Causality: a decision theoretic foundation*

Pablo Schenone†
Arizona State University

This version: February 21, 2019.
First version: September 1, 2017

Abstract

We propose a decision theoretic model akin to that of Savage [14] 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 [12]).

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

*I wish to thank David Ahn and Jeff Ely 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 [14] 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 (who 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 axiomatic exercise requires that the DM/researcher can produce beliefs/probability distributions to feed into the numerical model, but how those distributions are obtained is an orthogonal question. The objective of this paper is to provide a theoretical foundation for the selection of numerical models to use in empirical studies of causality in economics.

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 2 describes the model, and Section 3 formally defines causality. Second, we conduct a standard decision theoretic analysis: we provide axioms on preferences and show that the axioms hold if, and only if, we can numerically represent the DM's beliefs. Since we are interested in causal effects, the conditional independence properties of the DM's joint beliefs are of special interest. A useful tool to represent the conditional independence properties of a probability law is the language of Bayesian graphs (see Section 4.1 for the axioms and Section 4.2 for the representation theorem). Thus, our representation has a graph theoretic component. Specifically, Proposition 1 and Theorem 1 show that our definition of causal effect, which, as argued before, is tailored to how economists use the term causality, admits a unique representation in terms of a Bayesian network. Finally, we provide an additional axiom under which causal effects can

be directly calculated from standard conditional probabilities in a decision problem with no intervention policies to make the model useful for empirical applications. When this is the case, we say that causal effects are *identified*. Proposition 2 and Theorem 2 in Section 5 prove the identification results.

As the reader may anticipate, the statistics, computer science, and economics literature addressing causal effects is extensive. The related literature is discussed in Section 6, 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 Model and notation

2.1 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.

2.2 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.

To understand the role of intervention policies in our model, consider the following example.

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".*

These are clearly different statements that do not imply one another. Therefore, we need language to distinguish these two distinct decision problems. Intervention policies provide such language.

The discussion above implies our DM's primitive choice domain must be ex-

panded to include intervention policies. To do so, 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.

Intervention preferences look like Savage conditional preferences but have important differences. Savage conditional preferences operate as follows. First, a DM selects an act of the form $\mathbb{1}_{x_\mathcal{J}} f$, where $x_\mathcal{J}$ is some realization of the variables in $\mathcal{J}$ and f is some Savage act. Then, the DM goes about life as usual while waiting for forces outside of his control (say, "Nature") to select a value for the state. When a state $s \in X$ is realized, if $s_\mathcal{J} = x_\mathcal{J}$, the DM is payed $f(s)$; otherwise, the DM is payed 0. Thus, Savage conditional preferences refer to such statements as in item [1.] of Example 1. In an intervention preference, the DM chooses a policy p and an act f over $\mathcal{N}(p)$. That is, Nature is brushed aside, and the DM actively *imposes* that variables in $\mathcal{N}^\complement(p)$ take the value $p_{\mathcal{N}^\complement(p)}$. After choosing p and f, the DM goes about life as usual and waits for Nature to select the value of the non-intervened variables, $\mathcal{N}(p)$. Thus, intervention preferences refer to such statements as in item [2.] of Example 1. Understanding the relation between Savage conditional preferences and intervention preferences is the core of our analysis of causal effects.

3 Definition of causal effect

We now introduce the definition of causal effect, which formalizes the intuitive definition given in Section 1. Conceptually, we say that a variable j *causes* a variable i when policies that *intervene* j at different levels have a *ceteris paribus* effect on the DM's beliefs about i. If the way a DM chooses acts over i is insensitive to whether j was intervened, we say j does not cause i. As mentioned in Section 1, this is precisely how "causation" is understood in economics: as any correlation which, subjectively for the analyst, is sufficient grounds for making policy recommendations.

The definition informally described above emphasizes two terms: "intervene" and "ceteris paribus". As discussed in Section 2, interventions must distinguish relations that are purely correlations (as decided by Nature) from other types of relations. The term ceteris paribus — meaning that variables other than i and j are intervened to a constant level — guarantees that we are defining direct causal effects rather than indirect effects.

Figure 1 illustrates the importance of using ceteris paribus interventions when defining causal effects. The figure captures 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. If $a, a' \in A$ are two ability levels and f, $g \in \mathbb{R}^{L\times E}$ are two affects on lifetime earnings that are constant in E, we might have the DM behave as follows: $f \succ_a g$ and $g \succ_{a'} f$. This behavior is selected because intervening A at different levels has an impact on E, and E has an impact on L. However, this result captures the indirect effect of A an L, whereas we are interested in the *direct* effect of A on L. The correct way to capture this effect is to look at intervention preferences $\succ_{(a,e)}$ as a function of a, for each fixed $e \in E$. This is the purpose of the term ceteris paribus in our informal description of causal effect. To overcome the naivety presented above, we define 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 mediated 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

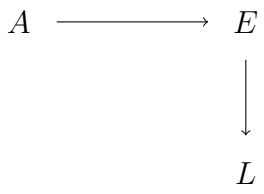


Figure 1: The naive definition of causal effect would have us wrongly believe that A causes E.

of j if the following holds: $(\forall x_{\mathcal{K}} \in X_{\mathcal{K}})$, $(\forall x_j, x'_j \in X_j)$, and $(\forall f, g \in \mathbb{R}^{X_i})$,

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

The first line indicates that having intervened $\mathcal{K}$, 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,

$$f \succ_{x_j, x_{\mathcal{K}}} g \Leftrightarrow \mathbb{1}_{X_j} f \succ_{x_{\mathcal{K}}} \mathbb{1}_{X_j} g \Leftrightarrow f \succ_{x'_j, x_{\mathcal{K}}} g,$$

so the first equation also holds. For future reference, we record this result in the following definition.

Definition 1. *For all $i, j \in \mathcal{N}$ and $\mathcal{K} \subset \{i, j\}^{\complement}$, we say variable i is* $\mathcal{K}$-independent *of variable j if for all $f, g \in \mathbb{R}^{X_i}$,*

$$f \succ_{x_j, x_{\mathcal{K}}} g \Leftrightarrow \mathbb{1}_{X_j} f \succ_{x_{\mathcal{K}}} \mathbb{1}_{X_j} g,$$

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 $Ca(i) = \{j \in \mathcal{N} : j$ 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 $j_T = i$.

We make two observations regarding Definition 2. First, the definition refers to direct (or unmediated) effects: j causes i if, *all other variables held constant*, interventions on j affect the DMs choice of acts in $\mathbb{R}^{X_i}$. This definition does not rule out the possibility that interventions on i *without keeping other variables constant* might also affect choices in $\mathbb{R}^{X_i}$ (see Figure 1). However, we do not call such an effect a causal effect since it might be an effect that i has on j *mediated* through other variables. Second, 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 2.

Finally, we define the causal graph associated with a preference, $\bar{\succ}$. Causal graphs are an integral part of our representation, which is introduced in Section 4.2. 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 $E = \{(j, i) : j \in Ca(i)\}$.*

4 Axioms and representation

Our main theorem is stated below. Section 4.1 states and discusses the set of axioms. Section 4.2 discusses why our notion for when a DAG represents a preference relation is reasonable.

Theorem 1. *Let $\bar{\succ}$ satisfy Axiom 1, and let μ_p be the intervention beliefs elicited from $\succ_p$. The following statements are equivalent:*

i Axioms 2 and 3 hold,

ii $(\exists G)$ such that G is a DAG, and G represents $\bar{\succ}$ in the following sense:

$(\forall x \in \Pi_{i\in\mathcal{N}} X_i)$, *and all* $\mathcal{J} \subset \mathcal{N}$

- $\mu_{x_{\mathcal{J}}}(x_{\mathcal{J}^{\complement}}) = \Pi_{i=1}^{N} \mu_{x_{\mathcal{J}}}(x_i | x_{ND(i)\setminus\mathcal{J}}) = \Pi_{i=1}^{N} \mu_{x_{\mathcal{J}}}(x_i | x_{Pa(i)\setminus\mathcal{J}})$
- $(\forall i)$ $Pa(i)$ *is the smallest set such that the above property holds*
- *If* $(i,j) \in E$ *then* $\mu_{x_{\{i,j\}^{\complement}}}(x_j | x_i) = \mu_{x_{\{j\}^{\complement}}}(x_j)$.

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

4.1 Axioms

We consider three basic axioms on $\bar{\succ}$. Axiom 1 is standard and states that for each policy p, the DM's preferences induced by p are probabilistically sophisticated. Axioms 2 and 3 restrict how policies affect the choice of Savage acts over non-intervened variables; these restrictions convey the qualitative properties of causal effects.

Axiom 1 establishes the following three properties: (i) our DM is a subjective expected utility maximizer, (ii) all relevant variables are included in the state space, and (iii) all states have positive probability. We include these properties as part of a basic axiom since they are more primitive issues than causation. Before introducing Axiom 1, we recall Gul's [4] axioms (Appendix D contains a formal statement of the Axioms). The reader is referred to the original paper for a normative discussion of these axioms.

Axiom (Gul '95). *Let* T *be a finite set and* $\succ$ *a binary relation on* $\mathbb{R}^T$, *with weak part* $\succsim$ *and symmetric part* $\sim$.

G1. $\succsim$ *is complete and transitive.*

G2. Independence axiom: $(\forall f, g, h \in \mathbb{R}^T)$, $(\forall t \in T)$, $f \succ g \Leftrightarrow f' \succ g'$, *where* f' *and* g' *are appropriately constructed mixtures of* f *with* h *and* g *with* h, *respectively. (See appendix for details on the mixing operation.)*

G3. $\succsim$ *is monotone increasing. If* f *and* f' *are constant acts and* f *pays more than* f', *then* $f \succ f'$. *Furthermore, the state space can be partitioned into two equally likely events; that is, there exists an event* $A \subset T$ *such that, for any two values* $x, y \in \mathbb{R}$, $xAy \sim yAx$ *(where* xAy *is the act that pays* x *in* A

and y elsewhere, and yAx is analogously defined).

G4. $\succsim$ *is continuous (upper and lower contour sets are closed).*

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

i- *For each $p \in \mathcal{P}$, the preferences induced by $\succ_p$ satisfy the Gul [4] axioms; that is, for each $\mathcal{J} \subset \mathcal{N}$, $x_{\mathcal{J}} \in X_{\mathcal{J}}$, the preference $\succ_{x_{\mathcal{J}}}$ on $\mathbb{R}^{X_{\mathcal{J}^c}}$ satisfies the Gul axioms when setting $T = X_{\mathcal{J}^c}$.*

ii- *$(\forall i, j \in \mathcal{N})$, $(\forall x_{\{i\}^c} \in X_{\{i\}^c})$, and $(\forall f, g \in \mathbb{R}^{X_i})$, if $j \in Ca(i)$, then $f \succ_{x_{\{i\}^c}} g \Leftrightarrow \mathbb{1}_{x_j} f \succ_{x_{\{i,j\}^c}} \mathbb{1}_{x_j} g$.*

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

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

Item $[i.]$ in Axiom 1 is standard and implies that agents are subjective expected utility maximizers in all choice problems. Specifically, for each possible intervention, agents are subjective expected utility maximizers in the associated choice problem. Formally, if Axiom 1 holds, then for each $\mathcal{J} \subset \mathcal{N}$ and for each element $x_{\mathcal{J}} \in X_{\mathcal{J}}$, there exists an increasing and continuous function $u_{x_{\mathcal{J}}} : \mathbb{R} \to \mathbb{R}$ and a probability distribution $\mu_{x_{\mathcal{J}}}$ over $X_{\mathcal{J}^c}$ such that

$$f \succ_{x_{\mathcal{J}}} g \Leftrightarrow \sum_{x_{\mathcal{J}^c}} u_{x_{\mathcal{J}}}(f(x_{\mathcal{J}^c}))\mu_{x_{\mathcal{J}}}(x_{\mathcal{J}^c}) > \sum_{x_{\mathcal{J}^c}} u_{x_{\mathcal{J}}}(g(x_{\mathcal{J}^c}))\mu_{x_{\mathcal{J}}}(x_{\mathcal{J}^c}).$$

Furthermore, each $\mu_{x_{\mathcal{J}}}$ is unique, and each $u_{x_{\mathcal{J}}}$ is unique up to positive affine transformations. Specifically, G3 implies that all Bernoulli utilities represent the same ordinal preference (namely, more money is better), but they might differ in cardinal ways, such as risk aversion coefficients. However, the central objects of our study are the probability distributions, not the Bernoulli indexes. Thus, item $[iv.]$ rules out the possibility that policies have a direct impact on the Bernoulli utility indexes.

Note that Axiom 1 implies that all intervention preferences are probabilistically sophisticated. Henceforth, we refer to the probability distributions induced by intervention preferences as *intervention beliefs*.

Item [*ii.*] in Axiom 1 says that the state space is complete in the following sense: given two variables (say, i and j), the state space includes all variables that could mediate effects between i and j. Indeed, given i and j, two possibilities exist: either one of the variables affects the other directly (say, j is a cause of i) or the variables are unrelated. 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 [14], we assume that the state space is complete.

Axiom 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 2 through 4 provide such discipline.

Axiom 2. *For all $i \in \mathcal{N}$, i is not an indirect cause of i.*

Axiom 2 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 explain 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 2 states that the DM's state space includes only logically independent variables.

Axiom 2 could be seen as precluding the DM from viewing the world as a system of recursive structural equations. As such, Axiom 2 could be seen as precluding the DM from reasoning in terms of equilibrium equations (see, for example, the critique in Heckman and Pinto [8]). 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 de-

fined 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 equilibrium 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.

Another 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.

To present Axiom 3, we need some notation. Let $ND(i)$ denote the set of variables that are not indirectly caused by some variable i. We use this notation since this set will correspond to the set of i's non-descendants in our representation.

$$ND(i) = \{j \in \{i\}^{\complement} : i \text{ is not an indirect cause of } j\}$$

With this notation in mind, we present Axiom 3.

Axiom 3. $(\forall i \in \mathcal{N})$ *and* $(\forall f, g \in \mathbb{R}^{X_i})$*, a set* $\mathcal{I} \subset \{i\}^{\complement}$ *satisfies the condition below if, and only if,* $Ca(i) \subset \mathcal{I}$*.*

$(\forall \mathcal{J} \subset ND(i) \cap \mathcal{I}^{\complement}),\ (\forall \mathcal{K} \subset \{i\}^{\complement},\ \mathcal{J} \cap \mathcal{K} = \varnothing),\ and\ (\forall x_{\mathcal{I}\cup\mathcal{J}\cup\mathcal{K}} \in X_{\mathcal{I}\cup\mathcal{J}\cup\mathcal{K}}),$

$$\mathbb{1}_{x_{\mathcal{J}}}\mathbb{1}_{x_{\mathcal{I}\setminus\mathcal{K}}} f \succ_{x_{\mathcal{K}}} \mathbb{1}_{x_{\mathcal{J}}}\mathbb{1}_{x_{\mathcal{I}\setminus\mathcal{K}}} g \Leftrightarrow \mathbb{1}_{x_{\mathcal{I}\setminus\mathcal{K}}} f \succ_{x_{\mathcal{K}}} \mathbb{1}_{x_{\mathcal{I}\setminus\mathcal{K}}} g$$

To understand Axiom 3, consider the sources of information that a DM finds useful for the purpose of choosing acts over a variable i. Clearly, the causes of i should provide information about i; by implication, indirect causes of i should also provide information about i. However, the information about i encoded in the indirect causes of i should already be contained in the causes of i. Therefore, conditional on the causes of i, indirect causes of i provide no additional useful information about i. Furthermore, the variables of which i is an indirect cause also provide information about i: by definition, such variables indirectly depend on i, so knowing their value provides useful information about i. Axiom 3 formalizes these ideas. It states that only two fundamental sources of information about i exist: its causes and the variables that are indirectly caused by i. Specifically, Axiom 3 implies that a variable is never statistically independent of its causes. Both of these properties are normative properties that a set of causes should have.

To understand why Axiom 3 captures the discussion presented above, fix a variable i and consider three separate cases: when $\mathcal{K} = \varnothing$ and $\mathcal{I} = Ca(i)$, when $\mathcal{K} \neq \varnothing$ and $\mathcal{I} = Ca(i)$, and when $\mathcal{I} \subsetneq Ca(i)$.

First, let $\mathcal{K} = \varnothing$ and $\mathcal{I} = Ca(i)$. Axiom 3 states that the DM treats the following two problems in the same way. The first problem is to choose between two acts over i (say, f and g), whose payments are contingent on the causes of i obtaining a certain value, $x_{Ca(i)}$. The second problem is to choose between the same acts f and g when the payments are contingent on the causes of i obtaining the same value $x_{Ca(i)}$, as well as variables in $\mathcal{J} \subset ND(i)\backslash Ca(i)$ obtaining some value, $x_{\mathcal{J}}$. In the example illustrated in Figure 2 below, let $\mathcal{I} = Ca(i) = \{w, j\}$ and let $\mathcal{J} = \{k\}$. Thus, the first problem is to choose between $\mathbb{1}_{x_{w,j}} f$ and $\mathbb{1}_{x_{w,j}} g$, and the second problem is to choose between $\mathbb{1}_{x_{w,j,k}} f$ and $\mathbb{1}_{x_{w,j,k}} g$. Our DM treats these two problems in the same way because when the payments of f and g are contingent on the values of $\{w, j\}$, the only variables that could affect his beliefs over X_i are variables that are indirectly caused by i. Since no variable in $\mathcal{J}$ is indirectly caused by i (in Figure 2, $\mathcal{J} = \{k\}$), none of these variables provides

useful information about i that is not already included in $\{w, j\}$. Hence, the two decision problems are equivalent.

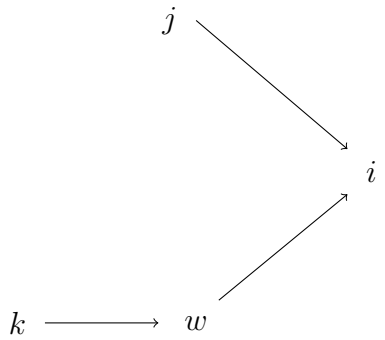


Figure 2: If i is independent of $ND(i)\backslash\mathcal{I}$, then $\mathcal{I}$ must contain $\{w, j\}$.

When $\mathcal{K} \neq \varnothing$ and $\mathcal{I} = Ca(i)$, the logic in the above paragraph remains valid even if we intervene the variables in $\mathcal{K}$ and work with the preference induced by this intervention. Since the intervention $\mathcal{K}$ might include variables in $\mathcal{I}$, we choose only contingent acts on the remaining variables, $\mathcal{I}\backslash\mathcal{K}$. For instance, in terms of Figure 2, let $\mathcal{J} = \{k\}$, $\mathcal{K} = \{w\}$, and $\mathcal{I} = Ca(i) = \{w, j\}$. Axiom 3 states that $\mathbb{1}_{x_j} f \succ_{x_w} \mathbb{1}_{x_j} g$ if, and only if, $\mathbb{1}_{x_{j,k}} f \succ_{x_w} \mathbb{1}_{x_{j,k}} g$: once w is intervened, k is uninformative about i since the only information k provides about i is that which is mediated through w.

Finally, suppose that $\mathcal{I} \subsetneq Ca(i)$. Since $Ca(i)\backslash\mathcal{I} \neq \varnothing$ and no variable in $Ca(i)\backslash\mathcal{I}$ is indirectly caused by i, we can choose $\mathcal{J} = Ca(i)\backslash\mathcal{I}$. In the first decision problem, the payments of f and g are contingent on $\mathcal{J} = Ca(i)\backslash\mathcal{I}$ and $\mathcal{I}$; therefore, the payments of f and g are contingent on the full set of causes of i. In the second problem, the payments of f and g are contingent on only $\mathcal{I}$; hence, they are not contingent on $Ca(i)\backslash\mathcal{I}$. In terms of Figure 2, let $\mathcal{I} = \{w\} \subsetneq \{w, j\}$, $\mathcal{J} = \{j\} \subset ND(i) \cap \mathcal{I}^{\mathbb{C}}$ and $\mathcal{K} = \varnothing$. The first decision problem is to choose between $\mathbb{1}_{x_{w,j}} f$ and $\mathbb{1}_{x_{w,j}} g$, whereas the second problem is to choose between $\mathbb{1}_{x_w} f$ and $\mathbb{1}_{x_w} f$. Since the variables in $Ca(i)\backslash\mathcal{I}$ ($\{j\}$ in the example) *do* provide useful information about i, these two problems should *not* be equivalent. Thus, the equivalence established in Axiom 3 should hold when $Ca(i) \subset \mathcal{I}$ but not when $\mathcal{I} \subsetneq Ca(i)$.

While Axioms 1 through 3 are our basic axioms, Axiom 4 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.

Axiom 4. $(\forall i \in \mathcal{N})$, $(\forall \mathcal{J} \subset \{i\}^{\complement})$ $(\forall f, g \in \mathbb{R}^{X_i})$, $(\forall x_{Ca(i)\cup\mathcal{J}} \in X_{Ca(i)\cup\mathcal{J}})$,

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

Similar to Axiom 3, Axiom 4 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 $x_{\mathcal{J}^{\complement}}$ being realized.

However, contrary to Axiom 3, Axiom 4 imposes restrictions *across* different intervention preferences. Indeed, for any fixed variable i and intervention $x_{\mathcal{K}}$, Axiom 3 provides conditions under which the choice of acts over i is insensitive to the observation of variables in $\mathcal{J}$. Therefore, Axiom 3 addresses the statistical independence properties of a *fixed* intervention belief but is silent regarding the connection between the intervention preference $\succ_{x_{\mathcal{K}}}$ and the standard Savage preference $\succ$. Axiom 4 considers the connection between any particular intervention preference, $\succ_{x_{\mathcal{K}}}$, and the conditional Savage preference, $\succ_{\cdot|x_{\mathcal{K}}}$. Thus, contrary to Axiom 3, Axiom 4 imposes restrictions *across* different intervention preferences.

We use Figure 3 below to illustrate why Axiom 4 is normatively appealing.

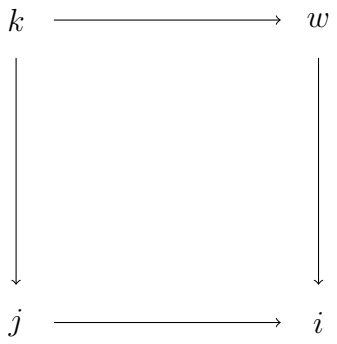


Figure 3: Observing or intervening j makes the DM update differently about k. This difference in updating may affect the DM's beliefs about i.

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 $\mathbb{1}_{x_j} g$. 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.

Now consider the situation above but where the payments of f and g are contingent on the values of both j *and* w. That is, for some x_j and x_w, the DM must choose 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? Note that 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 4 requires.

To complete our discussion of Axiom 4, we must consider the case where $\mathcal{J}$ contains non-causes of i. Axiom 4 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 3, 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.

Finally, Axiom 4 imposes an important restriction that interventions have no structural impact on the DM's probabilistic model; that is, only the value of the causes of a variable matter, not whether those values arose by means of an observation or an intervention. In the context of Figure 3, what matters for choosing acts in $\mathbb{R}^{X_i}$ are the numerical values obtained by j and w. Axiom 4 implies that whether the payments of an act $f \in \mathbb{R}^{X_i}$ are contingent on the values of j and w or if the DM intervened the values of j and w is in itself irrelevant. Situations where

interventions have an intrinsic effect are incompatible with Axiom 4.

4.2 Representation

We now define our representation of $\succsim$. Since $\succsim$ is associated with a collection of probability distributions, we first define what it means for a DAG to represent a family of probability distributions (see Section 2.1 for a reminder of our graph theoretic notation.)

Lauritzen et al. [10] provide a definition for when a DAG represents a probability distribution. Let $p \in \Delta(\Pi_{i\in\mathcal{N}} X_i)$, and let $G = (\{1, ..., N\}, E)$ be a DAG. The chain rule implies that $(\forall x \in \Pi_{i\in\mathcal{N}} X_i)$, $p(x) = \Pi_{i=1}^{N} p(x_i|ND(i))$. However, the only non-descendants that provide direct information about i are its parents; thus, we have the following definition.

Definition 4. *Let $p \in \Delta(\Pi_{i\in\mathcal{N}} X_i)$. A DAG $(\{1, ..., N\}, E)$ represents p if, and only if, the following hold:*
$(\forall x \in \Pi_{i\in\mathcal{N}} X_i)$,

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

Definition 4 makes two statements. First, a DAG represents a probability distribution if, and only if, the DAG summarizes the conditional independence properties of p; namely, that conditional on its parents, a variable is independent of its non-descendants. Second, the set of parents is the smallest set that allows for such a decomposition. Indeed, consider a set of nodes $V = \{1, 2, 3\}$ and a probability distribution $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$. Since all variables are statistically independent, both DAGs in Figure 4 represent this p. Indeed, both $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$ and $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$ are true statements. However, the first representation includes irrelevant arrows: the minimality requirement prevents this.

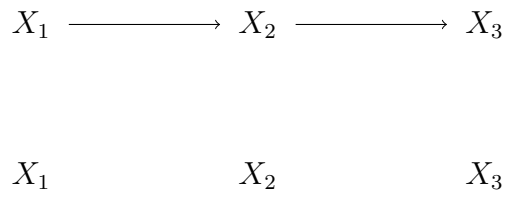


Figure 4: Both DAGs above represent the same probability distribution, $p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3)$, but the top one includes irrelevant arrows.

Note that $\bar{\succ}$ is associated with a collection of beliefs, one for each induced $\succ_p$, 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. We do so by defining 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 uninformative about the value of their causes. 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.

Definition 5. *Let $\bar{\succ}$ satisfy Axiom 1, and let $G = (\mathcal{N}, E)$ be a DAG. For each $\mathcal{T} \subset \mathcal{N}$ and each $x_{\mathcal{T}} \in X_{\mathcal{T}}$, let $\mu_{x_{\mathcal{T}}}$ be the probability distribution over $X_{\mathcal{T}^{\complement}}$ that represents the DM's beliefs after $X_{\mathcal{T}}$ is intervened to the value $x_{\mathcal{T}}$. We say that G* represents *$\bar{\succ}$ if the following are true for each $\mathcal{K} \subset \mathcal{N}$ and each $x \in X$:*

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

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

Definition 5 has two components. First, once we eliminate all intervened variables from the graphical description of the DM's probabilistic model, then the remaining graph represents (in the sense of Definition 4) the intervention beliefs over the remaining variables. However, that a graph represents a probability distribution means that the graph summarizes the conditional independence properties of that probability distribution. Since independence is a symmetric relation, this information is insufficient to identify the direction of causality.[1] Indeed, if j and i are two variables that are not $\{i,j\}^{\complement}$-independent, then either i causes j or vice versa; however, our model is yet unable to identify which is the cause and which is the consequence. The second item in Definition 5 addresses the identification of the direction of causality. A crucial difference between cause and consequence is that *observing* the value of a consequence is informative about the value of the cause, whereas *intervening* the value of a consequence is uninformative about the value of the cause. Therefore, if i is the cause and j is the consequence, then intervening i or conditioning on i should lead to the same (ceteris paribus) beliefs on j. That is, if $i \to j$, then $\mu_{x_{\{i,j\}^{\complement}}}(x_j|x_i) = \mu_{x_{\{j\}^{\complement}}}(x_j)$, which is precisely what Definition 5 requires.

Proposition 1 below is our first result.

Proposition 1. *Let $\bar{\succ}$ 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 Axiom 1, then the following are true:*

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

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

We say that a graph $G = (\mathcal{N}, E)$ has no 2-cycles if, for each pair $i, j \in \mathcal{N}$, $(i,j) \in E \Rightarrow (j,i) \notin E$; that is, no two variables can mutually point at each other. Note that Axiom 2 implies $G(\bar{\succ})$ has no 2-cycles. We then have the following corollary.

Corollary 1. *Under the assumptions of Proposition 1, if $G(\bar{\succ})$ has no 2-cycles,*

[1] For a discussion of how the lack of uniqueness is addressed in the basic Pearl model, see Pearl [19].

then $G(\bar{\succ}) = G$. *Specifically, if Axiom 2 holds, then* $G = G(\bar{\succ})$.

Proposition 1 makes two assertions. First, any representing graph includes all causal arrows, in the sense that Definition 2 assigns to the expression "causal arrow". Second, a representing graph $G = (\mathcal{N}, F)$ might fail to include some causal links; however, this situation only arises when two variables mutually cause each other. As in our discussion of Axiom 2, if $i \to j$ and $j \to i$, then i and j are not logically independent variables, so either i or j should not have been included in the state space to begin with.

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

i Axioms 2 and 3 hold,

ii $(\exists G)$ *such that* G *is a DAG and represents* $\bar{\succ}$.

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

Note that for any collection of probability laws $\{\mu_{x_\mathcal{J}} \Delta(X_{\mathcal{N}\setminus\mathcal{J}}) : \mathcal{J} \subset \mathcal{N}, x_\mathcal{J} \in X_\mathcal{J}\}$, there exists a preference $\bar{\succ}$ such that $\mu_{x_\mathcal{J}}$ is the intervention belief associated with $\succ_{x_\mathcal{J}}$.[2] From the uniqueness claim, we have the following Corollary.

Corollary 2. *If a DAG* G *represents* $\{\mu_{x_\mathcal{J}} \Delta(X_{\mathcal{N}\setminus\mathcal{J}}) : \mathcal{J} \subset \mathcal{N}, x_\mathcal{J} \in X_\mathcal{J}\}$, *then there exists* $\bar{\succ}$ *such that* $G = G(\bar{\succ})$.

The immediate contribution of Theorem 1 is to provide a foundation for representing a DM's causal model. In this way, Theorem 1 shows that causality can be seamlessly incorporated into the standard Savage framework in a way that is consistent with how economists understand the term "causality".

For practical applications, the first implication of Theorem 1 comes from Corollary 2. Consider an empirical researcher who models a causal structure by means of a set of probabilities $\{\mu_p : p \in \mathcal{P}\}$ and a DAG that represents said probabilities. The analyst would like to interpret each μ_p as an intervention probability and the arrows in the DAG as representing a reasonable definition of causality. Corollary 2 states that there must be a preference $\bar{\succ}$ such that $G = G(\bar{\succ})$ and each μ_p is the intervention belief induced by policy $p \in \mathcal{P}$. Thus, the researcher's model of causal

[2] For any $\mathcal{J} \subset \mathcal{N}$, any $x_\mathcal{J}$, and any $f, g \in \mathbb{R}^{X_{\mathcal{J}^\complement}}$, define $\succ_{x_\mathcal{J}}$ as $f \succ_{x_\mathcal{J}} g$ iff $\sum_{x_{\mathcal{J}^\complement}} f(x_{\mathcal{J}^\complement}) \mu_{x_\mathcal{J}}(x_{\mathcal{J}^\complement}) > \sum_{x_{\mathcal{J}^\complement}} g(x_{\mathcal{J}^\complement}) \mu_{x_\mathcal{J}}(x_{\mathcal{J}^\complement})$.

effects captures "causality" as defined in Definition 2 and not some other notion of "causality". This result is valuable since Definition 2 was tailored for use in economics (for a discussion of Definition 2, see Section 1 and 3). Theorem 1 states that causality models based on DAGs identify a unique and useful definition of causal effect, which is the theorem's first contribution.

Theorem 1 also provides a foundation for unifying and structuring our understanding of causation. The theorem states that any formal discussion of causality must begin 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, as understood by Definition 2. In this way, Theorem 1 provides a foundation for selecting among models with which to empirically study causal effects.

Finally, as discussed in Section 1, the purpose of this axiomatic exercise is to provide a link between numerical methods (in this case, DAG representations of a family of probability distributions) and the way a rational DM approaches the issue of interest (in this case, causation). The goal is to guarantee that the numerical method treats causality the way a perfectly rational DM would. In the context of empirical research, the DM is the empirical researcher and the DM's beliefs are the probability laws that the researcher feeds into the numerical model. Typically, those probabilities are calculated from empirical data using a statistical method that the DM/researcher deems acceptable. Data regarding intervention probabilities are generally unavailable outside of experimental settings where different policies $p \in \mathcal{P}$ can be implemented. Therefore, while Theorem 1 provides a foundation for qualitatively representing a causal structure, it does not provide a foundation for methods with which to quantify causal effects.

The focus of Section 5 is to provide foundations for methods with which to quantify causal effects. Theorem 2 in Section 5 provides an extra axiom under which, for various policies p, μ_p can be expressed purely as a function of the non-intervention probability, μ. In such cases, we say μ_p is *identified*. Since the data required to calculate non-intervention probabilities, μ, is generally available and since identification provides a way to express intervention probabilities, μ_p, in terms of μ,

identification provides a way to calculate μ_p using non-intervention data.

5 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 2 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. Specifically, if the DM is an analyst, the model allows the DM to use data from outside a controlled setting to calculate intervention beliefs, which is paramount to how empirical causal analysis is conducted.

When added to Axioms 1 through 3, Axiom 4 yields a model in which different intervention beliefs, μ_p, can be expressed in terms of μ. In the following, we remind the reader of Axiom 4 and illustrate Theorem 2 by means of three simple examples. Then, we state and discuss the general form of Theorem 2.

Axiom 4. $(\forall i \in \mathcal{N})$, $(\forall \mathcal{J} \subset \{i\}^{\complement})$ $(\forall f, g \in \mathbb{R}^{X_i})$, $(\forall x_{Ca(i)\cup\mathcal{J}} \in X_{Ca(i)\cup\mathcal{J}})$,

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

Example 2. *Consider Alex, from Section 1, who studies the relation between Ability, Education, and Lifetime earnings. Alex understands causal effects as defined in Definition 2, and she acknowledges that Axioms 1 through 3 are appealing normative properties of causal models. Thus, Alex models causal effects via a DAG that represents a family of intervention beliefs. Specifically, Alex's causal mode is that ability causes both education and lifetime earnings but that education does not cause lifetime earnings. This model is graphically depicted in Figure 5. Thus, identifying the direct causal effect of education on earnings is simple: there is none.*

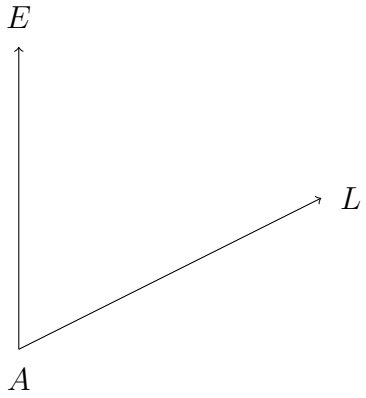


Figure 5: E has no causal effect on L.

Example 3. *Blake is a colleague of Alex who believes that ability causes both education and lifetime earnings and also that education causes lifetime earnings. This model is graphically depicted in Figure 6. To understand the 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* $\mu_{(a,e)}$ *with which to form his beliefs.Under Axiom 4, however, these data are unnecessary. By setting* $\mathcal{J} = \{A, E\}$*, Axiom 4 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*. Blake can therefore use data from outside a controlled environment to form his intervention beliefs.*

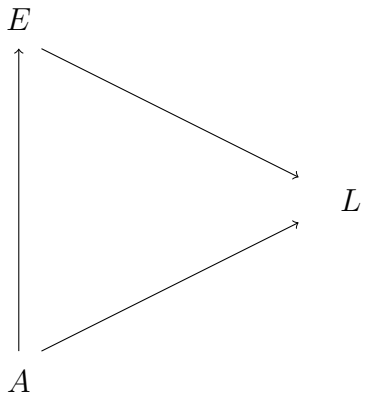


Figure 6: Causal effects are identified: $\mu_{(a,e)(l)=\mu(l|a,e)}$.

Example 4. *Charlie is another colleague of Alex. However, Charlie believes that education causes ability and that ability is the sole cause of lifetime earnings, as depicted in Figure 7. Charlie is interested in studying the* indirect *effect that education policies have on lifetime earnings, which can be done by applying Axiom 4 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* $\mu(a|e)$. *Even if access to a controlled environment is unavailable, the identification of* μ_e *implies that such data are unnecessary.*

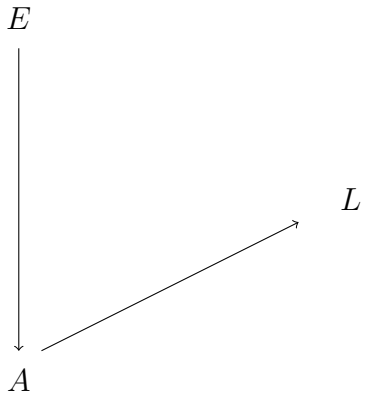


Figure 7: Indirect causal effect of E on L is identified: $\mu_e(l) = \sum_a \mu(l|a)\mu(a|e)$.

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.

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? From Axiom 4 we can deduce two formulas such that for some policy intervention $p \in \mathcal{P}$, μ_p is identified if, and only if, it is identified by iterative application of those two formulas. However, in order to state the formulas we need 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.

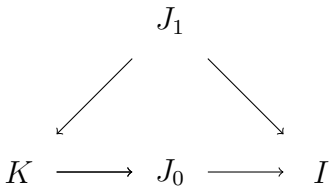
Definition 6. *Given G and three disjoint sets of variables $\mathcal{I}, \mathcal{J}, \mathcal{K} \subset \mathcal{N}$, the truncated DAGs $G_{\mathcal{I}^{in}}$, $G_{\mathcal{J}^{out}}$, and $G_{\mathcal{I}^{in},\mathcal{J}(\mathcal{K})^{out}}$ are defined as follows:*

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

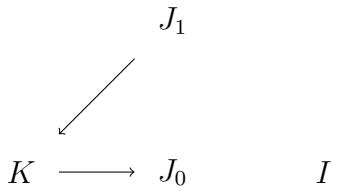
2 $G_{\mathcal{I}^{in},\mathcal{J}^{out}}$ is obtained from G by eliminating all arrows emerging from nodes in $\mathcal{J}$ and all arrows pointing to nodes in $\mathcal{I}$,

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 $G_{\mathcal{I}^{in}}$.

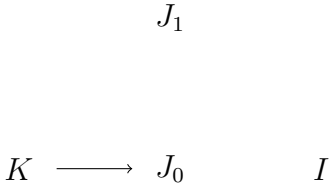
The following figures show the base DAG, G, and its corresponding truncations. In all cases, $\mathcal{J} = \{J_0, J_1\}$, $\mathcal{I} = \{I\}$, $\mathcal{K} = \{K\}$.



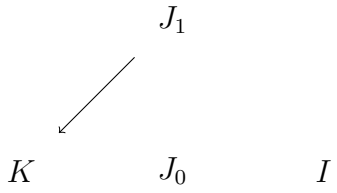
(a) The base DAG, G.



(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 8: 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 8 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 7. *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 $q \in \mathcal{K}$.*

Proposition 2 below provides two rules with which to identify intervention beliefs. While the rules themselves are known in the statistical causality literature (for a discussion, see Section 6), the proposition makes two contributions. First, the rules are known to be valid only in the context of a probabilistic model called the *do-probability* model, which we describe in Section 5.1. However, we derive the rules without reference to do-probabilities. Furthermore, it is known that the do-probability model is sufficient to guarantee that all identified intervention beliefs are identified by iterative application of Rules 1 and 2; however, whether the do-probability model is also necessary for obtaining that conclusion is unknown. The proposition states that in the context of Axioms 1 through 3, Axiom 4 is a necessary and sufficient condition to guarantee that all identified intervention beliefs are identified by iterative application of Rules 1 and 2. This implies that the do-probability model is the only model for which Rules 1 and 2 completely summarize all identification results.

Proposition 2. *Let $\bar{\succ}$ satisfy Axioms 1 through 3, let G represent $\bar{\succ}$, and let $\{\mu_p : p \in \mathcal{P}\}$ be the DM's intervention beliefs. Then, the following statements are equivalent.*

- *$\bar{\succ}$ satisfies Axiom 4.*
- *Rules 1 and 2 below hold. Furthermore, if μ_p is identified for some $p \in \mathcal{P}$, then the identification is obtained by iterative application of these two rules.*

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*

$$\mu_{x_{\mathcal{I}_1}, x_{\mathcal{I}_2}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}) \quad = \quad \mu_{x_{\mathcal{I}_1}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_2}, x_{\mathcal{I}_3}). \tag{1}$$

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*

$$\mu_{x_{\mathcal{I}_1}, x_{\mathcal{I}_2}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}) \quad = \quad \mu_{x_{\mathcal{I}_1}}(x_{\mathcal{I}_0} | x_{\mathcal{I}_3}). \tag{2}$$

With Proposition 2, we can refer to Example 4 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 $\mu_e(l) = \sum_a \mu(l|a)\mu(a|e)$.

5.1 Markov representations and do-probabilities

The results in the previous section are obtained purely from adding Axiom 4 to the list of Axioms imposed on $\succsim$. As such, they depend only on the Axioms. Moreover, Pearl [12] and Huang and Valtorta [9] show that if a DAG represents a family of *do-probabilities* (to be formally defined shortly), then any do-probability that can be identified is identified by iterative application of Rules 1 and 2. In this section, we prove that Axiom 4 is necessary and sufficient for intervention beliefs to admit a representation in terms of do-probabilities. This result establishes the link between the rules of causal calculus studied in the previous section, the do-probability model, and Axiom 4. In the following, we define do-probabilities, state Theorem 2, and discuss its implications.

Definition 8. *Let $p \in \Delta(X)$. For each $i \in \mathcal{N}$, let p_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 p 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,*
- *p can be recovered implicitly as a solution to the following system of equations:*

$$p(x_i) = \phi(\{\varepsilon : h_i(x_{Pa(i)}, \varepsilon_i) = x_i\}), \quad (i \in \{1, ..., N\}). \tag{3}$$

Markov representations are used in statistical causality to numerically represent causal effects (see Pearl [12]). 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 [12]. See examples 5 and 6 in Appendix C for a concrete illustration of how to calculate do-probabilities and how they differ from standard conditional probabilities.

Definition 9. *Let $p \in \Delta(X)$ be a probability distribution, and let $((h_1, ..., h_N), (\varepsilon_1, ..., \varepsilon_N))$ be a Markov representation of p. Given two disjoint sets of variables, $\mathcal{I}$ and $\mathcal{J}$, the do-probability $p(x_{\mathcal{I}}|do(x_{\mathcal{J}}))$ is calculated as follows:*

1. *For all $j \in \mathcal{J}$, eliminate from system (3) in Definition 8 all the formulas $p(x_j) = \phi(\{\varepsilon : h_j(x_{Pa(j)}, \varepsilon_i) = x_j\}$.*
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 (3) of Definition 8.*
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 9. 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 9: 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 Axiom 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 2, 3, and 4 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 $\succsim$.*

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

The crucial contribution of Theorem 2 is that it clarifies the role of do-probabilities in the understanding of causal effects. Do-probabilities are presented as natural representations of post-intervention beliefs; however, we show that DAGs can legitimately represent a causal model based on intervention beliefs without needing to invoke the do-probability formalism. This result is analogous to the exercise conducted by Machina-Schmeidler [11]: 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 assumption about when interventions and simple observations can be used interchangeably.

Theorem 2 further clarifies that Axiom 4 is the fundamental property that links do-probabilities with intervention beliefs. Indeed, as discussed after introducing Axiom 4, some types of causal effect may be incompatible with Axiom 4 and, thus, incompatible with do-probabilities. As such, Axiom 4 clarifies which types of causal effects can be represented via do-probabilities.

Jointly, Proposition 2 and Theorem 2 imply that Pearl's rules of causal calculus serve as an axiomatization of do-probability. Indeed, Huang and Valtorta [9] 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 4 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 8 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 [5]), under Axiom 4, this result is without loss of generality.

6 Related literature

This paper is related to two main fields of research: economics (both theory and applied) and work done in statistics and computer science.

In economic theory, the work most closely related to ours is Spiegler ([15], [16], [17]). 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. Proposition 1 provides the point of contact between both papers. If a graph G both represents a DM's correlation structure and is interpreted causally (in the sense that arrows point from cause to effect), then the definition of causal effect must be as in Definition 2.

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. [3], Lauritzen et al. [10]). Specifically, Dawid [1] and Geiger et al. [3] 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 7 and 10) 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. Lauritzen et al. [10] provide alternative graphical tests for D-separation based on the cut-sets of G.

In causal statistics, the most closely related papers are those in the Bayesian networks literature (see Spirtes [18], Pearl [12], 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 [13]: " *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 3, 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 3 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 4, 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 [2], Halpern [5], and Halpern and Pearl ([6], [7]) provide an alternative approach. Specifically, Halpern [5] expands on Galles and Pearl [2] and axiomatizes a more general model. Rather than a decision theoretic approach, Halpern [5] 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 [12] 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] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998.

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

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

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

[6] 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.

[7] 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.

[8] J. J. Heckman, S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz. A new cost-benefit and rate of return analysis for the perry preschool program: A summary. Technical report, National Bureau of Economic Research, 2010.

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

[10] 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.

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

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

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

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

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

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

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

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

[19] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence*, pages 220–227, 1991.

A Proofs

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

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

Proof. Let $\succsim$ be as in the statement of the proposition, $G(\succsim)$ 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 $\succsim$. 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 Axiom 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 4, 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 ([12]) 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 10. *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 $ND(i)\backslash\mathcal{K} \equiv \{\hat{j} \in \mathcal{K}^{\complement} : \ i \text{ is not an indirect cause of } \hat{j}. \ \}$.*

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, $Ca(i)\backslash\mathcal{K} \subset \mathcal{T}$. □

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

- *Axioms 2 and 3 hold,*
- *$(\exists G)$ such that G is a DAG, and represents $\bar{\succ}$.*

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

Proof. The uniqueness claim is proved in Proposition 1.

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 2, 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 $\mathcal{I} = Ca(i)$, $\mathcal{J} = \{j \notin \mathcal{K} : j < i \text{ and } j \notin Ca(i)\}$. By axiom 3 applied to these sets $\mathcal{K}$, $\mathcal{I}$, $\mathcal{J}$, $\mu_{x_{\mathcal{K}}}(x_i | \{j \notin \mathcal{K} : j < i\}) = \mu_{x_{\mathcal{K}}}(x_i | Ca(i) \cap \mathcal{K}^{\complement})$. 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 $Ca(i)$. Suppose $(\mathcal{T}_i)_{i \notin \mathcal{K}}$, $\mathcal{T}_i \subset \{1, ..., i-1\} \cap \mathcal{K}^{\complement}$ satisfies

$$\mu_{x_{\mathcal{K}}}(x_{\mathcal{K}^{\complement}}) = \Pi_{i \in \mathcal{K}^{\complement}} \mu_{x_{\mathcal{K}}}(x_i | \mathcal{T}_i). \tag{4}$$

We need to show that $Ca(i) \backslash \mathcal{K} \subset \mathcal{T}_i$. That $\mathcal{T}_i \subset \{1, ..., i-1\} \backslash \mathcal{K}$ satisfies 4 implies that $\mathcal{T}_i$ D-separates $\{i\}$ from $ND(i)$. By Lemma 2, $Ca(i) \backslash \mathcal{K} \subset T_i$.

The last step is to show that if $\mu_{x_{\{i,j\}^{\complement}}}(x_j | x_i) \neq \mu_{x_{\{j\}^{\complement}}}(x_j)$ then $i \nrightarrow j$ (or, equivalently, that $i \notin Ca(j)$). This follows from the contrapositive statement of Axiom 1 item $[ii]$.

Now, suppose G is a DAG that represents $\bar{\succ}$. By our uniqueness claim, without loss of generality G is such that $Pa(i) = Ca(i)$. By contrapositive, that G is acyclic implies Axiom 2 holds. If Axiom 2 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. To see that Axiom 3 holds, consider Lemmas 1 and 2. For each $\mathcal{K}$, each $i \notin \mathcal{K}$, and each $\mathcal{J} \subset \{\hat{j} \in \mathcal{K}^{\complement} : \ i \text{ is not an indirect cause of } \hat{j} \ \}$, $Ca(i) \backslash \mathcal{K}$ D-separates $\{i\}$ from $\mathcal{J}$. Furthermore, it is the smallest set with this property. Therefore, i is independent of $\mathcal{J}$ conditional on $Ca(i) \backslash \mathcal{K}$, and this is the smallest set with this property. Thus, Axiom 3 holds. □

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

- *Axioms 2, 3, and 4 hold,*

- $\exists$ *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* $G = G(\bar{\succ})$.

Proof. The uniqueness claim was proven in 1.
We first show the axioms imply the representation. By Theorem 1, Axioms 2 and 3 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 4, for each $x_i \in X_i$, and each $x_{Ca(i) \cup \mathcal{J}} \in X_{Ca(i) \cup \mathcal{J}}$, we obtain

$$\mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i) \setminus \mathcal{J}}) = \mu(x_i | x_{Ca(i)}). \tag{5}$$

Our Markov representation implies

$$\begin{aligned} \mu(x_i | x_{Ca(i)}) &= \phi(\{\varepsilon : h_i(x_{Ca(i)}, \varepsilon_i) = x_i\}) \\ &= \mu(x_i | do(x_{\mathcal{J}}), x_{Ca(i) \setminus \mathcal{J}}). \end{aligned} \tag{6}$$

By 5 and 6, $\mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i) \setminus \mathcal{J}}) = \mu(x_i | do(x_{\mathcal{J}}), x_{Ca(i) \setminus \mathcal{J}})$. 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 $\succsim$ then axioms 2 and 3 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 4 holds. □

Proposition 2 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 1 and 2 hold, then Axiom 4 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 1 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 $j \in \mathcal{J} \backslash \mathcal{J}^*$. By contradiction, assume that $j \in \mathcal{J} \backslash \mathcal{J}^*$ 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 2, $\mu_{x_{\mathcal{J}^*}}(x_i | x_{Ca(i)\backslash \mathcal{J}^*}) = \mu_{x_{\mathcal{J}^* \cup (\mathcal{J} \backslash \mathcal{J}^*)}}(x_i | x_{Ca(i)\backslash \mathcal{J}^*}) = \mu_{x_{\mathcal{J}}}(x_i | x_{Ca(i)\backslash \mathcal{J}})$.

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

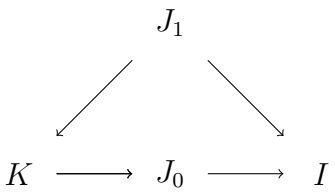
In this appendix we give some intuition behind why the notion of a block i relevant for analyzing conditional independence. Furthermore, we give intuition as to why the truncations in Figure 10a are the relevant truncations for identifying intervention beliefs. We begin by reminding the reader of the definition of a block.

Definition 11. *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:*

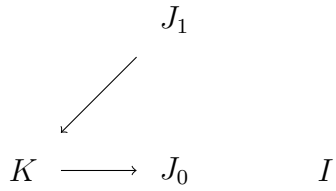
- *K has converging arrows along p, and neither K nor any of its descendants is in $\mathcal{K}$, or*
- *K does not have converging arrows in p, and K is in $\mathcal{K}$.*

To illustrate the notion of a block, see Figure 8, 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

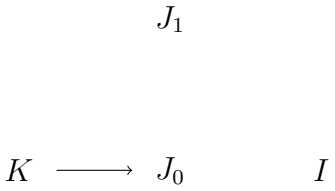
that there are converging arrows at I, but neither I nor any of its descendants are in $\{K\}$.



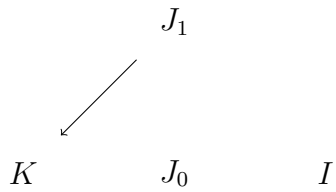
(a) A base DAG, G.



(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 10: 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 $i \leftarrow w \to j$, 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 $i \to w \to j$, 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 ([3]), Lauritzen et. al ([10]) 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 5. *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:*

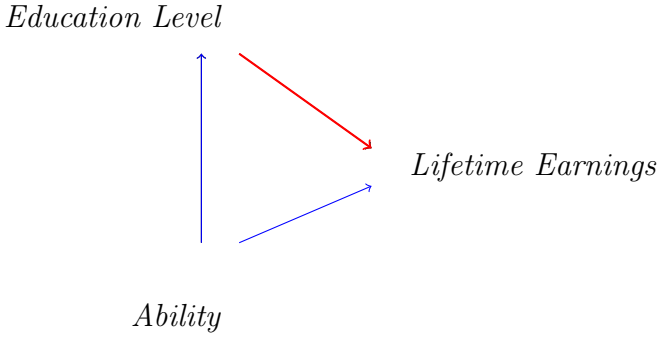
$$p(x_3|do(x_2)) = \phi(\{\varepsilon : \varepsilon_1 - \varepsilon_3 = x_3\}), \tag{7}$$

$$p(x_3|x_2) = \phi(\{\varepsilon : \varepsilon_1 - \varepsilon_3 = x_3\}|\{\varepsilon : \varepsilon_1 + \varepsilon_2 = x_2\}). \tag{8}$$

In 7, 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 (8), 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 (7) is independent of the value* x_2*, whereas the conditional probability in (8) 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 6 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 5.

Example 6. *Assume a DM's preferences can be represented by the DAG below.*



If this DAG represents a probability distribution that admits a Markov representa-

tion 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 (indicated by the red path). The second is through the effect that* A *has on both* E *and* L*: observing the valaue 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:

$$\begin{aligned} p(L = l|do(E = e)) &= \sum_a p(A = a, L = l|do(E = e)), \\ &= \sum_a p(A = a|do(E = e))p(L = l|do(E = e), A = a), \\ &= \sum_a p(A = a)p(L = l|A = a, E = e), \qquad (9) \\ &\neq p(L = l|E = e), \end{aligned}$$

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 (10). Notice also that the above equation array can be replicated in terms of intervention beliefs and Axiom*

4:

$$\begin{aligned} p_e(L=l) &= \sum_a p_e(A=a, L=l), \\ &= \sum_a p_e(A=a)p_e(L=l|A=a), \\ &= \sum_a p(A=a)p(L=l|A=a, E=e), \end{aligned} \tag{10}$$

where the last line applies after applying Axiom 4 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.

D Gul [4] axioms

In this appendix we formally define Gul's ([4]) axiomatization of subjective EU for finite state spaces.

Axiom (Gul '95). *Let T be a finite set, and $\succ$ a binary relation on $\mathbb{R}^T$, with weak part $\succsim$ and symmetric part $\sim$. For each set $Q \subset T$, let $\mathbb{1}_Q$ be the indicator function of Q (that is, $\mathbb{1}_Q(t) = 1$ if $t \in Q$, and $\mathbb{1}_Q(t) = 0$ if $t \notin Q$). The following are the axioms in Gul ([4])*

G1. $\succsim$ is complete and transitive.

G2. $(\forall f, g, h \in \mathbb{R}^T)$, $(\forall t \in T)$, and $\forall A \subset T$, construct f' and g' so that

$$\begin{aligned} \mathbb{1}_T(\cdot)f'(t) &\sim \mathbb{1}_A(\cdot)f(t) + (1 - \mathbb{1}_A(\cdot))h(t), \\ &\textit{and} \\ \mathbb{1}_T(\cdot)g'(t) &\sim \mathbb{1}_A(\cdot)g(t) + (1 - \mathbb{1}_A(\cdot))h(t). \end{aligned}$$

Then, $f \succ g \Leftrightarrow f' \succ g'$. If f', g' are impossible to construct, this item holds vacuously.

- *Note: The act $\mathbb{1}_T(\cdot)f'(t)$ is an act that pays $f'(t)$ at each state $t' \in T$, and analogously for $\mathbb{1}_T(\cdot)g'(t)$.*

– *Note: The act* $\mathbb{1}_A(\cdot)f(t) + (1 - \mathbb{1}_A(\cdot))h(t)$ *is an act that pays* $f(t)$ *at each state* $t' \in A$*, and* $h(t)$ *at each state* $t' \notin A$ *(analogously for for* $\mathbb{1}_A(\cdot)g(t) + (1 - \mathbb{1}_A(\cdot))h(t)$*).*

G3. $(\forall z, z' \in \mathbb{R})$*,* $z > z' \Leftrightarrow \mathbb{1}_T(\cdot)z \succ \mathbb{1}_T(\cdot)z'$*. Furthermore, there exists* $A \subset T$ *such that* $\mathbb{1}_A(\cdot)x + (1 - \mathbb{1}_A(\cdot))y \sim \mathbb{1}_A(\cdot)y + (1 - \mathbb{1}_A(\cdot))x$ *for all* $x, y \in \mathbb{R}$*.*

G4. $(\forall f \in \mathbb{R}^T)$*, the sets* $B(f) = \{g \in \mathbb{R}^T : g \succsim f\}$ *and* $W(f) = \{g \in \mathbb{R}^T : f \succsim g\}$ *are closed (please see the original paper for the topological definition of closed and open).*