}$ be two disjoint set of variables that do not contain i. It is known from Dawid ([1]) and Pearl ([16]) that i is independent to $\mathcal{I}$ conditional on $\mathcal{J}$ if, and only if, $\mathcal{J}$ D-separates $\{i\}$ from $\mathcal{I}$ (see below for a definition of D-separation). The next two lemmas prove that, for each variable i, $Ca(i)$ D-separates $\{i\}$ from all sets $\mathcal{J}$ that

satisfy $\mathcal{J} \subset ND(i)$, where $ND(i)$ is the set of non-descendants of i. Furthermore, $Ca(i)$ is the smallest set that has this property.

Definition 11. *Let $\mathcal{I}, \mathcal{J}, \mathcal{K} \subset \mathcal{N}$ be three disjoint set of variables. We say $\mathcal{K}$ D- separates $\mathcal{I}$ from $\mathcal{J}$ if for each undirected path between a variable in $\mathcal{I}$ and a variable in $\mathcal{J}$, one of the following properties holds:*

- *There is a node w along the path such that w is a collider (that is, there are nodes w_0, w_1 in the path such that $w_0 \to w \leftarrow w_1$), and such that $w \notin \mathcal{K}$ and $\mathcal{K} \subset ND(w)$.*
- *There is a node w along the path such that w is not a collider, and such that $w \in \mathcal{K}$.*

Lemma 1. *Fix $\mathcal{K} \subset \mathcal{N}$ and $x_{\mathcal{K}} \in X_{\mathcal{K}}$. Let $G_{\mathcal{K}}$ represent $\succ_{x_{\mathcal{K}}}$. For each $i \in \mathcal{N}$, $Ca(i) \backslash \mathcal{K}$ D-separates $\{i\}$ from $ND(i) \backslash \mathcal{K} \equiv \{\hat{j} \in \mathcal{K}^{\complement} : \ i$ is not an indirect cause of $\hat{j}$. $\}$.*

Proof. Pick $j \in \{\hat{j} \in \mathcal{K}^{\complement} : \ i$ is not an indirect cause of $\hat{j}$. $\}$ Pick an undirected trail t from j to i. That is, $t = (i_0, ..., i_N)$ where $i_0 = j$, i_N=i, and, for each $n \in \{1, ..., N\}$, either $(i_{n-1}, i_n) \in E$ or $(i_n, i_{n-1}) \in E$. First, since i is not an indirect cause of j, then t cannot be a directed path from i to j. That is, t cannot be such that $(i_n, i_{n-1}) \in E$ for each n. Second, if t is a directed path from j to i (that is, $(i_{n-1}, i_n) \in E$ for each n), then t is blocked by $i_{N-1} \in Ca(i) \backslash \mathcal{K}$. Third, assume that t is not directed in any direction. Then, t has colliders and/or tail-to-tail nodes. Let i_n be the last node that is either a collider or a tail-to-tail node. Let $q = (i_n, ..., i_N)$ be the trail starting at i_n. By definition of i_n, q must be directed. Assume that q is directed from i_n to i. Then, i_n is tail-to tail. Then, t, is blocked by i_{N-1}. Finally, assume that q is directed from i to i_n. Then, i_n is a collider. If $i_n \in Ca(i) \backslash \mathcal{K}$ then (i_n, i, q) is a cycle. Thus, $i_n \notin Ca(i) \backslash \mathcal{K}$. By a similar argument, no descendants of i_n can be in $Ca(i) \backslash \mathcal{K}$. Therefore, i_n blocks t. Since each trail joining j to i is blocked, this concludes the proof. □

Lemma 2. *Fix $\mathcal{K} \subset \mathcal{N}$, $x_{\mathcal{K}} \in X_{\mathcal{K}}$, and $i \in \mathcal{K}^{\complement}$. Let $G_{\mathcal{K}}$ represent $\succ_{x_{\mathcal{K}}}$. If $\mathcal{T} \subset \mathcal{K}^{\complement}$ satisfies that $\mathcal{T}$ D-separates $\{i\}$ from $ND(i)$, then $Ca(i) \backslash \mathcal{K} \subset \mathcal{T}$*

Proof. Let $\mathcal{K}$, i, and $\mathcal{T}$ be as in the statement of the Lemma. Assume $w \in Ca(i) \backslash \mathcal{K}$.

Then, $w \in ND(i)$ because otherwise $G_{\mathcal{K}}$ would not be acyclic. Consider the path $w \to i$. Then, $\mathcal{T}$ can D-separate this path only if $w \in \mathcal{T}$. Thus, $Ca(i) \backslash \mathcal{K} \subset \mathcal{T}$. $\square$

Theorem 1. *Let $\succsim$ satisfy Assumption 1. The following are equivalent:*

- *Axioms 1 and 4 hold,*
- *$(\exists G)$ such that G is a DAG, and represents $\succsim$.*

Furthermore, if G represents $\succsim$, then $G = G(\succsim)$.

Proof. The uniqueness claim is proved in Proposition 2.

We now prove that the axioms imply the existence of a representation. Without loss of generality, label the variables so that $i < j$ implies $j \in ND(i)$. Construct G by setting $Pa(i) = Ca(i)$. By axiom 1, G is acyclic. Indeed, if for some length $k \in \mathbb{N}$ there was a cycle $e = ((i_1, i_2), (i_2, i_3), ..., (i_k, i_1))$, then i_1 would be an indirect cause of itself. Pick any set $\mathcal{K} \subset \mathcal{N}$ and any realization $x_{\mathcal{K}} \in X_{\mathcal{K}}$. Let $K = \#\mathcal{K}$. We need to show that $\mu_{x_{\mathcal{K}}}(x_{\mathcal{K}^\complement}) = \Pi_{i \in \mathcal{K}^\complement} \mu_{x_{\mathcal{K}}}(x_i | Ca(i) \cap \mathcal{K}^\complement)$. By our enumeration, $\{j \notin \mathcal{K} : j < i\} \subset \{j \in \mathcal{N} : i \text{ is not an indirect cause of } j\}$. Let $\mathcal{I} = Ca(i)$, $\mathcal{J} = \{j \notin \mathcal{K} : j < i \text{ and } j \notin Ca(i)\}$, and pick $j \in \mathcal{J}$. Axiom 3 implies that $\mu_{x_{\mathcal{K}}}(x_i | x_{(Ca(i) \cup Ca(j) \cup \{j\}) \cap \mathcal{K}^\complement}) = \mu_{x_{\mathcal{K}}}(x_i | x_{(Ca(i) \cup Ca(j)) \cap \mathcal{K}^\complement})$. Axiom 4 implies that $\mu_{x_{\mathcal{K}}}(x_i | x_{(Ca(i) \cup Ca(j)) \cap \mathcal{K}^\complement}) = \mu_{x_{\mathcal{K}}}(x_i | x_{(Ca(i)) \cap \mathcal{K}^\complement})$. By the intersection property of conditional probability this implies that $\mu_{x_{\mathcal{K}}}(x_i | x_{(Ca(i) \cup \{j\}) \cap \mathcal{K}^\complement}) = \mu_{x_{\mathcal{K}}}(x_i | x_{Ca(i) \cap \mathcal{K}^\complement})$. By the chain rule, we know $\mu_{x_{\mathcal{K}}}(x_{\mathcal{K}^\complement}) = \Pi_{i=1, i \notin \mathcal{K}}^{N} \mu_{x_{\mathcal{K}}}(x_i | \{j \notin \mathcal{K} : j < i\})$. Combining the last two claims, $\mu_{x_{\mathcal{K}}}(x_{\mathcal{K}^\complement}) = \Pi_{i=1, i \notin \mathcal{K}}^{N} \mu_{x_{\mathcal{K}}}(x_i | Ca(i) \cap \mathcal{K}^\complement)$, which is what we wanted to prove. We now prove minimality of $Ca(i)$. Assume $D \subsetneq Ca(i)$. Then there is $j \in Ca(i) \backslash D$. Because G is acyclic, $j \in ND(i)$. Because $j \in Ca(i)$, axiom 2 states that for all sets H, $i \not\perp j | H$. Hence, there exists $j \in ND(i)$ such that $i \not\perp j | D$, completing the proof.

Now, suppose G is a DAG that represents $\succsim$. By our uniqueness claim, without loss of generality G is such that $Pa(i) = Ca(i)$. By contrapositive, that G is acyclic implies Axiom 1 holds. If Axiom 1 did not hold, there exists i and there exists a sequence $(i, i_1, ..., i_T, i)$ such that $i \in Ca(i_1)$, for all $t \in \{1, ..., T-1\}$, $i_t \in Ca(i_{t+1})$, and $i_T \in Ca(i)$. Thus, $((i, i_1), ..., (i_{t-1}, i_t), ..., (i_T, i))$ is a cycle in G. Axiom 2 holds

by definition of representation. Indeed, if $i \to j$, then this constitutes a path that can never be blocked. Thus, $i \not\perp_{\mathcal{K}} j|\mathcal{H}$ for all variables i, j and all disjoint sets $\mathcal{H}, \mathcal{K}$, with $i, j \notin \mathcal{K}, \mathcal{H}$. To see that Axiom 3 holds, notice that $Ca(i) \cup Ca(j) \cap \mathcal{K}^{\complement}$ block all paths from i to j in $G_{\mathcal{K}}$. Indeed, assume p is an undirected path from i to j that is not blocked, and enumerate $p = (i, i_0, ..., i_T, j)$. Because p is not blocked, $i_0 \notin Ca(i)$ and $i_T \notin Ca(j)$. Indeed, if either $i_0 \in Ca(i)$ or $i_T \in Ca(j)$, then either i_0 is not a collider or i_T is not a collider, thus implying that p is blocked by $Ca(i) \cup Ca(j)$. Therefore, p has a collider. Let n be the smallest number such that i_n is a collider and m be the largest number such that i_m is a collider (possibly $n = m$). Note that, because G is acyclic, $i_n \notin Ca(i)$ and $i_m \notin Ca(j)$. Because of this, and because p is not blocked, the following must be true:

- (i) $i_n \in Ca(j)$,
- (ii) $i_m \in Ca(i)$.

Then, the directed path that goes from i to i_n, jumps to j, comes back to i_m, and skips back to i, is a cycle. [3] This constitutes a contradiction. Thus, there every path p from i to j is blocked by $Ca(i) \cup Ca(j)$. Thus, axiom 3 holds. Similarly, $Ca(i) \cup \{j\} \cap \mathcal{K}^{\complement}$ blocks all paths from i to $Ca(j) \cap \mathcal{K}^{\complement}$ in $G_{\mathcal{K}}$. Indeed, let p be an undirected path from i to $Ca(j)$, and assume p is not blocked by $Ca(i) \cup \{j\}$. Enumerate $p = (i, i_0, ..., i_T, k)$ where $k \in Ca(j)$. Because $j \in ND(i)$ then p cannot be directed from i to k. If $i_0 \in Ca(i)$ then i_0 blocks p. If $i_0 \notin Ca(i)$, since p is not directed, p has a collider. Let n be the smallest number such that i_n is a collider. First, note that $i_n \notin Ca(i)$ because this would constitute a cycle. Second, if $i_n = j$ then j would be a descendant of i, a contradiction. Thus, $i_n \notin Ca(i) \cup \{j\}$, and hence p is blocked. Thus, axiom 4 holds. □

Theorem 2. *Let $\succsim$ satisfy Assumption 1, and let $(\mu_{x_{\mathcal{I}}})_{\mathcal{I} \subset \mathcal{N}}$ be the subjective beliefs elicited from $\succsim$. The following are equivalent:*

- *Axioms 1 through 5 hold,*
- *$\exists$ a Markov representation of μ, $(G, (h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$, such that*

[3] Formally, this is the path $q = (i, i_0, ...i_n, j, i_T, i_{T-1}, ..., i_m, i)$

– $(\forall \mathcal{J} \subset \mathcal{N})$, $(\forall x_{\mathcal{J}} \in X_{\mathcal{J}})$; $\mu_{x_{\mathcal{J}}} = \mu(\cdot|do(x_{\mathcal{J}})) \in \Delta(X_{\mathcal{J}^{\complement}})$,

– *G represents $\bar{\succ}$.*

Furthermore, if G represents $\bar{\succ}$, then $G = G(\bar{\succ})$.

Proof. The uniqueness claim was proven in 2.
We first show the axioms imply the representation. By Theorem 1, Axioms 1 and 2 imply there exists a DAG G such that G represents $\bar{\succ}$. For each $i \in \mathcal{N}$ let $Pa(i)$ be the set of parents of i in G. Note $Pa(i) = Ca(i)$ by the uniqueness claim. For each $i \in \mathcal{N}$, let $\varepsilon_i \sim U[0,1]$. For each realization $x_i \in X_i$ and each $x_{Pa(i)} \in X_{Pa(i)}$, let $I(x_i, x_{Pa(i)}) \subset [0,1]$ be an interval of length $\mu_{x_{Pa(i)}}(x_i)$. Because $\sum_{x_i \in X_i} \mu_{x_{Pa(i)}}(x_i) = 1$ for each $x_{Pa(i)}$, then $I(\cdot, x_{Pa(i)})$ can be chosen to form a partition of $[0,1]$. Fix any variable $i \in \mathcal{N}$, let $h_i(x_{Pa(i)}, \varepsilon_i) = \sum_{x_i \in X_i} x_i \mathbb{1}_{I(x_i, x_{Pa(i)})}(\varepsilon_i)$. By construction, $(G, (h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$ is a Markov representation of the beliefs elicited from $\bar{\succ}$. Pick any $\mathcal{J} \subset \mathcal{N}$ and any $i \in \mathcal{J}^{\complement}$. By Axiom 5, for each $x_i \in X_i$, and each $x_{Ca(i) \cup \mathcal{J}} \in X_{Ca(i) \cup \mathcal{J}}$, we obtain

$$\mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i) \setminus \mathcal{J}}) = \mu(x_i | x_{Ca(i)}). \tag{10}$$

Our Markov representation implies

$$\begin{aligned} \mu(x_i | x_{Ca(i)}) &= \phi(\{\varepsilon : h_i(x_{Ca(i)}, \varepsilon_i) = x_i\}) \\ &= \mu(x_i | do(x_{\mathcal{J}}), x_{Ca(i) \setminus \mathcal{J}}). \end{aligned} \tag{11}$$

By 10 and 11, $\mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i) \setminus \mathcal{J}}) = \mu(x_i | do(x_{\mathcal{J}}), x_{Ca(i) \setminus \mathcal{J}})$. Because G represents $\bar{\succ}$, for each $x \in X$,

$$\begin{aligned} \mu_{x_{\mathcal{J}}}(x_{\mathcal{J}^{\complement}}) &= \Pi_{i=1, i \notin \mathcal{J}}^{N} \mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i) \setminus \mathcal{J}}) \\ &= \Pi_{i=1, i \notin \mathcal{J}}^{N} \mu(x_i | do(x_{\mathcal{J}}), x_{Ca(i) \setminus \mathcal{J}}) = \mu(x_{\mathcal{J}^{\complement}} | do(x_{\mathcal{J}})). \end{aligned}$$

Thus, $\mu_{x_{\mathcal{J}}}(\cdot) = \mu(\cdot | do(x_{\mathcal{J}})) \in \Delta(X_{\mathcal{J}^{\complement}})$.
We now show the representation implies the axioms. If there exists a DAG G that represents $\bar{\succ}$ then axioms 1 and 2 hold is proven in Theorem 1. Let $i \in \mathcal{N}$, $\mathcal{J} \subset \{i\}^{\complement}$, $f, g \in \mathbb{R}^{X_i}$, $x_{\mathcal{J}} \in X_{\mathcal{J}}$ and $x_{Ca(i) \setminus \mathcal{J}} \in X_{Ca(i) \setminus \mathcal{J}}$ be arbitrarily selected. We know form

the Markov representation that for each $x_i \in X_i$, $\mu^i_{x_{\mathcal{J}}}(x_i|x_{Ca(i)\setminus\mathcal{J}}) = \mu^i(x_i|x_{Ca(i)})$, where μ^i denotes the marginal of μ on X_i. Thus, Axiom 5 holds. □

Proposition 1 is a direct consequence of Theorem 2 and Theorem 3, which is stated and proven below.

Theorem 3. *Let $\bar{\mu} = \{\mu_p : p \in \mathcal{P}\}$ be a collection of intervention beliefs, and let G be a DAG that represent $\bar{\mu}$. If equations 7 and 8 hold, then Axiom 5 holds.*

Proof. Let $\bar{\mu}$ and G be as in the theorem. Let $i \in \mathcal{N}$ and $\mathcal{J} \subset \{i\}^{\complement}$. We want to show that $\mu(x_i|x_{Ca(i)}) = \mu_{x_{\mathcal{J}}}(x_i|x_{Ca(i)\setminus\mathcal{J}})$. Let $\mathcal{J}^* \equiv \mathcal{J} \cap Ca(i)$; that is, $\mathcal{J}^*$ are those variables in $\mathcal{J}$ that are direct causes of i. Thus, we need to show that $\mu(x_i|x_{Ca(i)}) = \mu_{x_{\mathcal{J}}}(x_i|x_{Ca(i)\setminus\mathcal{J}^*})$; we do this in two steps.

First we show that $\mu(x_i|x_{Ca(i)}) = \mu_{x_{\mathcal{J}^*}}(x_i|x_{Ca(i)\setminus\mathcal{J}^*})$. To so this, notice that $Ca(i)\setminus\mathcal{J}^*$ blocks any path from $\{i\}$ to $\mathcal{J}^*$ in the graph $G_{(\mathcal{J}\setminus\mathcal{J}^*)^{in},(\mathcal{J}^*)^{out}}$. Indeed, let p be any path from i to some $j \in \mathcal{J}^*$ in graph $G_{(\mathcal{J}\setminus\mathcal{J}^*)^{in},(\mathcal{J}^*)^{out}}$. Write $p = (i_0, ..., i_T)$ where $i_0 = i$ and $i_T = j$. Because $j \in Ca(i)$, then p cannot be a directed path from i to j, or else G would have a cycle. Likewise, p cannot be a directed path from j to i since $G_{(\mathcal{J}\setminus\mathcal{J}^*)^{in},(\mathcal{J}^*)^{out}}$ has no arrows emerging from j. Therefore, p has a collider or a tail-to-tail node. Let w be the first node that is either a collider or a tail-to-tail node. First, assume w is tail-to-tail. Then p is of the form $i \leftarrow i_1(...) \leftarrow w \rightarrow (...)j$. Then $i_1 \in Ca(i)\setminus\mathcal{J}^*$: indeed, $i_1 \in Ca(i)$ and $i_1 \notin \mathcal{J}^*$ (since there are no arrows emerging from nodes in $\mathcal{J}^*$). Furthermore, i_1 is not a collider. Then, i_1 blocks p. Now, assume w is a collider rather than tail-to-tail. Then p is of the form $i \rightarrow i_1(...) \rightarrow w \leftarrow (...)j$. Then, w is a descendant of i, so neither w nor any w descendant is in $Ca(i)$. A fortori, neither w nor any descendant of w is in $Ca(i)\setminus\mathcal{J}^*$. Thus, w blocks p. Therefore, by formula 7 we have $\mu(x_i|x_{Ca(i)}) = \mu_{x_{\mathcal{J}^*}}(x_i|x_{Ca(i)\setminus\mathcal{J}^*})$.

Second, we show $\mu_{x_{\mathcal{J}^*}}(x_i|x_{Ca(i)\setminus\mathcal{J}^*}) = \mu_{x_{\mathcal{J}^*\cup(\mathcal{J}\setminus\mathcal{J}^*)}}(x_i|x_{Ca(i)\setminus\mathcal{J}^*})$. This is because $Ca(i)$ blocks all paths between i and $\mathcal{J}\setminus\mathcal{J}^*$ in graph $G_{\mathcal{J}\setminus\mathcal{J}^*(Ca(i)\setminus\mathcal{J}^*)^{in}}$. To see this, notice that if $\mathcal{J}\setminus\mathcal{J}^*$ contains only non-descendants of i, then the result is a direct consequence of lemma 1. Let p be a path (not necessarily directed) between i and $j \in \mathcal{J}\setminus\mathcal{J}^*$. By contradiction, assume that $j \in \mathcal{J}\setminus\mathcal{J}^*$ is a descendant of i. Then, $j \notin Ca(i)$ and j is not an ancestor of any node in $Ca(i)$. Therefore,

$j \in \mathcal{J} \backslash \mathcal{J}^*(Ca(i) \backslash \mathcal{J}^*)$, so there are no arrows into j. Therefore, no path from i to j can be directed in any direction, so there is at least one collider or tail-to-tail node. Let w be the first such node, and assume w is a collider. Then, p is of the form $i \to (...) \to w \leftarrow (...) \leftarrow j$. Then, neither w nor any descendant of w can be in $Ca(i)$, so p is blocked by $Ca(i)$. Alternatively, say w is a tail-to-tail node. Then, p is of the form $i \leftarrow i_1(...) \leftarrow w \to (...) \leftarrow j$ (with possibly $w = i_1$). Then, $i_1 \in Ca(i)$ and i_1 is not a collider. Thus, $Ca(i) = \mathcal{J}^* \cup (Ca(i) \backslash \mathcal{J}^*)$ blocks p. Thus, by formula 8, $\mu_{x_{\mathcal{J}^*}}(x_i | x_{Ca(i)\backslash \mathcal{J}^*}) = \mu_{x_{\mathcal{J}^* \cup (\mathcal{J} \backslash \mathcal{J}^*)}}(x_i | x_{Ca(i)\backslash \mathcal{J}^*}) = \mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i)\backslash \mathcal{J}})$.

Combining this with the first step, we conclude $\mu(x_i | x_{Ca(i)}) = \mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i)\backslash \mathcal{J}})$ as desired.

□

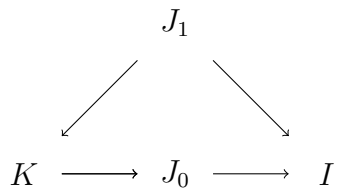
B The rules of causal calculus

In this appendix we give some intuition behind why the notion of a block i relevant for analyzing conditional independence. Furthermore, we give intuition as to why the truncations in Figure 13a are the relevant truncations for identifying intervention beliefs. We begin by reminding the reader of the definition of a block.

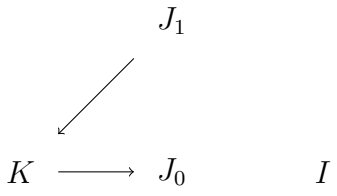
Definition 12. *Let $\mathcal{I}, \mathcal{J}, \mathcal{K}$ be three disjoint sets of variables, and let p be any path (not necessarily directed) between a node in $\mathcal{I}$ and a node in $\mathcal{J}$. We say $\mathcal{K}$* blocks *p if there is a node K on p such that one of the following conditions holds:*

- *K has converging arrows along p, and neither K nor any of its descendants is in $\mathcal{K}$, or*
- *K does not have converging arrows in p, and K is in $\mathcal{K}$.*

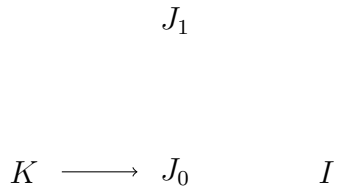
To illustrate the notion of a block, see Figure 11, replicated below for convenience. In that case, the singleton $\{K\}$ blocks all paths from J_1 to J_0. Indeed, one such path is $J_1 \to K \to J_0$. This path is blocked by $\{K\}$ because (i) the path has no converging arrows at K, and (ii) $K \in \{K\}$. The other path from J_1 to J_0 is $J_1 \to I \leftarrow J_0$. This path is blocked by $\{K\}$ because I is a node along the path such that there are converging arrows at I, but neither I nor any of its descendants are in $\{K\}$.



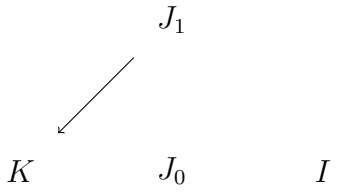
(a) A base DAG, G.



(b) DAG $G_{\mathcal{I}^{in}}$ obtained by eliminating all arrows into I.



(c) DAG $G_{\mathcal{I}^{in},\mathcal{J}^{out}}$ obtained by:
(i) eliminating arrows into I, and (ii) all arrows emerging from J_0 and J_1.



(d) DAG $G_{\mathcal{I}^{in},\mathcal{J}(\mathcal{K})^{in}}$ obtained by:
(i) eliminating all arrows into $\mathcal{I}$, and (ii) then eliminating all arrows into J_0 since J_0 is the only $\mathcal{J}$ node which is not an ancestor of a $\mathcal{K}$ node.

Figure 13: Different truncations of a *DAG*.

The notion of a block is a graphical depiction of conditional independence. Indeed, that a path exists between two sets of variables, $\mathcal{I}$ and $\mathcal{J}$, implies $\mathcal{I}$ and $\mathcal{J}$ are (a priori) statistically dependent: any variable w present in a path from $\mathcal{I}$ to $\mathcal{J}$ may potentially act as a correlating device between $\mathcal{I}$ and $\mathcal{J}$.

In particular, the position of a variable w in a path between $\mathcal{I}$ and $\mathcal{J}$ is relevant to the way in which w correlates these variables. Say that there is a path $i \to w \leftarrow j$, where $i \in \mathcal{I}$ and $j \in \mathcal{J}$; *i.e.* there is a path joining $\mathcal{I}$ and $\mathcal{J}$ that has converging arrows at w. This implies that observations of w (and its descendants) are informative about i and j simultaneously. However, interventions of w are useless for the purposes of predicting the value of either i or j, since neither w nor any of its descendants are a cause of either i nor j. By contrast, if there is a path of the from $i \to w \to j$ or $i \leftarrow w \to j$ (*i.e.* a path with non-converging arrows) then we know that both observations *and* interventions of w are useful for predicting the values of i and j, though in different ways. In the case where

$i \leftarrow w \rightarrow j$, observing or intervening w provides the same joint information about i and j, since w is a common direct cause of j and i. However, if $i \rightarrow w \rightarrow j$, intervening w provides information about j (since w is a direct cause of j) but provides no information about i (since w is neither a direct nor an indirect a cause of i). In this case, intervening w breaks down the statistical dependence of i and j in a way that is different to simply conditioning on observations of w. This sparks a natural question: can the structure of the graph tell us something about the conditional independence properties of the underlying conditional and do-probability distributions? This is the object of study in Dawid ([1]), Geiger, Pearl, and Verma ([4]), Lauritzen et. al ([14]) and others. The rules of causal calculus are a particular way in which the structure of the graph is informative about intervention beliefs.

C Two examples of do-probability

Example 4. *Consider a set $\mathcal{N} = \{1, 2, 3\}$, and a distribution $p \in \Delta(X_1 \times X_2 \times X_3)$. Suppose p has the following Markov Representation:*

$$\begin{aligned} Pa(1) &= \varnothing, \quad h_1(\varepsilon_1) = \varepsilon_1, \\ Pa(2) &= \{1\}, \quad h_2(x_1, \varepsilon_2) = x_1 + \varepsilon_2, \\ Pa(3) &= \{1\}, \quad h_3(x_1, x_2, \varepsilon_3) = x_1 - \varepsilon_3. \end{aligned}$$

Then, p can be represented as follows:

$$\begin{aligned} p(x_1) &= \phi(\{\varepsilon : \varepsilon_1 = x_1\}), \\ p(x_2) &= \phi(\{\varepsilon : x_1 + \varepsilon_2 = x_2\}) = \phi(\{\varepsilon : \varepsilon_1 + \varepsilon_2 = x_2\}), \\ p(x_3) &= \phi(\{\varepsilon : x_1 - \varepsilon_3 = x_3\}) = \phi(\{\varepsilon : \varepsilon_1 - \varepsilon_3 = x_3\}). \end{aligned}$$

Therefore, we can calculate $p(x_3|do(x_2))$, and $p(x_3|x_2)$ as follows:

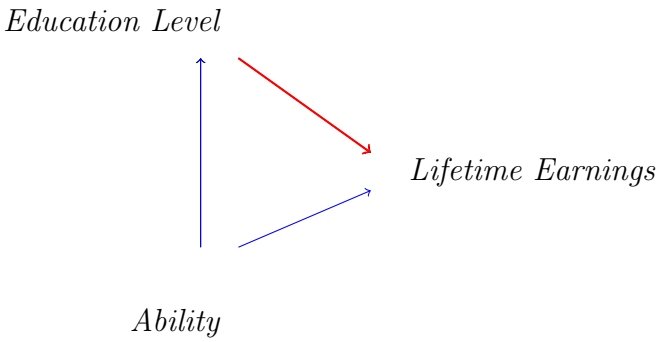
$$p(x_3|do(x_2)) = \phi(\{\varepsilon : \varepsilon_1 - \varepsilon_3 = x_3\}), \tag{12}$$

$$p(x_3|x_2) = \phi(\{\varepsilon : \varepsilon_1 - \varepsilon_3 = x_3\} | \{\varepsilon : \varepsilon_1 + \varepsilon_2 = x_2\}). \tag{13}$$

In 12, the equation determining the value x_2 is eliminated from the Markov representation. This makes the value x_2 uninformative about the value of ε_1. In (13), we recognize that variable 2 depends on ε_1, so the value x_2 gives information about the value of ε_1. Therefore, the do-probability in (12) is independent of the value x_2, whereas the conditional probability in (13) does depend on x_2. That $p(x_2|do(x_2))$ is a constant function (when viewed as a function of x_2) is intended to reflect that variable 2 is not a cause of variable 3. That x_2 does affect $p(x_3|x_2)$ captures that there is a correlation between these two variables (in this example, mediated by variable 1). The difference between these two calculations highlights the difference between causation and correlation.

Example 5 below illustrates how to use do-probabilities to identify causal effects in terms of conditional probabilities only. By connecting intervention beliefs to do-probabilities, Theorem 2 effectively provides all tools for identifying causal effects from conditional probabilities. For more detail on this see section 7.1.

Example 5. *Assume a DM's preferences can be represented by the DAG below.*



If this DAG represents a probability distribution that admits a Markov representation then there exist functions h_A, h_E, h_L such that the following holds:

$$\begin{aligned} p(L=l|E=e) &= Pr(\{\varepsilon : h_L(h_A(\varepsilon_A), h_E(h_A(\varepsilon_A), \varepsilon_E), \varepsilon_L) = l\} | h_E(h_A(\varepsilon_A), \varepsilon_E) = e), \\ p(L=l|do(E=e)) &= p(\{\varepsilon : h_L(h_A(\varepsilon_A), \boldsymbol{e}, \varepsilon_L) = l\}). \end{aligned}$$

Suppose we are interested in quantifying the direct effect that education has on earnings (graphically represented by the red arrow). However, as the graph shows, E provides information about L in two ways. The first, is the direct effect (in-

dicated by the red path). The second is through the effect that A has on both E and L: observing the value of E provides information about A, and A provides direct information on L (as indicated by the blue path). In the first equation, which corresponds to a conditional probability, we explicitly see that h_L depends on A through h_E. In the second line, we eliminate the equation determining education, and instead directly impute a value of $E = e$. In this way we block the dependence of L on A via E, and only the red effect remains.

As far as quantifying this effect, algebraically manipulating the equations above yields the following:

$$\begin{aligned} p(L = l|do(E = e)) &= \sum_a p(A = a, L = l|do(E = e)), \\ &= \sum_a p(A = a|do(E = e))p(L = l|do(E = e), A = a), \\ &= \sum_a p(A = a)p(L = l|A = a, E = e), \qquad (14) \\ &\neq p(L = l|E = e), \end{aligned}$$

Therefore, if we wish to elicit the direct effect that E has on L, all we need data on $p(A = a)$, $p(L = l|A = a, E = e)$, and to apply equation (15). Notice also that the above equation array can be replicated in terms of intervention beliefs and Axiom 5:

$$\begin{aligned} p_e(L = l) &= \sum_a p_e(A = a, L = l), \\ &= \sum_a p_e(A = a)p_e(L = l|A = a), \\ &= \sum_a p(A = a)p(L = l|A = a, E = e), \qquad (15) \end{aligned}$$

where the last line applies after applying Axiom 5 noting that (i) the marginal of A is the same regardless of whether we intervene E or not because $Ca(A) = \varnothing$ and (ii) since $Ca(L) = \{A, E\}$, then conditioning on A and E or conditioning on A and intervening E yield the same marginal over L.