Causal Coherence*

Andreas Duus Pape†

February 28, 2007

Abstract

Agents with the same information and same preferences can make different choices. Agents differ not only with respect to their preferences and information, but their causal interpretations of that information. This can lead to what agents with the correct causal model would perceive as "irrational mistakes" committed by others.

I apply an axiomatic representation to develop the *causally coherent* agent, who has a causal model about a causally ambiguous phenomenon that is consistent with data, makes choices rationally, but is unaware of alternative models. In essence, her model is not identified so she hazards a guess. The causal model is a causal bayesian network.

In this framework, I show how agents with the same information and the same preferences will make different choices. Moreover, with this framework, I can construct a set of reasonable theories that emerge from data the agents see. This provides a framework for constructing agents' conjectures in a general setting. I apply this framework to an auction to show that agents with wrong models suffer a 'causal curse' similar in kind to the winner's curse.

*I would like to thank Daniel Benjamin, Joshua Cherry, Peter DiCola, Michael Elsby, Robert Gazzale, David Greenstreet, Patrick Kline, Greg Lewis, Dmitry Lubensky, Yusufcan Masatlioglu, Daisuke Nakajima, Emily Shelton, Doug Smith, Charles Taragin, Seminar participants at Michigan, and especially Emre Ozdenoren, Scott Page, and John DiNardo for helpful comments.

†University of Michigan, Department of Economics

1 Introduction: Why would people suffer causal confusion?

Two potential CEOs, Sam and Quincy, are equally talented leaders and equally adept at picking successful companies on the stock market. They see the same data and they pick the same winners in the market. Sam has a chance to take over a company. His theories of what makes a company successful have all been confirmed, so he knows what choices to make. However, he's a failure. At the same time, Quincy takes over a company. Quincy's theories have also been confirmed, so he knows what choices to make: but his choices are not the same as Sam's, and Quincy is a success. Why would Quincy make different choices after seeing the same data, and why would Quincy succeed where Sam failed?

The difference between playing the stock market and starting a company is the difference between prediction and intervention. Sam and Quincy don't make the same causal inferences, and hence disagree on counterfactuals, though they are equally good at predicting the market without their intervention.

This is an example of agents with different causal models that may arise from, and be consistent with, the same data. In this case, agents with the same information and same preferences can make different choices. Agents may have the same preferences and information, but differ with respect to their causal interpretations of that information. Agents could be confused for a variety of reasons. Here I provide one: their models are not identified, and they hazard a guess. There is no missing data nor variables, and yet they draw different conclusions and make different choices. The framework of causal Bayesian networks provides us a reasonable set of theories that agents might believe given common data.

In section 2, I describe causal bayesian networks, used in artificial intelligence and statistics (Pearl 2000). In statistics, they are used by model makers to estimate causal effects. In artificial intelligence, they are used to represent an agent's mental model of a problem they face. I describe the standard framework of interventions in phenomena to then describe an agent's optimal behavior when endowed with such a model. I describe what it means for an agent to be *causally coherent* with respect to data (i.e., have a causal model consistent with the data, and act in a manner consistent with it). These agents are rational in the sense that their beliefs, actions, and data all logically cohere. They are not aware of alternative models, however—this captures the idea that people may be inductive, that is, have a theory about how something works and act in accordance with that theory until they are disabused of it. Alternatively, it can be said that they confuse evidence consistent with their model with evidence for their model.

A version of the causal model structure emerges in the utility representation provided in section 3, which, given agent's choices over interventions and bets on outcomes, allows one to construct a utility function, probability distributions, and causal structure which rationalize those choices. This representation is an application of the representation theorem in Karni (2005), which is a utility representation in the Savage style without reference to a state space. I provide an additional axiom of choice which provides for the causal Bayesian network. The version in this section provides for a case when there are two variables.

In section 4, I provide two applications of the causally coherent agent. I first show a decision problem in which agents agree on the data and have the same preferences, but make different choices. Then, I introduce causally coherent equilibria to investigate interactions of agents with different causal models. Causally coherent equilibria arise from considering agents with different causal models of the same information. Causally coherent equilibria are, in general, short-run phenomena; they arise from the different understandings of a phenomenon that can be settled when the right experiment is run. Agent behavior will sometimes implicitly run that experiment. Causally coherent equilibria are therefore appropriate for irregular events or the initial stages of a repeated game. I apply causal coherence to an auction, and find the causally coherent equilibrium, as if between Sam and Quincy above. The auction yields a result similar in kind to the winner's curse. Why? Consider how one nullifies the curse: by constructing one's opponent's information by mapping from the bid to data. Since Sam and Quincy draw different inferences from the same data, they, conversely, map the same inference to different data. In the presence of causal disagreement (and ignorance of it), Sam and Quincy cannot correct for the winner's curse; the chain is broken, and the winner will suffer a 'causal curse.'

In section 5, I discuss the cognitive science evidence for the value of such a model, the role of rationality in causal coherence, and some possible extensions to the model. The first extension would use causal models to explain apparent preference differences in a median voter setting. The second extension would construct agents who are ambiguity averse in the sense of Ellsberg (1961), who treat causal ambiguity in a manner similar to Gilboa and Schmeidler's (1989) Maxmin expected utility agents. The third would use this framework to construct agents who act in accordance with Quattrone and Tversky's (1984) empirical finding that people attribute causation to correlation.

I conclude in section 6.

2 Background in causal Bayesian networks as a decision-making framework

A causal model is a model of the internal workings of some phenomenon that the agent confronts. For example, the phenomenon could be "the firm," and the causal model could describe what causes a firm to be a success or a failure. The agent has the opportunity to observe this phenomenon, then has the opportunity to interact with that phenomenon. The agent could observe a cross-section of firms in the world, some of which were successes and some failures, and may observe other characteristics of these firms. The agent's model of the phenomenon provides, given the agent's observations about the phenomenon, a forecast for the outcome of each of her actions. From that, she decides how to optimally act. Given the correlations between these characteristics and success, the agent could conjecture a way that these characteristics cause success or failure. Given her conjecture, she can forecast what would happen if she were to change a characteristic:

for example, replace the CEO. From that, she could determine the optimal CEO for her firm.

The goal of this framework is to provide a set of reasonable causal models that the agent might consider in this setting: when she is called upon to intervene on an arbitrary phenomenon but her model is not identified in a statistical sense. When the correct model is not identified, then the set of possible models is infinite. I would like to make minimal, reasonable assumptions on the agent's cognition of the phenomenon to construct a more tractable, finite set of models. I suppose, given evidence from the cognitive science literature and intuitive appeal of the framework, that the finite set of causal Bayesian networks that are consistent with the data provides a good estimate of the set of reasonable models the agent might consider. Below, I describe causal Bayesian networks. As described by Sloman and Lagnado (2004): "A formal framework has recently been developed based on Bayesian graphical probability models to reason about causal systems (Spirtes, Glymour, and Scheines (1993); reviewed in Pearl (2000)). In this formalism, a directed graph is used to represent the causal structure of a system, with nodes corresponding to system variables, and direct links between nodes corresponding to causal relation[ships]." This presentation follows Pearl (2000). In the discussion section I discuss the cognitive science evidence regarding these structures.

Briefly, causal Bayesian networks are graphs of directed, causal relationships among variables in a phenomenon, and a mapping from the observed joint data about the phenomenon to particular relationships among variables. The set of these relationships are not unlike structural equation models, so with a brief introduction they should look familiar. The hypothesis, then, is that agents, when called upon to forecast outcomes of actions, construct models of phenomena akin to our economic models. Causal Bayesian networks provide a formal language to express these models.

The mapping from these network graphs to particular relationships, that is, moving beyond knowing X causes Y to predicting what happens to Y when X changes, is based on a few key assumptions. The first key assumption is that no variables are excluded. The agents suppose that what they see is the complete phenomenon they have to work with. The second key assumption is that relationships are acyclic: that is, that if X causes Y, then Y does not cause X. This is essentially the codification of an assumption that agents are not good at understanding feedback loops in arbitrary systems.

This section proceeds in three parts. First, the components of the causal Bayesian network are introduced and defined. Second, I describe, given a causal Bayesian network of a phenomenon, the effect of manipulations or interventions on phenomena, and hence what an agent will believe will happen for each of her available actions. One typically understands causality as being revealed under some kind of interventionist experiment(Heckman 2005) and this section describes what the outcome of such an experiment is given a causal Bayesian network. Third, I construct the set of 'reasonable' causal models from minimal assumptions about the agent's cognition of the phenomenon. These assumptions follow Spirtes, Glymour, and Scheines (1993) and Pearl (2000).

2.1 Causal Bayesian Networks

Let $\mathbb{V}$ be a set of random variables with support $\text{supp}(\mathbb{V}) = \prod_{V \in \mathbb{V}} \text{supp}(V)$. Let F be a joint distribution over $\mathbb{V}$. For the firm example, $\mathbb{V}_{firms}$ might be $\{\text{CEO skill } S, \text{firm quality } Q, \text{firm performance } P, \text{firm value } V\}$.

Define a *directed acyclic graph* as a collection of points ("nodes") and lines with arrowheads ("edges") connecting some (possibly empty) subset of the nodes, and suppose no series of arrows will lead a node to itself. (A series of arrows which lead from a node to itself would be a cycle). Figure 1 depicts a directed acyclic graph.

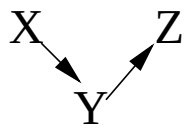


Figure 1: A causal structure

Define a *causal structure* $\boldsymbol{\mathcal{C}}$ *of* $\mathbb{V}$ to be a directed acyclic graph, in which each node corresponds to a distinct element of $\mathbb{V}$, and each link represents a direct functional relationship among the corresponding variables (Pearl 2000). Figure 1 depicts a causal structure over the variables $\{X, Y, Z\}$. if X causes Y and Y causes Z, but X does not directly cause Z. (If you believed this about X,Y, and Z, you would say that X is a good instrument.)

Define a $\boldsymbol{\mathcal{C}}$-causal parent of $V \in \mathbb{V}$ as any variable $W \in \mathbb{V}$ such that there is an arrow in $\boldsymbol{\mathcal{C}}$ which runs from W to V. The $\boldsymbol{\mathcal{C}}$-causal parents of a variable are the direct causes of that variable under structure $\boldsymbol{\mathcal{C}}$. In Figure 1, Y is a $\boldsymbol{\mathcal{C}}$-causal parent of Z, while X is not a $\boldsymbol{\mathcal{C}}$-causal parent of Z. Define $Pa_{\boldsymbol{\mathcal{C}}}(V)$ as the (possibly empty) set of all $\boldsymbol{\mathcal{C}}$-causal parents of V.[1]

Let $\Delta(V)$ be the set of all distributions over $V \in \mathbb{V}$. Let a causal probability function $\Phi_V^{\boldsymbol{\mathcal{C}}}$ be a mapping from the $\boldsymbol{\mathcal{C}}$-causal parents of V to the set of distributions over V.

$$\Phi_V^{\boldsymbol{\mathcal{C}}} : \text{supp}(Pa_{\boldsymbol{\mathcal{C}}}(V)) \rightarrow \Delta(V)$$

The causal probability function answers the following question: "Suppose the variables $Pa_{\boldsymbol{\mathcal{C}}}(V)$ achieved the values $\vec{pa}$. What distribution would they induce on V?"[2]

One example of a causal probability function is a Savage act (Savage 1954), that is, a choice over lotteries. The agent is asked to choose lotteries which deliver different distributions over money. The choice over lotteries represents a function that assigns, for each value of *Lottery*, a distribution over all possible dollar winnings.

[1] Other familial relationships can be similarly defined, namely $\boldsymbol{\mathcal{C}}$-causal child, ancestor, and descendant.

[2] I abuse notation slightly by supposing that, since $\Phi_V(Pa_{\boldsymbol{\mathcal{C}}}(V))$ assigns a distribution over V, that $\Phi_V(v|Pa_{\boldsymbol{\mathcal{C}}}(V))$ is that distribution (note the v).

Another example of a causal probability function is the classic econometric linear regression. Consider the regression $Y = \alpha + \beta X + \epsilon$, where ϵ is distributed normally with mean zero and variance σ. Suppose that regression properly captured causality. Then:

$$\Phi_Y(X = x) = Normal(\alpha + \beta x, \sigma)$$

In this sense, that regression represents the claim that setting X to x will induce a normal distribution over Y with appropriate mean and variance. Let us consider that interpretation carefully. As said above, here causal effects are stochastic: the effect of changing X may not be a fixed change in Y, but rather a draw from a new distribution. This is not the interpretation usually given to regressions: typically, the "error term" represents omitted variables and the true effect is supposed to be deterministic. That interpretation can be brought into this framework by including the "error term" *explicitly* as an additional variable:

$$\begin{aligned} \Phi_Y(X = x, \epsilon) &= \alpha + \beta x + \epsilon \\ \Phi_\epsilon(\emptyset) &= Normal(0, \sigma) \end{aligned}$$

The implications of omitted variables on behavior are excluded from this paper, although this is clearly interesting and worth developing in other work. But in this paper, I wish to highlight disagreement that can result without missing variables.

Now for the definition of a causal Bayesian network:

Definition 1. *A causal Bayesian network is a pair* $M = \left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{C}}}\right\}$ *consisting of a causal structure* $\boldsymbol{\mathcal{C}}$ *and a set of causal probability functions* $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{C}}} = \left\{\ldots \Phi_V^{\boldsymbol{\mathcal{C}}} \ldots\right\}$*, one for each variable* $V \in \mathbb{V}$*.*[3]

An example of a causal Bayesian network can be brought to the earlier example of the firm.

$$\mathbb{V}_{firms} = \{\text{CEO skill } S, \text{firm quality } Q, \text{firm performance } P, \text{firm value } V\}$$

One causal structure $\boldsymbol{\mathcal{S}}$ over those variables is represented by Figure 2, which represents the claim that Skill causes Quality, both Skill and Quality cause Performance, and Performance alone causes Value. This embeds classic causal claims: for example, that changing Performance has no effect on either Skill or Quality.

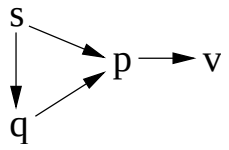


Figure 2: Causal structure $\boldsymbol{\mathcal{S}}$

[3]Adapted from Pearl (2000).

One could then write down a kind of structural equation model describing this system. A typical structural equation model of this system would look as follows, supposing the ϵs were error terms normally distributed around zero and the βs linear coefficients.

$$
\begin{aligned}
S &= \beta_{S,0} + \epsilon_S \\
Q &= \beta_{Q,0} + \beta_{Q,1} S + \epsilon_Q \\
P &= \beta_{P,0} + \beta_{P,1} S + \beta_{P,2} Q + \epsilon_P \\
V &= \beta_{V,0} + \beta_{V,1} P + \epsilon_V
\end{aligned}
$$

The causal Bayesian Network allows for a more general relationship between the variables and their causes. Instead of distributions around the means of the parent variables, the distributions can be arbitrary:

$$
\begin{aligned}
S &\sim \Phi_S^{\mathcal{C}} \\
Q &\sim \Phi_Q^{\mathcal{C}}(S) \\
P &\sim \Phi_P^{\mathcal{C}}(S, Q) \\
V &\sim \Phi_V^{\mathcal{C}}(P)
\end{aligned}
$$

Note that this allows for a different distribution on Q, for example, for each value of S.

I claim the causal Bayesian network provides a complete causal model of the phenomenon represented by the variables $\mathbb{V}$. It says what characteristics cause other characteristics in the phenomenon and, stochastically, how much one characteristic causes another.

2.2 Intervention actions and causal Bayesian Networks

Sam and Quincy in the opening example took over a firm and replaced the CEO with themselves. This disturbs the otherwise stable system: it takes a firm from the population, out of its current context, and changes or manipulates or intervenes on it.

Definition 2. *A* **intervention** *on variables $\mathbb{W} \subseteq \mathbb{V}$ is the setting of variables $\mathbb{W}$ from some current values $\vec{w}$ to some set of values $\vec{w}'$. The variables $\mathbb{W}$ will be called the intervention variables, and $\vec{w}'$ the intervention values, and variables $\mathbb{V} - \mathbb{W}$ the non-intervention variables.*

Pearl (2000) denotes the act of setting the intervention variables, appropriately enough, as $do(\mathbb{W})$. In the case of the firms, an intervention might be an agent replacing the skill of the CEO or changing the quality of the firm.[4]

An intervention breaks at least some of the current causal relationships that exist in the system at rest. Consider a barometer and the weather: the weather causes the barometer to change, and there is an

[4] "Interventions" are equivalent to "manipulations" in the econometrics literature.

observable, stable, natural, and stochastic steady state that $\{weather, barometer\}$ exist in: the weather and the barometer have some joint distribution. Now, suppose I intervene and squeeze the barometer. Now the causal relationship between the weather and the barometer is broken: whatever causal influence the weather had on the barometer has been usurped by my hand. The phenomenon has been pushed out of its natural state, and now the distribution over $\{weather, barometer\}$ is new... but not wholly unrelated to the original distribution. After all, the marginal distribution over weather continues unabated.[5]

The causal Bayesian network provides both which variables change and how much they change. The algorithm to determine these is as follows.

Theorem 1. *Suppose a vector $\vec{x}$ is drawn from $\mathbb{V}$, with entry x_V corresponding to variable $V \in \mathbb{V}$. Suppose the intervention $do(\mathbb{W} = \vec{w})$ is performed. Let $\mathbb{Y} \subseteq \mathbb{V}$ be the set of all variables which are descendants of at least one variable in $\mathbb{W}$. Let $\vec{x'}$ be the outcome vector of intervention. Then:*

1. *$m_V = w_V \in \vec{w}$ if $V \in \mathbb{W}$,*

2. *$m_V = v_V \in \vec{v}$ if $V \in \mathbb{V} - (\mathbb{Y} \cup \mathbb{W})$*

and otherwise, if $V \in \mathbb{Y} - \mathbb{W}$, then m_V has a distribution. The distribution is determined by:

$$F(\mathbb{Y} \backslash \mathbb{W} | do(\mathbb{W})) = \prod_{Y \in \mathbb{Y} \backslash \mathbb{W}} \Phi_Y(y | pa_{\mathcal{C}}(Y) \subset \vec{m})$$

This states the following: that when the set $\mathbb{W}$ of variables are manipulated by being set to particular values, those variables change to the new values (point number 1.) Other than those variables, only variables which are descendants of the variables in $\mathbb{W}$ change. The descendant variables are those variables which are caused by variables in $\mathbb{W}$, or the variables which are caused by those variables, etc. Hence, non-descendant variables which are also not in $\mathbb{W}$ do not change (point number 2.) Finally, the causal Bayesian network delivers what those descendant variables change to. The probability of a selection of variables, conditional on a particular variable, can be constructed by chaining the relevant conditional distributions. For example,

$$f(x, y|z) \quad = \quad f(x|y, z) f(y|z)$$

I will illustrate the all the objects discussed so far in this framework with the investor example: suppose a CEO is taking over a firm and investors are trying to forecast what will happen to the value of this company when she takes over.

Here I suppose for simplicity that firms are defined by three only values:

1. S, the skill of the CEO ("she is a talented manager");

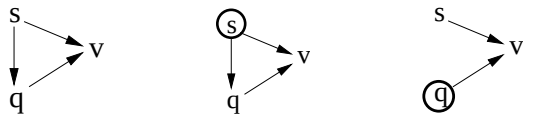
2. Q, the quality of the firm ("quality of the product this firm produces");

[5]The weather/barometer example is in both Druzel and Simon (1993) and Pearl (2000).

3. V, the value of the firm ("the current market assessment of the value of this firm")

So $\mathbb{V} = \{S, Q, V\}$. Suppose investors know the skill of the new CEO is some level s_c. Then the investors are trying to forecast the effect on V of $do(S = s_c)$.

One conjecture is that CEO skill and firm quality create value and, in addition, CEO skill causes firm quality: a good CEO causes the firm to be better managed and create more or better output. The causal structure $\boldsymbol{\mathcal{S}}$ which represents that conjecture is depicted in Figure 3(a).



(a) Without Intervention. (b) An Intervention on S (c) An Intervention on Q

Figure 3: Directed Acyclic Graphs representing $\boldsymbol{\mathcal{S}}$

$\boldsymbol{\mathcal{S}}$ fully captures the causality that the theory puts forth. What is the investor's problem? The investor is trying to forecast the effect on V of an intervention on S. Figure 3(b) represents that intervention: one would expect both Q and V to change. How much do they change? Suppose the investor's causal relation $\boldsymbol{\mathcal{S}}$ was augmented with a set of parameters $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{S}}}$ (i.e., the $\boldsymbol{\mathcal{S}}$-causal probability functions):

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{S}}} \quad = \quad \left\{\Phi_S^{\boldsymbol{\mathcal{S}}}(s), \Phi_Q^{\boldsymbol{\mathcal{S}}}(q|S=s), \Phi_V^{\boldsymbol{\mathcal{S}}}(v|S=s, Q=q)\right\}$$

What will be the distribution of Q under $do(S = s_c)$? Q will be distributed according to $\Phi_Q(q|S = s')$:

$$F_{do(S=s_c)}(q|s_c) = \Phi_Q^{\boldsymbol{\mathcal{S}}}(q|S = s_c)$$

This follows the original definition of the causal probability function.

In calculating the distribution over V there is a direct effect through the fact that S causes V, and an indirect effect, from the fact that S causes Q causes V.

$$F_{do(S=s_c)}(v|s) = \Phi_V^{\boldsymbol{\mathcal{S}}}(v|s_c, q)\Phi_Q^{\boldsymbol{\mathcal{S}}}(q|S = s_c)$$

For comparison, what if the investors were solving a different problem: one in which there was a known exogenous change in quality Q (depicted in Figure 3(c)). Then, under $\boldsymbol{\mathcal{S}}$, the investors would expect only V to change. The missing arrow (as per Pearl's convention) represents the fact that the intervention in $\mathbb{V}$ interrupts one of the existing causal relationships: the effect of s on q.

Suppose the initial values of $\mathbb{V}$, before intervention, are $\vec{v} = \{s_k, q_k, v_k\}$. The distribution of S would be atomic: $S = s_k$. On the other hand, the distribution of V would be defined by:

$$F_{do(Q=q')}(v|s) = \Phi_V^{\boldsymbol{\mathcal{S}}}(v|s_k, q')$$

Now suppose further that the observer intervened on $\mathbb{W}'$, but had not observed s_k. The distribution over V that the observer would expect would therefore need to incorporate the observer's ignorance over the true value of s_k. In that case, the distribution this observer would expect to see over V would be:

$$F'_{do(\mathbb{W}=\vec{w})}(v) = \sum_s \Phi_V^{\boldsymbol{\mathcal{S}}}(v|s,q')\Phi_S^{\boldsymbol{\mathcal{S}}}(s)$$

2.3 The set of reasonable models

A restatement of the goal of this framework: to construct, for each $\{\mathbb{V}, F\}$ pair, a set $\mathbb{M}$ of $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{C}}}\right\}$ pairs that are 'reasonable' for the agent to believe given $\{\mathbb{V}, F\}$.

First, I show what data (F) are generated by a particular causal Bayesian network. Then one can ask the question: What other causal Bayesian networks could generate those same data? The answer to that question, coupled with assumptions of minimalism and stability (which I explain below) defines the set of reasonable models.

Mapping from causal Bayesian networks to data:

Lemma 2. $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}\right\}$ *defines a unique distribution* F *over supp*$(\mathbb{V})$*, where:*

1. $dF(\mathbb{V} = \vec{v}) = \prod_{V\in\mathbb{V}} d\Phi_V(v|pa_{\boldsymbol{\mathcal{C}}}(v))$,
2. $pa_{\boldsymbol{\mathcal{C}}}(V)$ *be an associated instance of* $Pa_{\boldsymbol{\mathcal{C}}}(V)$*; i.e.,* $pa_{\boldsymbol{\mathcal{C}}}(V) \in supp(Pa_{\boldsymbol{\mathcal{C}}}(V))$*,*
3. *and* $d\Phi_V(v|\vec{w})$ *is the pdf at* v *associated with* $\Phi_V(pa_{\boldsymbol{\mathcal{C}}}(V))$

Proof. Without loss of generality, suppose there are n variables in $\mathbb{V}$, and they are ordered such that parents have lower indices than children. That is, for all $V_i, V_j \in \mathbb{V}$, if $V_i \in Pa(V_j)$ then $i < j$. Since there are no cycles, this is well-defined. Suppose that there are n variables in $\mathbb{V}$. Then let f be the joint probability density function (and used to represent marginal density functions). Then it must be the case that $dF(v_j|V_0,\ldots,V_{j-1}) = d\Phi_j(V_j|pa(V_j))$. This is true by the definition of a causal probability function: any time that $Pa(V_j) = pa(V_j)$, the distribution $\Phi_j(V_j|pa(V_j))$ is induced on V_j. When the parents of V_j have values $pa(V_j)$, then the distribution $\Phi_j(V_j|pa(V_j))$ is assigned to V_j. No other variables affect the distribution of V_j, and by virtue of the ordering, $Pa(V_j) \subseteq \{V_k,\ldots,V_{j-1}\}$.

The repeated application of Bayes's Rule demonstrates the equivalence claimed in the lemma.

$$\begin{aligned} dF(v_1, v_2, \ldots, v_n) &= dF(v_n|v_1, \ldots, v_{n-1})dF(v_1, \ldots, v_{n-1}) \\ &= d\Phi_n(v_n|pa(V_n))dF(v_1, \ldots, v_{n-1}) \\ &= d\Phi_n(v_n|pa(V_n))dF(v_{n-1}|v_1, \ldots, v_{n-2})dF(v_1, \ldots, v_{n-2}) \\ &= d\Phi_n(v_n|pa(V_n))d\Phi_{n-1}(v_{n-1}|pa(V_{n-1})dF(v_1, \ldots, v_{n-2}) \\ &\ldots \\ \implies dF(v_1, v_2, \ldots, v_n) &= \prod_{\{i|1\leq i\leq n\}} d\Phi_i(v_i|pa(V_i)) \end{aligned}$$

□

To take a causal relation $\boldsymbol{\mathcal{C}}$ and a joint distribution F, and construct $\widehat{\boldsymbol{\Phi}}$ which is consistent with both, is the act of calibrating the causal structure to data, or, calibrating the causal Bayesian network. Namely, suppose that the data F is observed. Now for each Φ_V, assign the following:

$$\Phi_V(pa(V)) = F(V|pa(V))$$

For a $\boldsymbol{\mathcal{C}}$-exogenous variable V (for which $Pa(V)$ is empty), the appropriate calibration is that $\Phi_V = F(V)$, that is, simply the marginal distribution, conditional on nothing.

By this mechanism, for a given F and $\boldsymbol{\mathcal{C}}$, the $\widehat{\boldsymbol{\Phi}}$ is unique (Pearl 2000). However, it is not the case that for every $\boldsymbol{\mathcal{C}}$ with the appropriate variables can a $\widehat{\boldsymbol{\Phi}}$ be constructed. If an F exhibits a $\widehat{\boldsymbol{\Phi}}$ relative to $\boldsymbol{\mathcal{C}}$, then it is said that F is *Markov relative to* $\boldsymbol{\mathcal{C}}$. This is important in statistical modeling because it is "a necessary and sufficient condition for a DAG $\boldsymbol{\mathcal{C}}$ to expalin a body of empirical data produced by F (Pearl 2000)."

A graph $\boldsymbol{\mathcal{C}}$ represents F if the following is true: For every two variables X and Y in $\mathbb{V}$, if X and Y are independent or conditionally independent, given any set of other variables in $\mathbb{V}$, then they are not connected by an edge, otherwise they are. And, for every two variables X and Y, if they are independent, then there must not be one path of arrows running from one to the other.

This is best illustrated with an example. Suppose $\mathbb{V} = \{X, Y, Z\}$ and F is such that Y renders X and Z conditionally independent, but X and Z are otherwise dependent in the data. Then the following graphical configurations are possible:

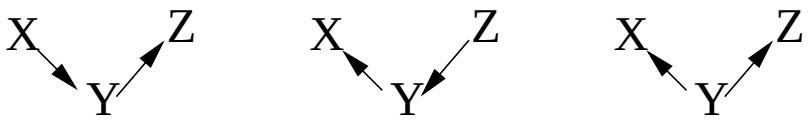


Figure 4: Directed Acyclic Graphs that are compatible with F

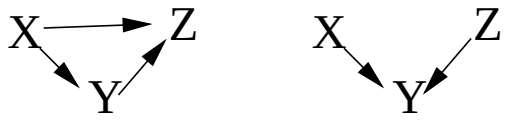


Figure 5: Directed Acyclic Graphs that are incompatible with F

Figures 4(a)-4(c) are all compatible with F. Figure 5(a) is rejected because X and Z are conditionally independent given Y, which suggests that any effect X has on Z goes through Y, unless by mere coincidence they cancel each other out (that such coincidences are ruled out is the assumption of what Pearl calls *stability*.) It is also ruled out since the extra branch is not needed to generate appropriate causal probability functions, which is ruled out by *minimalism* (if two models explain the same data, than the less complicated should be preferred.) Figure 5(b) is ruled out because if X and Z have no, even indirect, causal effect on each other than they should be completely independent in the data (recall, I suppose there are no omitted variables.)

So the sets of causal probability functions associated with Figures 4(a)-4(c) form the three reasonable models given F.

The set of reasonable models is characterized by this theorem, by Verma and Pearl (1990):

Theorem 3. *(Verma and Pearl). Two DAGs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow.*

2.4 Causal Coherence

With regard to the CEO-replacement problem, consider an investor I$\boldsymbol{\mathcal{S}}$ who believes the causal relation $\boldsymbol{\mathcal{S}}$. $\boldsymbol{\mathcal{S}}$ states that skill causes quality (depicted in Figure 3(a) and discussed in the previous section). Taking her causal model to be correct, she acts rationally. This is defined as causal coherence.

Definition 3. *An agent i is* **causally coherent with** $\left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i\right\}$ *if she behaves rationally supposing* $\left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i\right\}$ *were true.*

A causally coherent agent believes that interventions into the phenomenon V will be resolved according to $\left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i\right\}$.

Now I can precisely define agents who might agree about common information—that is, the observed distribution F—but disagree about causal models. I define agents who are causally coherent with a relation $\boldsymbol{\mathcal{C}}_i$ and have calibrated it to some distribution F as causally coherent with data.

Definition 4. *An agent i is* **causally coherent with data** F *if she is causally coherent with* $\left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i\right\}$, *where:*

1. $\boldsymbol{\mathcal{C}}_i \in \mathbb{M}(F)$

2. $\widehat{\boldsymbol{\Phi}}_i$ *results from* $\boldsymbol{\mathcal{C}}_i$ *calibrated to* F

Causal coherence can represent the behavioral claim that agents may confuse evidence consistent with their model with evidence for their model. Suppose the agent was taught the theory that $(\boldsymbol{\mathcal{S}}, \widehat{\boldsymbol{\Phi}})$ described the phenomenon $\mathbb{V}$. Her observation of the phenomenon would be consistent with her theory. This may naturally increase her confidence in this theory, although it is in fact not evidence, because the alternative model $(\boldsymbol{\mathcal{Q}}, \widehat{\boldsymbol{\Phi}}')$ fits the data equally well.

Now I have described the causal Bayesian network framework for representing agents' mental models of phenomena. This framework is appropriate for modeling decision-making under causal ambiguity. In the following section, I provide a utility representation of an agent from observing her choices.

3 A Utility Representation of the Causally Coherent Agent

In this section, I adapt Karni's (2005) representation theorem to the causal model setting to gain a utility representation for the causally coherent agent. Karni (2005) provides a framework in which a subjective expected utility representation emerges without reference to a state space; instead, he uses an event space, which has a probability distribution the agent can manipulate. Note that both causal and evidential decision theory (e.g. Jeffrey (1964), Joyce (1999)) are extensive literatures in decision theoretic structures which allow for actions to manipulate probabilities over states, in contrast to Savage and in a manner similar to Karni. However, those literatures differ in that choice behavior in those cases do not completely determine utility functions and probabilities in the sense of Savage (they have other advantages, however.) I seek a representation in which choice behavior completely determines the utility function and probabilities, and hence follow Karni's model.

Let X and Y be random variables with finite support, and let Z be an arbitrary variable among the two. These random variables will provide both actions and *effects*, which in Karni's framework take the place of states. Let $\Theta = \text{supp}(X) \times \text{supp}(Y)$ be the set of effects. Let $do(Z = z)$ be the intervention action which sets variable Z to some value z. The intervention induces a distribution on effect space:

$$do(Z = z) : \text{supp}(Z) \to \Delta(\Theta)$$

where $\Delta(\Theta)$ is a set of distributions over Θ. $\mathbb{I}_Z$ is the set of intervention actions on the variable X.

$$\mathbb{I}_Z = \{do(Z = z) | z \in \text{supp}(Z)\}$$

In addition, $do(\emptyset)$ is the non-intervention action, when the agent does not intervene in the system and instead allows the system to run its natural course. Let the set of all intervention acts be $\mathbb{I} = \mathbb{I}_X \cup \mathbb{I}_Y \cup \{do(\emptyset)\}$

with arbitrary element *do*. Let $\mathbb{B}$ be the set of all functions $b : \Theta \to \mathbb{R}$, where b is called a *bet* which yields a real-valued payoff for each effect. Bets are exactly like Savage acts where effects play the role of states, denoted here as bets instead to be consistent with Karni and to differentiate them from the intervention actions, which have more of the spirit of the "activity" the agent is engaged with. Denote $(b_{-(x,y)}, r)$ as the bet which awards b in all effects $(x', y') \neq (x, y)$, and awards r in effect (x, y). In other words, it is the bet b, with the (x, y)th entry replaced with r.

Agents will chose over (intervention, bet) pairs. The set of all (intervention,bet) pairs will be denoted $\mathbb{C} = \mathbb{I} \times \mathbb{B}$, and the agent's preference relation over $\mathbb{C}$ will be denoted $\succsim$.

The agent might believe that under some action *do*, an effect (x, y) might be impossible. This will be captured by the notion of null effects. An effect (x, y) is *null given do* if $(do, (b_{-(x,y)}, r)) \sim (do, (b_{-(x,y)}, r'))$ for all r, r'. Following Karni, I assume every effect is nonnull given some action. Let $\Theta(do; \succsim)$ be the set of effects that are nonnull given action *do*.

Recall the ongoing example of Investor I$\boldsymbol{\mathcal{S}}$ and Investor I$\boldsymbol{\mathcal{Q}}$, who invest in a company and have the opportunity to change the Skill of the CEO or the Quality of the firm. The *actions* in this context are the various $do(S)$ and $do(Q)$, that is, the act of intervening on the phenomenon in its natural state and setting the value of Skill or Quality to a specified level. The set of effects Θ are all possible values (s, q) that the phenomenon might attain. The decision maker is allowed to choose pairs of interventions $do(S = s), do(Q = q)$ and bets over (s, q) outcomes. These bets can be thought of as representing the role of V in the firm model; the bets are, for example, going short or long on the company's stock. The payoff is determined jointly by the CEO skill and firm quality.

When the choices of (intervention, bet) pairs are observed, and satisfy the axioms below, then a representation emerges. This representation specifies

1. a unique utility function u over money (the bets), and

2. A unique set of probability distributions π that each intervention induces.

Karni's original four axioms will deliver a representation with the properties above. With my additional axioms 5 and 6, it can be determined whether the agent adheres to the model $X\boldsymbol{\mathcal{C}}Y$ or $Y\boldsymbol{\mathcal{C}}X$, or perhaps neither. Causal coherence with, for example, model $X\boldsymbol{\mathcal{C}}Y$, has three additional requirements.

First of all, the intervention $do(Z = z)$ fixes Z and induces a distribution over the other variable W. This means that causal coherence will require that that, for all distributions induced on Θ by $do(Z = z)$, the distribution puts zero probability on effects (z', w), for $z' \neq z$. Axiom A5, the intervention axiom, will deliver this requirement by rendering such impossible effects as null.

Second, if $X\boldsymbol{\mathcal{C}}Y$ then Y not $\boldsymbol{\mathcal{C}}X$. This means that causal coherence with $X\boldsymbol{\mathcal{C}}Y$ will require that $do(Y = y)$ and $do(Y = y')$ will induce the same distribution over X. This will be delivered by axiom A6, the axiom of causal irreversibility.

Third, the overall joint distribution over Θ and the causal distributions will have to satisfy $P(x,y) = P(y|do(x))P(x)$. The joint distribution $P(x,y)$ is $\pi(x,y|do(\emptyset))$. The distribution $P(y|do(x))$ is $\pi(x,y|do(x))$. And the distribution $P(x)$ is $\pi(x,y|do(y)) = \pi(x,y'|do(y'))\quad \forall y,y'$, by the previous axiom. The requirement that $P(x,y) = P(y|do(x))P(x)$ is not delivered axiomatically, and instead is left as a condition that must be checked for causal coherence.

3.1 Axioms

Here are the axioms. Axioms 0-4 are from Karni (2005). Axiom 0 is a structural axiom, and the rest are behavioral. The first two behavioral axioms are standard. The third and fourth are discussed at length in Karni and introduced there. They provide separability between bets, actions, and effects. I then introduce axioms A5, the intervention axiom, and A6, the causal irreversibility axiom, which provide for the causal model structure.

Following Karni, a bet $\hat{b}$ is a *constant-valuation bet* on Θ if $(do(x),\hat{b}) \sim (do(x'),\hat{b})$ for all $do(x), do(x')$ in some $\hat{\mathbb{I}} \subseteq \mathbb{I}$ and $\bigcap_{do(x)\in\hat{\mathbb{I}}} \{b' \in \mathbb{B} | (do(x),b') \sim (do(x),b)\} = \left\{\hat{b}\right\}$. In essence, constant valuation bets leave the agent indifferent across outcomes: the value of the bet is sufficient to offset the value of the effect. (There is an additional requirement that constant valuation bets are at least pairwise unique across actions.) Constant-valuation bets are used to allow utility to be effect-dependent, which is analogous to state-dependent utility. Since I assume that utility is effect-independent through axiom A4 below, I do not make much use of the constant-valuation bets.

Recall $\Theta(do(x), \succsim)$ are the set of effects which are nonnull given $do(x)$. Then two effects $(x,y),(x',y')$ are said to be *elementarily linked* if there exists actions $do(x), do(x')$ such that $(x,y),(x',y') \in \Theta(do(x),\succsim) \cap \Theta(do(x'),\succsim)$. And two events $(x,y),(x',y')$ are *linked* if there are a sequence of events, such that each is linked to its neighbor, and the first is linked to (x,y) and the last linked to (x',y'). In essence, two events are elementarily linked if there are two actions which weight both effects positively. Two events are linked if they are connected by some sequence of linked events. Linked events are required in Axiom A0 to establish comparability between events.

Given these definitions, Karni's Axiom A0 is:

Axiom. (A0) *(Karni) Every pair of effects is linked, there exist constant-valuation bets b, b' such that $b' \succsim b$ and, for every $(do(x),b) \in \mathbb{C}$, there is a constant-valuation bet $\hat{b}$ satisfying $(do(x),b) \sim \mathbb{C}$.*

This structural axiom first requires comparability across effects. This allows for the definition of a single utility function. Second, there must be one constant-valuation bet which is superior to another. This is akin to the standard Savage axiom that the decision problem is non-trivial, in particular when tied with the next point. Third, it requires a constant valuation bet for each choice: a constant valuation bet benchmark that

is indifferent to each possible (action,bet) choice. We see how these play the role of the constant acts in the Savage framework.

Axiom. (A1: Weak Order) $\succsim$ *on* $\mathbb{C}$ *is a complete and transitive binary relation.*

Axiom. (A2: Continuity) *For all* $(do(x), b) \in \mathbb{C}$*, the sets* $\{(do(x), b') \in \mathbb{C} | (do(x), b') \succsim (do(x), b)\}$ *and* $\{(do(x), b') \in \mathbb{C} | (do(x), b) \succsim (do(x), b')\}$ *are closed.*

These axioms are standard. First, that $\succsim$ is a preference relation, and second, that there is continuity in the bet (act) space.

Axiom. (A3: Action-independent betting preferences) *(Karni) For all* $do(z), do(z') \in \mathbb{I}, b, b', b'', b''' \in \mathbb{B}, \theta \in \Theta(do(z)) \cap \Theta(do(z'))$ *and* $r, r', r'', r''' \in \mathbb{R}$*, if* $(do(z), (b_{-\theta}, r)) \succsim (do(z), (b'_{-\theta}, r'))$*,* $(do(z), (b'_{-\theta}, r'')) \succsim (do(z), (b_{-\theta}, r'''))$*, and* $(do(z'), (b''_{-\theta}, r')) \succsim (do(z'), (b'''_{-\theta}, r))$*, then* $(do(z'), (b''_{-\theta}, r'')) \succsim (do(z'), (b'''_{-\theta}, r'''))$

Karni (2005) explains:[6] "To grasp the meaning of action-independent betting preferences, think of the preferences $(do(z), (b_{-\theta}, r)) \succsim (do(z), (b'_{-\theta}, r'))$ and $(do(z), (b'_{-\theta}, r'')) \succsim (do(z), (b_{-\theta}, r'''))$ as indicating that, given action $do(z)$ and effect θ, the intensity of the preferences r'' over r''' is sufficiently larger than that of r over r' as to reverse the preference ordering of the effect-contingent payoffs $b_{-\theta}$ and $b'_{-\theta}$. This axiom requires that these intensities not be contradicted when the action is $do(z')$ instead of $do(z)$." It means if r'' is sufficiently better than r''' under action $do(z)$ to reverse preferences, than it shouldn't make the bet less attractive under action $do(z')$. That is, how the agent values money doesn't change when the action changes.

Here is an example: Suppose under $do(x)$, the bet b yielding 2 in outcome (x, y) was preferred over the bet b' yielding 1 in outcome (x, y): $(b_{(x,y)}, 2) \succsim (b'_{(x,y)}, 1)$. Suppose further that replacing 2 with 3 and 1 with 4 was enough to reverse preferences, such that the modified second bet was preferred: $(b'_{(x,y)}, 4) \succsim (b_{(x,y)}, 3)$ Then, under $do(x')$, making that same change, from $\{2, 1\}$ with $\{3, 4\}$ for a different set of bets, should not make the second bet *less* attractive. (It may not make the second bet more attractive, but it shouldn't make it worse.) (If $(b''_{(x,y)}, 1) \succsim (b'''_{(x,y)}, 2)$, then $(b''_{(x,y)}, 4) \succsim (b'''_{(x,y)}, 3)$.)

Axiom. (A4: Effect-independent betting preferences) *(Karni) For all* $do(z) \in \mathbb{I}, b, b', b'', b''' \in \mathbb{B}, \theta, \theta' \in \Theta(do(z))$ *and* $r, r', r'', r''' \in \mathbb{R}$*, if* $(do(z), (b_{-\theta}, r)) \succsim (do(z), (b'_{-\theta}, r'))$*,* $(do(z), (b'_{-\theta}, r'')) \succsim (do(z), (b_{-\theta}, r'''))$*, and* $(do(z), (b''_{-\theta'}, r')) \succsim (do(z), (b'''_{-\theta'}, r))$*, then* $(do(z), (b''_{-\theta'}, r'')) \succsim (do(z), (b'''_{-\theta'}, r'''))$

The interpretation is similar to that of action-independent betting preferences: if r'' is sufficiently better than r''' under effect θ to reverse preferences, than it shouldn't make the bet less attractive under action θ'. That is, how the agent values money doesn't change when the *effect* changes.

Now define $\Theta_{Z=z}$ as those effects that are consistent with variable Z having value z. Namely, $\Theta_{X=x} = \{(x, y) | y \in \text{supp}(Y)\}$ and $\Theta_{Y=y} = \{(x, y) | x \in \text{supp}(X)\}$.

[6] Modified to have consistent notation.

Axiom. (A5: Interventions) $(do(x),(b_{-\theta},r)) \sim (do(x),(b_{-\theta},r'))$, *for all* $r,r' \in \mathbb{R}, \theta \in \Theta - \Theta_{X=x}$.

This axiom imposes the causal structure. Consider two (action, bet) pairs described above. Consider the action $do(X=x)$. Then this axiom requires that it doesn't matter what the rewards are in any effect $(x',y) \in \Theta - \Theta_{X=x}$, where $x \neq x'$. The agent knows with certainty that those effects (x',y) will never occur. Hence changing the rewards on those effects should do nothing to change preference.

Now, let $b_{z \leftrightarrow z'}$ be the bet b, which each entry $b(z,w)$ replaced with entry $b(z',w)$ and vice versa. Then:

Axiom. (A6: Causal Irreversibility) *If there exists* $b \in \mathbb{B}$ *such that* $(do(x),b) \succ (do(x'),b_{x \leftrightarrow x'})$, *then for all* $do(y) \in \mathbb{I}_Y$, $\bar{b} \in \mathbb{B}$,

$$(do(y),\bar{b}) \sim (do(y'),\bar{b}_{y \leftrightarrow y'})$$

for all $r \in \mathbb{R}$.

Similarly, if there exists $b \in \mathbb{B}$ *such that* $(do(y),b) \succsim (do(y'),b_{y \leftrightarrow y'})$, *then for all* $do(x) \in \mathbb{I}_X$, $\bar{b} \in \mathbb{B}$,

$$(do(x),\bar{b}) \sim (do(x'),\bar{b}_{x \leftrightarrow x'})$$

for all $r \in \mathbb{R}$.

This is the interpretation of this axiom: if there exists b such that $(do(x),b) \succ (do(x'),b_{x \leftrightarrow x'})$, that reveals that the agent believes that $do(x)$ causes a different probability distribution than $do(x')$ over Y. In that case, since the agent has revealed she believes $X \boldsymbol{\mathcal{C}} Y$, we would like to assure that Y does not cause X. Hence, the agent should consider it equally probable that $X=x$ under $do(y)$ as under $do(y')$. Therefore, a bet which yields the vector $\vec{b}$ for effects $(\cdot,y)$ will be as valuable under $do(y)$ as the bet which yields vector $\vec{b}$ under $do(y')$.

3.2 The Representation Theorem

This result is that, if a preference relation $\succsim$ adheres to axioms A1-A6, then there is a unique utility function u over bets and set of probabilities π over effects under interventions such that (do,b) is represented by $f_{do}(\sum_{(x,y)\in\Theta}[\sigma(x,y)u(b)+\kappa(x,y)]\,\pi(x,y)|do)$, and, additionally, if these probabilities additionally have the property that $\pi((x,y)|do(\emptyset)) = \pi(y|do(x))\pi(x)$, where $\pi(y|do(x)) = \pi(x,y|do(x))$ and $\pi(x) = \pi((x,y)|do(y))\forall y$, then the agent adheres to causal model $X \boldsymbol{\mathcal{C}} Y$ calibrated to $\pi((x,y)|do(\emptyset))$ in the sense that the revealed probabilities π conform to such a model. It is the Karni representation theorem, with the additional causal structure, which appears in statement 3.

Theorem 4. *Suppose axiom (A0) is satisfied, and* $|\Theta(a)| \geq 2 \quad \forall do(x) \in \mathbb{I}$. *Then:*

1. *The following are equivalent:*

(a) *The preference relation* $\succsim$ *on* $\mathbb{C}$ *satisfies A1-A6*

(b) *There exists*

i. *a continuous function* $u : \mathbb{O} \to \mathbb{R}$, *and for each* $\theta \in \Theta$, *there are numbers* $\sigma(\theta) > 0$ *and* $\kappa(\theta)$

ii. *a family of probability measures* $\{\pi(x, y|do(Z = z))\}$ *on* $supp(X) \times supp(Y)$, *and* $\pi(x, y|do(\emptyset))$, *and*

iii. *a family of continuous, increasing functions* $\left\{f_{do(x)}\right\}_{do(x)\in\mathbb{I}}$,

such that, for all $(do(W = w), b)$, $(do(Z = z), b') \in \mathbb{C}$,

$$(do(w), b) \succsim (do(z), b') \iff f_{do(w)}\left(\sum_{\{w,s\}\in\Theta} [\sigma(\{w, s\})u(b(\{w, s\}) + \kappa(\{w, s\})]\, \pi(\{w, s\}\,|do(x))\right) \geq f_{do(z)}\left(\sum_{\{z,s\}\in\Theta} [\sigma(\{z, s\})u(b'(\{z, s\}) + \kappa(\{z, s\})]\, \pi(\{z, s\}\,|do(z))\right)$$

2. u, σ, *and* κ *are unique and* $\left\{f_{do(x)}\right\}_{do(x)\in\mathbb{I}}$ *are unique up to a common, strictly monotonic increasing transformation.*

3. *For each* $do(x) \in \mathbb{I}$, $\pi(\{z, w\}\,|do(z))$ *is unique and* $\pi(\{z, s\}\,|do(z)) = 0$ *if and only if* $\{z, s\}$ *is null given* $do(z)$, *so* $\pi(\{z', s\}\,|do(z)) = 0, \forall z' \neq z$..

Furthermore, if the π *satisfy* $\pi((x, y)|do(\emptyset)) = \pi((x, y)|do(x))\pi((x, y)|do(y))$, *then:*

1. *If there exists* $b \in \mathbb{B}$ *such that* $(do(x), b) \succ (do(x'), b_{x\leftrightarrow x'})$ *then* π *satisfies* $\pi((x, y)|do(y)) = \pi((x, y')|do(y')) := \pi(x) \quad \forall y, y'$, $\pi\left((x, y)|do(x)\right)$ *can be rewritten as* $\pi(y|do(x))$, *and the agent adheres to causal model* $X\boldsymbol{\mathcal{C}}Y$,

2. *Else if there exists* $b \in \mathbb{B}$ *such that* $(do(y), b) \succ (do(y'), b_{y\leftrightarrow y'})$ *then* π *satisfies* $\pi((x, y)|do(x)) = \pi((x', y)|do(x')) := \pi(y) \quad \forall x, x'$, $\pi\left((x, y)|do(y)\right)$ *can be rewritten as* $\pi(y|do(y))$, *and the agent adheres to the causal model* $Y\boldsymbol{\mathcal{C}}X$,

3. *Else if for all* $b \in \mathbb{B}$, $(do(x), b) \sim (do(x'), b_{x\leftrightarrow x'})$ *and* $(do(y), b) \sim (do(y'), b_{y\leftrightarrow y'})$, *then* π *satisfies* $\pi((x, y)|do(y)) = \pi((x, y')|do(y')) := \pi(x) \quad \forall y, y'$, *and* $\pi\left((x, y)|do(x)\right)$, $\pi((x, y)|do(x)) = \pi((x', y)|do(x')) := \pi(y) \quad \forall x, x'$. *The agent then adheres to the causal model* $X\neg\boldsymbol{\mathcal{C}}Y$ *and* $Y\neg\boldsymbol{\mathcal{C}}X$.

The proof follows Karni, save for those parts that explicitly reference axioms 5 and 6 in the following description. For every $do \in \mathbb{I}$, Axioms 1-3 imply the existence of jointly cardinal, continuous, additive representations of $\succsim_{do}$, so that (do, b) is represented by $\sum_{\theta\in\Theta} w_{do}(b(x, y), (x, y))$. Axiom A6 allows that the w_{do}s can be chosen such that $w_{do(y)}(b(x, y), (x, y)) = w_{do(y')}(b(x, y'), (x, y'))$. Then, two arbitrary constant valuation bets, b^* and b^{**}, such that $b^{**} \succ b^*$, are chosen as reference points, and the following normalization

is made: $w_{do}(b^*(x,y),(x,y)) = 0$ and $\sum_{(x,y)\in\Theta} w_{do}(b^{**}(x,y),(x,y)) = 1$, for all $do \in \mathbb{I}$. The probability $\pi(x,y|do)$ is defined to be $w_{do}(b^{**}(x,y),(x,y))$ and u is constructed by dividing all w_{do} by π. Axiom 4 assures that the resulting utility is also almost independent of effect, in the following form: $\sigma(\{w,s\})u(b(\{w,s\}) + \kappa(\{w,s\}))$. Finally, an action $\bar{do} \in \mathbb{I}$ is chosen and f_{do} is constructed with the constant valuation bets, so that $f_{do}\left(\sum_{\theta\in\Theta}\left[\sigma(\theta)u(\bar{b}(\theta)) + \kappa(\theta)\right]\right)\pi(\theta|do) = \sum_{\theta\in\Theta}\left[\sigma(\theta)u(\bar{b}(\theta)) + \kappa(\theta)\right]\pi(\theta|\bar{do})$.

By axiom 5, $\pi((x',y)|do(x)) = 0$ for all $x \neq x'$, and therefore $\pi((x,y)|do(x))$ can be rewritten as $\pi(y|do(x))$. By the implication of Axiom 6, $\pi((x,y')|do(y'))$ can be written as $\pi(x)$.

Axiom 5 renders all effects that involve non-intervened values of the intervention value null. This means that, under $do(x)$, Axiom 5 renders effect (x',y), for all $y \in \text{supp}(Y)$ null. By Karni's representation, the probability distribution $\pi(x',y;do(x))$ is null for all $x' \neq x$. This means that $\pi(x,y;do(x))$ can be interpreted as the causal probability function on y of $do(x)$.

Axiom A5 appears to be potentially inconsistent with Axiom A0's requirement that all effects are linked, but that is not the case. Recall, effects are elementarily linked when, for some pair of actions do and do', both effects are non-null. Two effects (x,y) and (x',y') are linked if there is a sequence of linked events connecting (x,y) to (x',y'). Axiom A5 requires widespread and systematic nullification of effects. Therefore, it is important to demonstrate that A5 and the requirement that all effects are linked are not inconsistent.

	do(y)	do(y')	do(y'')
do(x)	(x,y)	(x,y')	(x,y'')
do(x')	(x',y)	(x',y')	(x',y'')
do(x'')	(x'',y)	(x'',y')	(x'',y'')

Figure 6: No effects are elementarily linked

As Figure 6 demonstrates, without the non-intervention act $do(\emptyset)$, no effects are elementarily linked. Effect (x',y') can be in, at most, $\Theta(do(x'),\succsim)$ and $\Theta(do(y'),\succsim)$, and no other effects are in that intersection. Even requiring the largest possible set of non-null effects consistent with the axiom A5, there are too many null effects to link effects. However, with the non-intervention act, then many effects might be elementarily linked. For example, (x',y') and (x',y'') are elementarily linked:$(x',y'),(x',y'') \in \Theta(do(x'),\succsim) \cap \Theta(do(\emptyset),\succsim)$. Hence every two effects which vary in only one coordinate are elementarily linked (and therefore linked), so all effects are linked.

Now, suppose that pi satisfies

$$\pi((x,y)|do(\emptyset)) = \pi((x,y)|do(x))\pi((x,y)|do(y))$$

Then axiom A6 allows for the construction of a causal structure, either $X\boldsymbol{\mathcal{C}}Y$ or $Y\boldsymbol{\mathcal{C}}X$, or neither. Consider first when there exists $b \in \mathbb{B}$ such that $(do(x), b) \succ (do(x'), b_{x \leftrightarrow x'})$. Then Axiom 6 requires that for all $do(y) \in \mathbb{I}_Y$, $\bar{b} \in \mathbb{B}, (do(y), \bar{b}) \sim (do(y'), \bar{b}_{y \leftrightarrow y'})$ for all $r \in \mathbb{R}$. Since $\pi((x, y)|do(y))$ can be rewritten as $\pi(x|do(y))$, then this requires moreover that $\pi(x|do(y)) = \pi(x|do(y'))$ and hence can be written as simply one function $\pi(x)$. This can be interpreted as what the agent believes is the exogenous distribution of x, and hence the causal probability function of x. Then, $\pi(y|do(x))$ can be interpreted as the causal probability function of y, and the causal model is $X\boldsymbol{\mathcal{C}}Y$. This process works identically in reverse if it is revealed that $Y\boldsymbol{\mathcal{C}}X$. It is worth noting, then, if both $b \in \mathbb{B}$, $(do(x), b) \sim (do(x'), b_{x \leftrightarrow x'})$ and $(do(y), b) \sim (do(y'), b_{y \leftrightarrow y'})$, then the agent has revealed that she believes that X and Y are independent, and therefore the correct causal model is $X\neg\boldsymbol{\mathcal{C}}Y$ and $Y\neg\boldsymbol{\mathcal{C}}X$. It is also worth noting that if Axiom A6 were to fail, then this would be a case of cyclic causality and one would not expect $\pi((x, y)|do(x))\pi((x, y)|do(y))$ to have any particular meaning.

4 Applications

I have described the parts of the causal Bayesian network framework to represent agent's models of phenomena. I have proposed an axiomatic framework by which one can, after observing an agents' choices of interventions and bets, deduce her utility function (a utility function which represents her behavior) and the unique stochastic effect she believes her interventions will result in. This representation in an application of a result by Karni (2005) with an two additional axioms to define the causal structure.

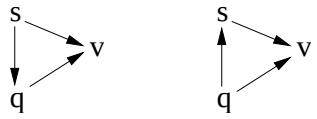
In this section, I place these agents in different scenarios. In the first, I address an example of a decision problem. I demonstrate that two agents who agree about observed data and have the same preferences may disagree about optimal interventions. Then, I take the agents to an auction, in which they participate in a causally coherent equilibrium, which I define below. In the auction, there is a causal curse in some ways similar to a winner's curse.

4.1 Information, Causal Coherence with Data, and Disagreement

Two agents who are causally coherent with the same data are precisely those agents who might agree about an (infinite) common source of information but disagree about best behaviors. This is because the same behavior–the same choice of intervention–is believed to map to different probability distributions over the phenomenon. The following example demonstrates the two investors of the ongoing example choosing different optimal interventions when their causal models differ, although their data F is common.

First let us consider investor I$\boldsymbol{\mathcal{S}}$, who believes the causal relation $\boldsymbol{\mathcal{S}}$ which states that skill causes quality. Suppose she calibrates her causal model to a distribution F. This calibration generates a unique $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{S}}}$.

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{S}}} = \left\{\Phi_S^{\boldsymbol{\mathcal{S}}}, \Phi_Q^{\boldsymbol{\mathcal{S}}}, \Phi_V^{\boldsymbol{\mathcal{S}}}\right\}$$



(a) Causal Relation $\boldsymbol{\mathcal{S}}$ (b) Causal Relation $\boldsymbol{\mathcal{Q}}$

Figure 7: Directed Acyclic Graphs $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{Q}}$

where

$$\Phi_S^{\boldsymbol{\mathcal{S}}}(s) = F(s)$$
$$\Phi_Q^{\boldsymbol{\mathcal{S}}}(q|S=s) = F(q|s)$$
$$\Phi_V^{\boldsymbol{\mathcal{S}}}(v|S=s, Q=q) = F(v|s,q)$$

Then investor I$\boldsymbol{\mathcal{S}}$ would choose $s^\star \in \text{supp}(S)$ to maximize:

$$\sum_v u_i(v) F^{\boldsymbol{\mathcal{S}}}_{do(S=s^\star)}(v|s^\star)$$
$$= \sum_v u_i(v) F(v|s^\star, q) F(q|S=s^\star)$$

Investor I$\boldsymbol{\mathcal{Q}}$ believes the causal relation $\boldsymbol{\mathcal{Q}}$, which is the belief that the quality of firms is inherent, and that high quality firms attract (cause) high-quality CEOs. The calibration to F generates a unique and distinct $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{Q}}}$:

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{Q}}} = \left\{ \Phi_S^{\boldsymbol{\mathcal{Q}}}, \Phi_Q^{\boldsymbol{\mathcal{Q}}}, \Phi_V^{\boldsymbol{\mathcal{Q}}} \right\}$$

where

$$\Phi_Q^{\boldsymbol{\mathcal{Q}}}(q) = F(q)$$
$$\Phi_S^{\boldsymbol{\mathcal{Q}}}(s|q) = F(s|q)$$
$$\Phi_V^{\boldsymbol{\mathcal{Q}}}(v|S=s, Q=q) = F(v|s,q)$$

Investor I$\boldsymbol{\mathcal{Q}}$, by contrast, would choose $s^{\star\star} \in \text{supp}(S)$ to maximize:

$$\sum_v u_j(v) F^{\boldsymbol{\mathcal{Q}}}_{do(S=s^{\star\star})}(v|s^{\star\star}, q^k)$$
$$= \sum_v u_j(v) F(v|S=s^{\star\star}, Q=q_k)$$

4.2 Describing the interaction of agents with different models

To describe the interaction of agents with different causal models, I introduce a new kind of equilibrium, called the causally coherent equilibrium. It has two main distinguishing features: that each agent believes

that her causal model is common knowledge, and that agents are only required to have an explanation for equilibrium events, as opposed to the stronger condition that expectations must be correct in equilibrium.

Causally coherent agents have a causal model and are unaware of alternative models. So if a causally coherent agent has a causal model $\boldsymbol{\mathcal{C}}_i$, it is natural that she also assumes other players to have that causal model $\boldsymbol{\mathcal{C}}_i$. This is the first component of the causally coherent equilibrium: that each agent i who forecasts outcomes with some causal model $\boldsymbol{\mathcal{C}}_i$ best responds to what she believes all other agents will play, assuming they forecast outcomes using causal model $\boldsymbol{\mathcal{C}}_i$.

In traditional equilibria, ones expectations are met at all information sets, or, in the case of weaker forms than Nash, such as Fudenberg and Levine's (1993) "Self-Confirming Equilibrium," exemplified in Eyster and Rabin (2005) and Esponda (2005), along the equilibrium path of play. Here, I relax that assumption and replace it with a weaker assumption that every agent must have *an explanation* for what she encounters in equilibrium. In other words, any profile of play and outcomes that is possible in equilibrium must be in the support of each agents' beliefs. This suggests correctly that the causally coherent equilibrium is a short-run phenomenon, and that agents might learn that some belief of theirs is wrong in the long run (although they may not be able to identify which one.)

Referring to the 'set of possibilities' as what an agent 'knows' to be possible, this assumption that the outcomes must be in the support of the beliefs is called "no knowledge violations."

Definition 5. *A coherent agent i's* **knowledge is violated** *when the agent observes an event that is impossible given her model, where "impossible" means an event that occurs with a non-positive probability (Pr=0 or zero density, as appropriate.)*

Given those two fundamental distinctions, here is the definition of the causally coherent equilibrium.

Definition 6. A Causally Coherent Equilibrium *of some game G, with associated phenomenon $\mathbb{V}$ and data F, is a set of actions played by each player i in which:*

1. *$\mathbb{V}$, F, and causal coherence of all agents are common knowledge.*
2. *Each agent i is endowed with a causal relation $\boldsymbol{\mathcal{C}}_i$, is causally coherent with F.*
3. *Each agent i plays an action consistent with some Bayes-Nash equilibrium E_i of the game implied by $G, \mathbb{V}, F$, and that all agents forecast using $\boldsymbol{\mathcal{C}}_i$, calibrated to F.*
4. *No agent's knowledge is violated in equilibrium.*

Item one reiterates that the phenomenon and associated distribution are common knowledge, so that all agents calibrate their causal models to the same (and true) set of information.

Since causally coherent agents take their theories as confirmed fact, and, since they are rational and believe that they are playing against other rational agents, they believe that their theory is commonly understood to be true.

Since agents have different causal models, and therefore at least some are wrong about the way the world works, they will typically under this framework not see the distribution of actions and payoffs that they expect. Hence this equilibrium is not stable in the long run. However, no knowledge violations means that, in the one-shot game, the agent can explain any event she encounters.

In the example below, there is an auction for a firm, and the winner of the firm replaces the current CEO with himself, which is an intervention on the firm. Each agent sees a signal about current quality of the firm, his opponent's bid, and, if she wins, the eventual draw, post-intervention, from supp($\mathbb{V}$). For no knowledge violations to hold, it must be the case that any event which obtains with positive probability in equilibrium–a particular signal, opponent bid, and vector $\vec{V} \in$ supp($\mathbb{V}$)–must also occur with positive probability under that agent's believed equilibrium. In the example below, this is true by the fact that support is infinite over supp($\mathbb{V}$), agents' skill (which determines their bid), and the signal, so their support sets are the same.

This equilibrium stands in contrast with Fudenberg and Levine's (1993) "Self-Confirming Equilibrium," exemplified in Eyster and Rabin (2005) and Esponda (2005). In those equilibria, the source of data is equilibrium play. In this model, the source of data is the phenomenon. The phenomenon exists apart from the play of the game. It is an external object which coordinates beliefs. In this equilibrium, agents make systematic mistakes, as one would expect from a misunderstanding of causal structure. However, these mistakes never result in an event that any agent deems impossible.

I construct an explicit example of a causally coherent equilibrium below.

4.3 Example: An auction for a company and a causal curse

Two aspiring CEOs, investor I$\boldsymbol{\mathcal{S}}$ and investor I$\boldsymbol{\mathcal{Q}}$, bid to take over a firm and replace the current CEO with himself. An infinite data stream about firms is public, so each agent believes she knows how CEO skill effects firm value: i.e., each agent has her own causal relation about the phenomenon of firm creation. The auction is a two-price auction, which is a simplified first-price auction. There is a single public signal about the quality of the firm, and each agent knows his own skill. Replacing the current CEO with the winner is an intervention, which exogenously changes skill of an existing firm, so the effect on quality is determined by the true causal model.

A two-price auction is a first-price, sealed-bid auction with only two allowable bids. It works as follows: each agent chooses one of two bids: $\$M > \$0$. The higher bid wins the object and ties are decided by the flip of a fair coin.

Both investors have seen an infinite data set of firms' Quality and CEO Skill. S and Q are binary

variables.[7] This is the observed symmetric joint distribution over S and Q, with associated marginal distributions, for some α, $\frac{2}{3} < \alpha < 1$:

$F(S,Q)$	$S=1$	$S=0$	$F(Q)$
$Q=1$	$\frac{1}{2}\alpha$	$\frac{1}{2}(1-\alpha)$	$\frac{1}{2}$
$Q=0$	$\frac{1}{2}(1-\alpha)$	$\frac{1}{2}\alpha$	$\frac{1}{2}$
$F(S)$	$\frac{1}{2}$	$\frac{1}{2}$	

Since $\alpha > \frac{1}{2}$, S and Q are correlated, so that good firms have good CEOs.

The value of the firm after the auction is determined by a bet on the (s,q) outcome. The bet $b(s,q)$ yields a payoff of \$1 if $S=Q=1$ and 0 otherwise.

$$b(s,q) = \begin{cases} 1 & \text{if } S=1, Q=1 \\ 0 & \text{otherwise} \end{cases}$$

Agents' Skill is drawn from a known distribution: Skill is 1 with probability $\frac{1}{2}$. This means that they are typical of the population of CEOs given by F.

The single publicly observable signal is σ. It follows a known distribution

$$G(\sigma = 1|q) = \begin{cases} \beta & \text{if } Q=1 \\ 1-\beta & \text{if } Q=0 \end{cases}$$

Note that $G(Q=1|\sigma=1) = \beta$, since $F(Q=1) = \frac{1}{2}$. In other words, when an agent of either type sees a signal of $\sigma = 1$ about the firm before intervention, the agent believes there is a β chance that the firm is (currently) of high Quality.

The players observe the single signal and place their bids simultaneously, then the winner is resolved. The winner performs the intervention of replacing the (unobserved) CEO Skill with his own skill. The new Quality is then resolved according to the true causal model: in the case of $\boldsymbol{\mathcal{S}}$, Q is determined by the distribution F, conditional on the winner's Skill. In the case of $\boldsymbol{\mathcal{Q}}$, the quality of the firm remains unchanged. Under $\boldsymbol{\mathcal{S}}$, since Q changes, the signal conveys no useful information. Under $\boldsymbol{\mathcal{Q}}$, the signal is useful. The winner then observes the new Quality of the firm, and the bet is resolved according to b above.

4.3.1 Play in the Causally Coherent Equilibrium

No agent will want to play M if he is of skill $S=0$, since the firm will be worth zero, so playing M can only make the agent worse off. It turns out that, for $0 < M \leq \frac{1}{2}\min\{\alpha, \beta\}$[8]

[7] This example has discrete types to be consistent with the representation theorem, which is for finite spaces $\text{supp}(S) \times \text{supp}(Q)$. In this example, the causally coherent equilibrium is also an ex-ante Nash equilibrium. In the continuous case, the causally coherent equilibrium is distinct from the Bayes-Nash. See section 4.4.

[8] See appendix section 7.1.1 for the details.

1. Investor I$\boldsymbol{\mathcal{S}}$ plays M only if $S_{\boldsymbol{\mathcal{S}}} = 1$

2. Investor I$\boldsymbol{\mathcal{Q}}$ plays M only if $S_{\boldsymbol{\mathcal{Q}}} = 1$ and $\sigma = 1$

In causally coherent equilibrium play, each agent plays the Bayes-Nash equilibrium associated with all agents having the same causal relation (a premise which is false.) Consider the Bayes-Nash equilibrium that investor I$\boldsymbol{\mathcal{S}}$ plays. He supposes that he plays against an agent who also believes $\boldsymbol{\mathcal{S}}$. Hence he believes that his opponent will only bid $\$M$ if he is of Skill 1 and only bid \$0 if he is of skill 0. Hence a high Skill investor I$\boldsymbol{\mathcal{S}}$ expects to win half the time against a fellow high Skill investor and, upon winning, win the bet α of the time.

Now consider the Bayes-Nash equilibrium that investor I$\boldsymbol{\mathcal{Q}}$ plays. He supposes that he is against an agent who also believes $\boldsymbol{\mathcal{Q}}$. Hence he believes that his opponent will only bid M if he is of Skill 1 and $\sigma = 1$, 0 otherwise. Hence a high Skill investor I$\boldsymbol{\mathcal{Q}}$ expects to win half the time when the signal is 1, and, upon winning, win the bet β of the time.

No agent encounters a knowledge violation when they play against each other. All agents have an explanation for any pattern of bids, wins, and losses. For example, since investor I$\boldsymbol{\mathcal{S}}$ does not know the type of his opponent, the first time that investor I$\boldsymbol{\mathcal{S}}$ loses to investor I$\boldsymbol{\mathcal{Q}}$, he 'learns' that his opponent is an investor I$\boldsymbol{\mathcal{S}}$ of the same skill. What he learns is false, but it is a coherent explanation for the event he witnessed.

I consider the case when $\alpha = \beta$, and suppose that $\frac{1}{2}\alpha = M$. This choice highlights the causal curse.

The auction is straight-forward for all pairings with one agent of skill $S = 0$. In that case, the agent with low skill always bids \$0. The interesting case is when investor I$\boldsymbol{\mathcal{S}}$ and investor I$\boldsymbol{\mathcal{Q}}$ both have skill $S = 1$.

Investor I$\boldsymbol{\mathcal{S}}$ bids M when his Skill is 1 in the Bayes-Nash equilibrium associated with all agents believing $\boldsymbol{\mathcal{S}}$; that is, he supposes that his opponent is plays the same strategy and that his (and his opponent's) payoff is determined by the causal relation $\boldsymbol{\mathcal{S}}$. He believes that if he wins, that he will get the \$1 payoff α of the time. This is not, however, the case if $\boldsymbol{\mathcal{Q}}$ is true. If he were not competing for the object, and simply getting it when he wanted to pay M, he would only get the \$1 payoff half of the time, which means he would still make a profit (since $M = \frac{1}{2}\alpha < \frac{1}{2}$). However, since he competes for the object, he ends up losing money on average, since investor I$\boldsymbol{\mathcal{Q}}$ is bids high precisely when Q is likely to be 1. Hence, investor I$\boldsymbol{\mathcal{S}}$ gets the \$1 payoff less than half the time. This violates his incentive constraint, and he would, were he to know this, be better off bidding 0. He would also, since he loses money on average, be better off getting out of the game entirely over bidding M.[9]

Investor I$\boldsymbol{\mathcal{Q}}$ bids M when his Skill is 1 and when he sees the signal $\sigma = 1$, and he also believes his opponent does the same. When $\boldsymbol{\mathcal{S}}$ is true, investor I$\boldsymbol{\mathcal{Q}}$ sees nothing he cannot explain. Whenever he wins the object, he gets what he expects: a payoff of \$1 exactly $\alpha = \beta$ of the time. Although he would also get

[9]Please see appendix about losing money on average.

that payoff when he bids \$0, he does not know this, nor ever learns it. Investor I$\boldsymbol{\mathcal{Q}}$ finds, however, that he never wins when he bids \$0. He has an explanation, since that is plausible (for any finite stream), simply unlikely.

Whether $\boldsymbol{\mathcal{S}}$ or $\boldsymbol{\mathcal{Q}}$ is true, the agent with the wrong causal model loses in some capacity: either on average losses in the case of investor I$\boldsymbol{\mathcal{S}}$ or by lost opportunity in the case of investor I$\boldsymbol{\mathcal{Q}}$. And each of them must rely on no knowledge violations instead of matching expectations about exactly one parameter. In the case of investor I$\boldsymbol{\mathcal{S}}$, that parameter is his payoff. In the case of investor I$\boldsymbol{\mathcal{Q}}$, that parameter is his win rate when he bids low.

Since investor I$\boldsymbol{\mathcal{S}}$ loses money on average, this is a kind of winner's curse. Note that there would be no winner's curse in this game, if all agents agreed on a causal model. The classic winners curse arises from incorrectly constructing opponent's estimates of the common component. Since both agents construct their values based on completely private information and completely public information, there are no deviant estimates of the common component. Instead, their different causal models serve, in some sense, as additional private 'signals' about the source of value of the firms.

4.4 Continuous type example

The set-up is similar to the previous example: two bidders for a firm with characteristics S and Q, and, in this case, an additional characteristic V, which is firm value (what, in this case, the agents are concerned with). There is a joint distribution over $\mathbb{V} = \{S, Q, V\}$ with the following properties:

1. S's marginal distribution is normal $(0, 1)$;
2. Q's conditional distribution on S is normal $(s, 1)$, that is, with a mean of s for each $s \in \text{supp}(S)$;
3. V's conditional distribution on S and Q is normal $(s + q, 1)$, that is, with mean $s + q$

The public signal is of a known distribution $G(q|\sigma)$, and is a mean-preserving spead of q, such that $E\left[G(q|\sigma)\right] = \sigma$. This means σ has been normalized such that one's expectation of q, after seeing σ, is just σ.

It is known that agents' skill is drawn from a distribution $H(s)$. It might be the case that $H(s)$ is $F(s)$, that is, the marginal distribution of s in the data, which would be the case if agents suspect that their opponents are typical of the population at large.[10]

[10] And believe either $\boldsymbol{\mathcal{S}}$, in which case skill is exogenous, or $\boldsymbol{\mathcal{Q}}$, and believed that skill is endowed, but that firms find good CEOs, but not actually cause otherwise bad CEOs to become good.

Then in the causally coherent equilibrium in which each player believes they are playing the symmetric Bayes-Nash, investor I$\boldsymbol{\mathcal{S}}$ and investor I$\boldsymbol{\mathcal{Q}}$ play according to:

$$b_{\boldsymbol{\mathcal{S}}}(s) = 2s - \frac{\int_{\underline{s}}^{s} H(t)dt}{H(s)}$$

$$b_{\boldsymbol{\mathcal{Q}}}(s,\sigma) = s + \sigma - \frac{\int_{\underline{s}}^{s} H(t)dt}{H(s)}$$

These are the symmetric Bayes-Nash equilibrium actions when both agents believe $\boldsymbol{\mathcal{S}}$ and both agents believe $\boldsymbol{\mathcal{Q}}$, respectively. The first term represents the expected value of the firm for an agent with skill s who sees signal σ. The agent who believes $\boldsymbol{\mathcal{Q}}$ believes that her own Skill and the original firm Quality each play equal roles. The agent who believes $\boldsymbol{\mathcal{S}}$ believes instead that her own skill counts directly in the value of V, and indirectly, through its impact on Q. So investors who believe $\boldsymbol{\mathcal{S}}$ feel their own skill plays a larger role.

Some agents also lose money on average, if it turns out that one of the causal relations, $\boldsymbol{\mathcal{S}}$ or $\boldsymbol{\mathcal{Q}}$, is correct, and they are wrong about the model. If the true causal relation is $\boldsymbol{\mathcal{Q}}$, agents i who believe $\boldsymbol{\mathcal{S}}$ and whose skill s_i is sufficiently above the average skill ($\hat{s}$) lose money on average. These agents over-attribute the value of the firm to their own skill; hence it is those high skill CEOs who will suffer the curse. On the other hand, if it is in fact $\boldsymbol{\mathcal{S}}$ which is true, those agents who believe $\boldsymbol{\mathcal{Q}}$ and whose skill s_i is sufficiently below the average will lose money on average.

5 Discussion

In this section, I first discuss the evidence for causal modeling as a good framework for agents' mental models from the cognitive science literature. Second, I discuss the relationship of this framework to the first principles of rationality, and what they imply for the value of this framework. I then discuss what it means for causally coherent agents to learn.

Sloman and Lagnado (2004) provide an excellent overview of the relevant cognitive science literature. Cognition, they claim, depends on what does not change: the separation of items of interest from noise, and that "Causal structure is part of the fundamental cognitive machinery." One piece of evidence to support that claim is that causal relationships become independent of the data from which they are derived: they cite a case from Anderson, Lepper, and Ross (1980), in which "they presented participants with a pair of firefighters, one of whom was successful and who was classified as a risk taker, the other unsuccessful and risk averse. After explaining the correlation between performance as a firefighter and risk preference, participants were informed that an error had been made, that in fact the pairings had been reversed and the true correlation was opposite to that explained. Nevertheless, participants persevered in their beliefs; they continued to assert the relation they had causally explained regardless of the updated information. Causal beliefs shape our thinking to such an extent that they dominate thought and judgment even when they are

known to be divorced from observation." This provides evidence for the fact that humans tend to encode information as causal models, since that is what persists.

The evidence from cognitive science provides one reason to consider this framework; the other is first principles from rationality. Does rationality require that agents agree about plausible explanations?

Rationality is typically defined in the economic theory literature to be coherence between beliefs and behavior: it is *psychological* rationality. These are examples of psychological rationality: that agents have well-defined goals that they pursue single-mindedly, have preferences that are complete and transitive, or that they choose actions they believe will optimize a well-defined objective function. This is often understood to be what rationality means within the theory literature.

Logical rationality stands at odds with psychological rationality. An agent is logically rational when she is making what is objectively the best choice. An example of logical rationality is rational expectations (Muth 1961). An agent who forms rational expections not only has some coherent and reasonable model; she has the right model (i.e., the economist's model). Logical rationality is of the Popper model (Popper 1966) of situational analysis, as opposed to *psychologism* "the view that one can explain all social processes solely by reference to the psychological states of individuals (Langlois 2001)." Logical rationality is a common (sometimes implicit) definition of rationality outside of the theory literature.

The phrase "logically rational agents" is not well-defined. Logical rationality requires a correspondence between the agent and the world, and therefore knowing the agent alone (and her behavior, preferences, information, etc) is insufficient to determine whether she is logically rational. You have to know the workings of the world, too. This makes psychological rationality more satisfying, since, unlike logical rationality, psychological rationality has meaning with reference to the agent alone. This is the downside: psychological rationality is not sufficient to generate common sets of plausible explanations.

There has been an unhappy marriage between psychological and logical rationality, in which beliefs were required to be true, or, at least, the set of possible explanations that the agent considers was required to include the truth.

Causal coherence does not make that assumption. Causal coherence investigates the case of psychologically rational agents with logically irrational beliefs about the world.[11] These agents do not have a common prior over the set of theories, since they don't put positive probability on each other's theories.

How do these agents learn? Although it is not yet made explicit in this model, an agent with one causal relation $\boldsymbol{\mathcal{C}}$ over a phenomenon $\mathbb{V}$ has an associated set of possible explanations: namely, the set of possible theories $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}\right\}$, for all possible $\widehat{\boldsymbol{\Phi}}$. If one were estimating the following regression:

$$V = \alpha + \beta S$$

a similar set would be all possible values for (α, β). Standard Bayesian updating will eliminate possibilities

[11] See Hacking (1967) for some related issues regarding construction of reasonable beliefs.

(in the long run) as the $\widehat{\mathbf{\Phi}}$ which corresponds to F is mapped out. In that sense, these agents are standard Bayesian updaters.

5.1 Extensions

Here I describe three possible extensions of this work. The first extension would use causal models to explain apparent preference differences in a median voter setting. This may provide insights into endogenizing otherwise exogenous preference shocks. The second extension would construct agents who are ambiguity averse in the sense of Ellsberg (1961), who treat causal ambiguity in a manner similar to Gilboa and Schmeidler's (1989) Maxmin expected utility agents. The third would use this framework to construct agents who act in accordance with Quattrone and Tversky's (1984) empirical finding that people attribute causation to correlation. These agents could be used to derive economic implications.

Differing causal models of a common phenomenon, when the agents themselves cannot perform the experiment, may allow us to meaningfully discuss what might otherwise be exogenous preference shifts. Suppose voters in a median voter setting disagree about a tax policy. Perhaps some voters prefer a low tax and others a high tax. One may be able to rationalize their differing preferences as common preferences, but with differing causal models. For example, it could be the case that some voters believe education causes skill, and other believe education signals (is caused by) skill. This may explain apparent preference dispersion in local public finance models, and, in particular, provide insight into how preferences may change as government behavior changes.(Anderson and Pape 2006)

Causal ambiguity is a form of ambiguity or Knightian uncertainty. Ellsberg (1961) discussed a behavioral implication of ambiguity aversion. In the Ellsberg Urn Experiment, Ellsberg describes uncertainty over the relative number of green and blue balls in an urn (versus a known number of red). When an agent is called upon to bet on the color of the next ball, Ellsberg recommends reasonable choices that are inconsistent with expected utility. Gilboa and Schmeidler (1989) provide an axiomatic representation of utility, which yields behavior consistent with the Ellsberg's recommended choices in the Urn Experiment. This representation results in an agent with a set of priors about a distribution. For example, instead of believing there are exactly 50 green and 50 blue balls in the urn, the agent believes that there might be as few as 20 green balls and as many as 80: hence the agent believes that there is a set of possible distributions of balls in the urn. When the agent is called upon to place a bet on the color of the next ball, she evaluates her utility under each distribution and acts as if she believes the worst-case scenario were true. For example, called upon to bet that the next ball is green, she acts as if there were only 20 green balls; called upon to bet that the next ball is blue, she acts as if there were 80 green balls (and therefore only 20 blue ones.)[12] If the representation

[12]This is an informal treatment of Gilboa and Schmeidler's (1989) work. Gilboa and Schmeidler's (1989) representation theorem identifies the set of priors and utility jointly from behavior, so the minimum prior chosen is not identified as the *worst case* per se.

is extended to incorporate these kind of preferences, it may be possible to generate a set of causal models that the agent treats in a similar way to a set of priors.

Finally, here is evidence from the psychology literature that the lay person's understanding of causality is limited. Quattrone and Tversky (1984) showed "that people often fail to distinguish between causal contingencies (acts that produce an outcome) and diagnostic contingencies (acts that are merely correlated with an outcome.)" In other words, have a habit of attributing correlation to causation. This kind of causal modeling is appropriate for investigating the economic implications of those behavioral claims: by constructing an alternative to causal coherence, in which agents act as if the variable that they intervene on is the root of the causal structure. That would allow the development of agents who exhibit this kind of causal bias.

6 Conclusion

Considering the agents Sam (investor I$\boldsymbol{\mathcal{S}}$) and Quincy (investor I$\boldsymbol{\mathcal{Q}}$): I use the framework of causal bayesian networks to represent their models of an arbitrary phenomenon, and have investigated their behavior when they are endowed with a particular model. A set of reasonable models can be constructed that the agents might consider, given data they see. One can consider their behavior when they participate in an auction, where one of them will perhaps emerge cursed. One can see why and how much they will disagree on optimal choices when they are confronted with the same problem, even though they have the same unlimited and complete data.

With the axiomatic representation, I am able to construct the utility function and probability distributions that the agent believes her interventions will cause, based on observed choices between interventions and bets over outcomes.

I have used the causal bayesian network framework in a game-theoretic setting to define a causally coherent equilibrium. This has allowed me to describe their behavior in these interactive games. This causal ambiguity can arise with infinite data without missing variables. When considering agent choice when models are not identified, the problem is how to characterize a plausible, general, and tractable set of "reasonable models" for agents' conjectures: I have argued that this framework allows for a general way to characterize sets of theories that agents might believe and empirically identify those theories from the data the agents see.

These agents are Bayesians and can never transcend their initial endowments of possibilities as Bayesians regularly cannot. They are psychologically rational without being logically rational. This framework then provides an alternative to bounded (psychological) rationality models to handle these kinds of issues. I have described the distinction between psychological rationality and logical rationality. This setting provides a rich ground for extensions: applications to public finance, an opportunity to capture causal ambiguity

aversion, and to represent causal bias.

References

Anderson, C. A., M. R. Lepper, and L. Ross (1980): "The perseverance of social theories: The role of explanation in the persistence of discredited information," *Journal of Personality and Social Psychology*, 39, 1037–1049.

Anderson, N., and A. Pape (2006): "An Insurance Model of Property Tax Limitations," Working Paper.

Druzel, M., and H. Simon (1993): "Causality in Bayesian Belief Networks," *Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence*.

Ellsberg, D. (1961): "Risk, ambiguity, and the Savage axioms," *Quarterly Journal of Economics*, 75, 643–669.

Esponda, I. (2005): "Behavioral Equilibrium In Economies with Adverse Selection," *unpublished*.

Eyster, E., and M. Rabin (2005): "Cursed Equilibrium," *Econometrica*, 73(5), 1623–1672.

Fudenberg, D., and D. K. Levine (1993): "Self-Confirming Equilibrium," *Econometrica*, 61(3), 523–45.

Gilboa, I., and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

Hacking, I. (1967): "Slightly More Realistic Personal Probability," *Philosophy of Science*, 34(4), 311–325.

Heckman, J. (2005): "The Scientific View of Causality," University of Chicago, University of College London, and the American Bar Association.

Jeffrey, R. (1964): *The Logic of Decision*. University of Chicago Press.

Joyce, J. M. (1999): *The Foundations of Causal Decision Theory*. Cambridge and New York: Cambridge University Press.

Karni, E. (2005): "Subjective Expected Utility Theory without States of the World," JHU WP523.

Langlois, R. (2001): *International Encyclopedia of Business & Management, 2nd edition*.chap. Entry on "Rationality in Economics". London: Thompson International Publishers.

Muth, J. (1961): "Rational Expectations and the Theory of Price Movements," *Econometrica*, 29(3), 315–335.

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

Popper, K. (1966): *The Open Society and Its Enemies*, vol. II. Princeton: Princeton University Press., 2nd edn.

Quattrone, G., and A. Tversky (1984): "Causal versus diagnostic contingencies: on self-deception and on the Voter's illusion," *Journal of Personality and Social Psychology*, 46(2), 237.

Savage, L. J. (1954): *The Foundations of Statistics*. Wiley.

Sloman, S., and D. Lagnado (2004): *The Psychology of Learning and Motivation*vol. 44, chap. Causal Invariance in Reasoning and Learning. San Diego: Academic Press.

Spirtes, P., C. Glymour, and R. Scheines (1993): *Causation, Prediction, and Search*. New York: Springer-Verlag.

Verma, T., and J. Pearl (1990): "Equivalence and synthesis of causal models," *Proceedings of the 6th Conference on the Uncertainty in Artificial Intelligence*, pp. 220–227.

7 Appendix

7.1 Notes Concerning the Two-Price Auction

7.1.1 Causally Coherent Equilibrium Play

In this section I demonstrate that, for $0 < M \leq \frac{1}{2}\min\{\alpha, \beta\}$, the causally coherent equilibrium play for investor I$\boldsymbol{\mathcal{S}}$ is "Bid M iff $S_{\boldsymbol{\mathcal{S}}} = 1$." Then I demonstrate that Investor I$\boldsymbol{\mathcal{Q}}$ plays M only if $S_{\boldsymbol{\mathcal{Q}}} = 1$ and $\sigma = 1$.

First, consider the payoffs for any low skill agent. This agent stands to win 0 under bid \$0 and $-M$ with some positive probability under bid $\$M$. Trivially, low skill agents bid \$0.

Now consider the high-skill investor I$\boldsymbol{\mathcal{S}}$. He has sufficient incentive to play $\$M$ iff:

$$PO_{\boldsymbol{\mathcal{S}}}(M) \geq PO_{\boldsymbol{\mathcal{S}}}(0) \tag{1}$$

$$Prob_{\boldsymbol{\mathcal{S}}}(win|M)\alpha - M \geq Prob_{\boldsymbol{\mathcal{S}}}(win|0)\alpha \tag{2}$$

$$\left(\frac{1}{2}\gamma_s + (1-\gamma_s)\right)\alpha - M \geq (1-\gamma_s)\frac{1}{2}\alpha \tag{3}$$

where γ_s is the probability that (he believes) his opponent plays M

$$\tag{4}$$

$$\frac{1}{2}\alpha \geq M \tag{5}$$

Consider the high-skill investor I$\boldsymbol{\mathcal{Q}}$. He has sufficient incentive to play M iff:

$$PO_{\boldsymbol{\mathcal{Q}}}(M) \geq PO_{\boldsymbol{\mathcal{Q}}}(0) \tag{6}$$

$$Prob_{\boldsymbol{\mathcal{Q}}}(win|M)\beta - M \geq Prob_{\boldsymbol{\mathcal{Q}}}(win|0)\beta \tag{7}$$

$$\left(\frac{1}{2}\gamma_q + (1-\gamma_q)\right)\beta - M \geq (1-\gamma_q)\frac{1}{2}\beta \tag{8}$$

where γ_q is the probability that (he believes) his opponent plays M

$$\tag{9}$$

$$\frac{1}{2}\beta \geq M \tag{10}$$

7.1.2 That investor I$\mathcal{S}$ loses money on average

We must establish the probability that $Q = 1$ given that S won, when $\boldsymbol{\mathcal{Q}}$ is true.

$$\begin{aligned} Prob(Q=1|Swon) = \frac{1}{2}&Prob(Q=1|\text{investor I}\boldsymbol{\mathcal{Q}}\text{ plays }M)Prob(\text{investor I}\boldsymbol{\mathcal{Q}}\text{ plays }M) \\ &+ Prob(q=1|\text{investor I}\boldsymbol{\mathcal{Q}}\text{ plays }0)Prob(\text{investor I}\boldsymbol{\mathcal{Q}}\text{ plays }0) \end{aligned}$$

$Prob(Q = 1|\text{investor I}\boldsymbol{\mathcal{Q}} \text{ plays } M)$ is β. $Prob(\text{investor I}\boldsymbol{\mathcal{Q}} \text{ plays } M) = \frac{1}{4}$; there is a half chance that the agent is of type $S_{\boldsymbol{\mathcal{Q}}} = 1$, and a half chance that the firm receives a signal of 1. Therefore,

$$\begin{aligned} Prob(Q = 1|Swon) &= \frac{1}{2}\beta\frac{1}{4} + (1-\beta)\frac{3}{4} \\ &= \frac{3}{4} - \frac{5}{8}\beta < M \end{aligned}$$

when $\frac{2}{3} < \beta$.