Causal Coherence[*]

Job Market Paper

Andreas Duus Pape[†]

December 20, 2006

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

In this section I discuss in detail the structure of causal bayesian networks as a decision-making structure. In the previous section I introduced the investors Sam and Quincy, who were equally good at predicting stocks but differed in choices when taking over a company because their models differed. It might be said that Sam and Quincy's models of how firms work are not identified in a statistical sense.

If we consider agent choice in a setting where models are not identified, we are confronted with a problem: how to construct a set of reasonable models that an agent might believe? Now, typically in economics we construct a context-specific set of reasonable models for each setting. For example, there might be one missing parameter that agents do not know. As an alternative, I use this causal modeling framework, which provides a tractable, well-defined set of reasonable models for a more general setting. This framework is used in both artificial intelligence and statistics. The set of models generated by this framework is plausible from a scientific standpoint, and has support from psychology and cognitive science. Hence it is a more general methodology for generating a "set of reasonable models."

First I present the reasoning framework, then discuss the implications for decision-making. 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]." Causal bayesian networks imagines variables as nodes connected by causal arrows.

The set of reasonable models as I use it here is ultimately defined by agents' disagreements over causal direction, i.e. which way the causal arrows point. The purpose is to highlight the kind of disagreements that can emerge among agents who have the same information (which amounts to an infinite data set) and are not plagued with problems of missing relevant variables.[1]

The structure, briefly. There is a set $\mathbb{V}$ of variables, called a *phenomenon*, which is the object of causal confusion. There is a network of directed arrows among these variables; the directed arrows represent causal relationships. This directed network is endowed with *causal probability functions.* The causal probability function assigns, when the causes $\mathbb{X}$ of variable Y have values $\vec{x}$, a distribution to Y. The directed network is called a *causal relation*, and when endowed with causal probability functions, it is a *causal relation with parameters.* The causal relation with parameters is identically a causal Bayesian network. The causal relation with parameters allows one to forecast the effect of changing a variable in a system. Changing a variable will be called an *intervention*, and is represented by breaking existing causal relationships and replacing them with the exogenous force. The causal relation is assumed to be acyclic. When that is the case, and there is an observed joint distribution F on $\mathbb{V}$, then a unique set of causal probability functions can be assigned to a causal relation. This calibrates the model to data. It is also the case that a causal relation with parameters generates a unique data set.

In the following subsection, I begin discussing the *phenomenon.*

2.1 What is the object of causal confusion?

The investors above, Sam and Quincy, differ in their causal models about how the value of firms is generated. In the general framework, there is an object about which these agents have models: a phenomenon in the world that is the object of causal confusion. In the case of the investors, that phenomenon is firms. A phenomenon is the bundle of characteristics of the object in question with an associated probability distribution over those characteristics.

[1] The framework provided by Spirtes, Glymour, and Scheines (1993) and Pearl (2000) provide for these cases.

Formally, a *phenomenon* is a collection of variables that has an associated joint distribution:

Definition 1. *A phenomenon $\mathbb{V}$ is a set of variables V with an associated distribution F over $\mathbb{V}$.*

In the case of the value of firms, $\mathbb{V}_{firms}$ might be $\{\text{CEO skill}, \text{firm quality}, \text{firm value}\}$.

As usual, a variable V also has defined an associated support. I suppose throughout that the support is finite. The support defines the event space associated with a set of variables $\mathbb{W}$:

$$\text{supp}(\mathbb{W}) \quad = \quad \prod_{W \in \mathbb{W}} \text{supp}(W)$$

The phenomenon represents the system in equilibrium. There is an underlying process which generates the phenomenon (that is, produces F), and each combination of characteristics occurs with some well-defined and observable frequency. In that sense, the phenomenon is like a data set.

The support space of phenomenon is similar to the state space of a decision problem in classic decision theory (Savage 1954). The elements of a phenomenon's support space do not, however, specify the causal model.

Here I describe the distinction between the state space of a decision problem and the support space of a phenomenon.

Decision problems are typically defined such that: (1) there is a state space Ω is exogenous to the agent, (2) the agent has a set A of actions, from which she chooses an action $a \in A$, and (3), the action a and the true $\omega \in \Omega$ define a payoff $\pi \in \Pi$. (Thus the payoff function maps A and Ω into distributions over Π). By analogy, one could say that Ω is like supp($\mathbb{V}$); all possible combinations of characteristics is the state space. For example, if $\mathbb{V}$ were $\{X, Y\}$, where X and Y were binary variables, then the state space Ω and the space supp($\mathbb{V}$) would both be:

$$\{XY, X\neg Y, \neg XY, \neg X\neg Y\}$$

In the classic decision problem, the agent believes there is a distribution F over Ω. When the agent choses her optimal action a, knowledge of F is all she needs: she is only interested in the likelihood of $\omega \in \Omega$, since that is what will determine her utility when combined with a. That is, she chooses a to maximize

$$\sum_{\Omega} u(\omega, a) F(\omega | a)$$

For problems of causal ambiguity, that is not the case. In this set-up, the action interacts with the space supp($\mathbb{V}$). Here, the actions A are choices of values of variables in $\mathbb{V}$. To maintain the analogy, one might say that A is $\{X, \neg X\}$; that is, that the agent can choose whether X is true or false. To extend the analogy, one might think that the agent's problem is to choose $x \in \{X, \neg X\}$ to maximize the expected utility, given x. That is:

$$Eu(x) = u(x, Y) * F(Y|x) + u(x, \neg Y) * F(\neg Y|x)$$

This is, however, false. Eu is referred to in the causal decision literature (Joyce 1999) as the news value of x; that is, the utility one would expect upon witnessing x. The news value is the received utility only if X causes Y. If X has no effect on Y, then the following will be the expected utility upon doing x:

$$Eu'(x) = u(x, Y) * F(Y) + u(x, \neg Y) * F(\neg Y)$$

There is no conditional distribution because the choice of x does not affect Y, hence does not change its probability.

In this section, I introduced the notion of the phenomenon, which is similar to the state space in the standard decision problem. However, it differs, because the phenomenon is not only states of the decision problem; it also defines the actions of the decision problem.

2.2 When is causal confusion made manifest?

Causal confusion is made manifest when the agent intervenes on an otherwise stable system described by the phenomenon. In those times the break is made between correlation and causation. The agent's understanding of the phenomenon must be augmented in order to consider what happens in that case, since the phenomenon and associated distribution are insufficient.

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.[2]

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 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.[3]

The causal effect of a intervention is (1) what changes and (2) how much.

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

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

The causal decision theory literature represents the causal effect of an action as a *causal probability function*. Recall the standard decision problem, in which agents choose actions $a \in A$ and receive distributions over payoffs $\pi \in \Pi$. The mapping from A to distributions over Π is a causal probability function: for each action $a \in A$, the causal probability function assigns a distribution over Π. Causal probability functions are defined from actions to outcomes. Recall that in this framework actions are *do* functions on selections of variables, so the same causal probability function can, and should, represent both *actions* and *changes in variables*, so it is appropriate to extend this beyond the causal effect of actions, to the causal effect of any variable (Joyce 1999).

So suppose that $\mathbb{W}$ causes V, and the action chosen is $do(\mathbb{W})$. Then:

Definition 3. *Φ_V is a causal probability function of V if $\Phi_V : supp(\mathbb{W}) \rightarrow \Delta_V$, where Δ_V is the set of all distributions over $Support(V)$, and $\mathbb{W}$ is some subset of $\mathbb{V}$.*

For any set of values $\vec{w} \in \text{supp}(\mathbb{W})$, a distribution is assigned to V.[4]

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 $do(Lottery)$, a distribution over all possible dollar winnings.

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.

[4] I abuse notation slightly by supposing that, since $\Phi_V(do(\mathbb{W}))$ assigns a distribution over V, that $\Phi_V(v|do(\mathbb{W}))$ is that distribution (note the v).

Above, I stated how the agent might choose $x \in \{X, \neg X\}$ to maximize

$$u(x, Y) * F(Y|x) + u(x, \neg Y) * F(\neg Y|x)$$

or

$$u(x, Y) * F(Y) + u(x, \neg Y) * F(\neg Y)$$

Depending on the causality of the system. Supposing the causal probability function is known, the ambiguity is resolved. The agent chooses x to maximize

$$u(x, Y) * \Phi_Y(Y|x) + u(x, \neg Y) * \Phi_Y(\neg Y|x)$$

Causal probability functions capture the magnitude of causality; they assign, for each set of causes, a distribution over the caused variable. Causal probability functions therefore allows one to describe the effect of an intervention in a phenomenon.

2.3 Constructing a mental model of interventions that accounts for agreement of information

A causal probability function for each variable forecasts what would happen to the phenomenon under any intervention (Pearl 2000). So an agent's complete mental model of a phenomenon consists of a causal probability function for each variable in that phenomenon. This structure captures the agent's sense of causal direction with a directed graph, which can be augmented with causal probability functions to capture "...and how much?" After all, two agents might agree that X causes Y, but disagree how much Y changes.

Pearl and Verma (1991) call the set of causal directions a "causal model" and the augmented causal model a "causal theory." Since I would like to use "model" and "theory" more generally, I define a "causal relation" for the set of directions of causality and "causal relation with parameters" for the relation augmented with causal probability functions. Intuitively, the causal relation represents a theory about how the phenomenon works and the causal relation with parameters represents that theory calibrated to data.[5]

A causal relationship has a cause and an effect. So a causal relation over a phenomenon $\mathbb{V}$ needs to capture that the relationship is directed. Binary relations capture directed relationships succinctly.

Definition 4. *A* **causal relation $\mathcal{C}$** *is a binary relation among variables in* $\mathbb{V}$*.*

Figure 1 depicts some causal relationships. There, X causes Y, Y causes Z. It is also the case that X causes Z, but only through Y. (If you found this in your data, you would say that X is a good instrument.)

[5] Barring these vocabulary differences, this structure follows Pearl and Verma closely (see Pearl and Verma (1991) and Pearl (2000)).

The causal relation captures only those direct causal relationships in the data. Namely, figure 1 depicts the causal relation $\boldsymbol{\mathcal{C}} = \{(X, Y), (Y, Z)\}$ where, for all other possible combinations, not $\boldsymbol{\mathcal{C}}$. In particular, not $X\boldsymbol{\mathcal{C}}Z$.

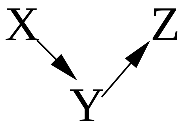


Figure 1: $X\boldsymbol{\mathcal{C}}Z$ and $Y\boldsymbol{\mathcal{C}}Z$, but not $X\boldsymbol{\mathcal{C}}Z$

Pearl defines objects like "the set of all variables in $\mathbb{V}$ that cause variable X." Following Pearl, a natural way to do this is to define the standard familial relationships of parent, child, ancestor, and descendent. They are defined in the obvious way. Consider again figure 1. When X causes Z, but only through Y, one would say that X is a $\boldsymbol{\mathcal{C}}$-causal parent of Y: there is a direct causal relationship between X and Y, and it runs from X to Y. One would say that X is a $\boldsymbol{\mathcal{C}}$-ancestor of Z, because there is a string of relationships from X to Z. Similarly, Z is a $\boldsymbol{\mathcal{C}}$-child of Y and a $\boldsymbol{\mathcal{C}}$-descendent of X. Formally:

Definition 5. *$W \in \mathbb{V}$ is a* **$\boldsymbol{\mathcal{C}}$-parent** *of $V \in \mathbb{V}$ if $W\boldsymbol{\mathcal{C}}V$. $W \in \mathbb{V}$ is a* **$\boldsymbol{\mathcal{C}}$-ancestor** *of $V \in \mathbb{V}$ if there exists some set of variables $W_1, \ldots, W_n \in \mathbb{V}$ such that*

$$W\boldsymbol{\mathcal{C}}W_1\boldsymbol{\mathcal{C}}\ldots\boldsymbol{\mathcal{C}}W_n\boldsymbol{\mathcal{C}}V$$

V a **$\boldsymbol{\mathcal{C}}$-child** *of W if W is a $\boldsymbol{\mathcal{C}}$-parent of V. V a* **$\boldsymbol{\mathcal{C}}$-descendent** *of W if W is a $\boldsymbol{\mathcal{C}}$-ancestor of V.*

I define $Pa_{\boldsymbol{\mathcal{C}}}(V)$ be the (possibly empty) set of all $\boldsymbol{\mathcal{C}}$-causal parents of V.[6]

Figure 1 depicts a directed graph. It can be thought of as a useful graphical depiction of a causal relation: it contains exactly the same information.

Definition 6. *A* ***directed graph*** *is a collection of points ("nodes") and lines with arrowheads ("edges") connecting some (possibly empty) subset of the nodes.*

A causal relation $\boldsymbol{\mathcal{C}}$ contains all the information required to determine what variables will change under an intervention. Consider once again figure 1. If the agent were to intervene and change X, one would expect both Y and Z to change. If one were to change Y, then one would expect only Z to change. In other words

Definition 7. *The* **$\boldsymbol{\mathcal{C}}$-causal effect of a intervention on variables** *$\mathbb{W}$ is the set of all $\boldsymbol{\mathcal{C}}$-descendants of any variable in $\mathbb{W}$.*

[6] Other traditional causality terms can be defined in this context: for example, if for all $W \in \mathbb{V}$, not $W\boldsymbol{\mathcal{C}}V$, then V is called $\boldsymbol{\mathcal{C}}$-exogenous. Similarly, if there exists $W \in \mathbb{V}$ such that $W\boldsymbol{\mathcal{C}}V$, then V is called $\boldsymbol{\mathcal{C}}$-endogenous.

The question of "how much" will be answered by assigning to each variable in $\mathbb{V}$ a causal probability function. Each variable needs exactly one causal probability function. The distribution on V is fully determined when one observes all the $\boldsymbol{\mathcal{C}}$-causal parents of V. A causal probability function associated with $\boldsymbol{\mathcal{C}}$ is called a $\boldsymbol{\mathcal{C}}$-causal probability function.

Definition 8. *A $\boldsymbol{\mathcal{C}}$-causal probability function is a causal probability function $\Phi_V : supp(Pa(V)) \to \Delta_V$. $Pa(V)$ is, as defined above, the set of $\boldsymbol{\mathcal{C}}$-causal parents of V.*

So now there is a complete representation of the agent's mental model of a phenomenon: a causal relation $\boldsymbol{\mathcal{C}}$ and a set of causal relations, one for each variable, which map from the parents of that variable to that variable. This is a causal relation with parameters:

Definition 9. *A* **causal relation with parameters** *is a set $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}\right\}$, where $\widehat{\boldsymbol{\Phi}}$ is a set of $\boldsymbol{\mathcal{C}}$-causal probability functions Φ_V, one for each $V \in \mathbb{V}$.*

Recall the domain of $\boldsymbol{\mathcal{C}}$-causal probability function is values of $\boldsymbol{\mathcal{C}}$-causal parents of the variable, so each of the causal probability functions in $\widehat{\boldsymbol{\Phi}}$ captures the mapping from $\boldsymbol{\mathcal{C}}$-causal parents to distributions over the $\boldsymbol{\mathcal{C}}$-causal child.

An example.

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. Consider the problem of: Investors are trying to forecast what will happen to the value of this company when she takes over.

Here is the mapping. Firms are defined by three 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");

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

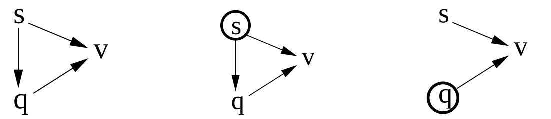
So the phenomenon is $\mathbb{V} = \{S, Q, V\}$. Investors are then interested in the distribution of V under the action $do(S)$.

There are various plausible theories that investors may have. One plausible theory 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 relation $\boldsymbol{\mathcal{S}}$ which represents that theory is:

$$S\boldsymbol{\mathcal{S}}Q \qquad\qquad Q\boldsymbol{\mathcal{S}}V \qquad\qquad S\boldsymbol{\mathcal{S}}V$$

Causal relation $\boldsymbol{\mathcal{S}}$ is depicted in figure 2(a).

$\boldsymbol{\mathcal{S}}$ 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. Suppose investors know the skill of the new



(a) Without Intervention.

(b) An Intervention on S

(c) An Intervention on Q

Figure 2: Directed Acyclic Graphs representing $\boldsymbol{\mathcal{S}}$

CEO is some level s_c. Then the investors are trying to forecast the effect on V of $do(S = s_c)$. Figure 2(b) represents that intervention: one would expect both Q and V to change. 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 2(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.

The final piece of this analysis is to judge "how much" variables change under an intervention. This structure is not well-defined, to produce effects of interventions or otherwise, unless the causal relation $\boldsymbol{\mathcal{C}}$ is *acyclic*. This means that: for all $V \in \mathbb{V}$, V is not a $\boldsymbol{\mathcal{C}}$-causal ancestor of V. It is interesting and probably realistic to consider models with cycles[7] but it is omitted from the inquiry in this paper.

Suppose that $\mathbb{V}$ has the values $\vec{v}$. Now suppose an agent manipulates variables $\mathbb{W}$ to $\vec{w}$. Then, as stated above, only those variables $\mathbb{Y}$ who are descendants of at least one variable in $\mathbb{W}$ change. What is the resulting distribution of the intervention $do(\mathbb{W} = \vec{w})$?

All variables in $\mathbb{Y} \cap \mathbb{W}$—that is, variables that are manipulated in the intervention and children of other manipulated variables—have an atomic distribution, because the value of those variables do not change. Hence, for all $Y \in \mathbb{Y} \cap \mathbb{W}$, the value $y_Y = w_Y \in \vec{w}$.

The nontrivial distribution is on the set $\mathbb{Y} \backslash \mathbb{W}$. Let us construct the set of values $\vec{m}$. $\vec{m}$ contains one value for each variable in $\mathbb{V}$. The entry m_V in $\vec{m}$ corresponds to variable V. Its value is:

1. $m_V = v$, an unspecified value, if $V \in \mathbb{Y} \backslash \mathbb{W}$,

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

3. $m_V = v_V \in \vec{v}$ otherwise.

Then the distribution over $\mathbb{Y} \backslash \mathbb{W}$ is constructed in the following way: for every non-intervention variable that is not caused by an intervention variable, the causal probability functions are multiplied, substituting in the values of those variables. Consider the causal probability functions as conditional distributions. Then

[7] See Pearl (2000) for some alternative views on this.

the joint distribution is their product. Formally:

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

The probability of a selection of variables, conditional on a particular variable, can be constructed by chaining the relevant conditional distributions. For example,

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

Returning to the example the firm under $do(S = s_c)$: Knowing that Q and V 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.

The distribution over V is the object of interest, however. Here, there is a direct effect through the fact that $S\boldsymbol{\mathcal{S}}V$, and an indirect effect, from the fact that $S\boldsymbol{\mathcal{S}}Q\boldsymbol{\mathcal{S}}V$.

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

Suppose the initial values of $\mathbb{V}$, before intervention, are $\vec{v} = \{s_k, q_k, v_k\}$.

Consider the intervention $\mathbb{W}' = \{Q\}$, where $\vec{w}' = \{q'\}$. Then 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:

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

The causal relation is a set of arrows among variables that describes the directions of causality among those variables. The causal relation represents an agent's mental model of a phenomenon. The causal relation augmented with causal probability functions, one for each variable in the phenomenon, is called a causal relation with parameters. A causal relation with parameters can be used to determine the distribution of changed variables after an intervention.

2.4 Describing an agent who behaves according to a causal model

Suppose agents have a set of knowledge K_i.

Definition 10. *A knowledge set K_i is a set of statements about the world that the agent believes to be true about the structure of the game, about the payoffs involved, about the number and nature of the other players, etc.*

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 2(a) and discussed in the previous section). Taking her causal model to be correct, she acts rationally. This is defined as causal coherence.

Definition 11. *An agent i is* **causally coherent with** $\left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i\right\}$ *if she behaves rationally given some K_i, where*

$$K_i = \left\{\boldsymbol{\mathcal{C}}_i, \widehat{\boldsymbol{\Phi}}_i, \ldots\right\}$$

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

There is a one-to-one mapping in the causal bayesian network literature between causal relations with parameters $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}\right\}$ and observed distributions F. As described in (Pearl 2000):

> $\left\{\boldsymbol{\mathcal{C}}, \widehat{\boldsymbol{\Phi}}\right\}$ defines a unique distribution F over supp($\mathbb{V}$), where:
>
> 1. $F(\mathbb{V} = \vec{v}) = \prod_{V \in \mathbb{V}} \Phi_V(v | pa_{\boldsymbol{\mathcal{C}}}(v))$,
> 2. $pa_{\boldsymbol{\mathcal{C}}}(V)$ be an associated instance of $Pa_{\boldsymbol{\mathcal{C}}}(V)$; i.e., $pa_{\boldsymbol{\mathcal{C}}}(V) \in \text{supp}(Pa_{\boldsymbol{\mathcal{C}}}(V))$

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.

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 relation to data, or, calibrating the causal model. Namely, suppose that the data F is observed. Now for each Φ_V, assign the following:

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

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 12. *An agent i is* **causally coherent with data** F *if she is causally coherent with a knowledge set K_i, which includes:*

1. *A causal relation* $\boldsymbol{\mathcal{C}}_i$
2. *The joint distribution* F
3. $\widehat{\mathbf{\Phi}}_i$*, which 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 $(\boldsymbol{\mathcal{S}}, \widehat{\mathbf{\Phi}})$ 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 $(\boldsymbol{\mathcal{Q}}, \widehat{\mathbf{\Phi}}')$ fits the data equally well.

A one-to-one mapping exists between, on one hand, a causal relation with parameters, and, on the other hand, a causal relation and a joint distribution F over the phenomenon. This means a causally coherent agent with causal relation $\boldsymbol{\mathcal{C}}$ can calibrate her causal relation to a data set, represented by F, and produce a well-defined set of parameters $\widehat{\mathbf{\Phi}}$. A causally coherent agent is rational given her causal relation with parameters, and can represent an agent who confuses evidence consistent with a theory with evidence for that theory.

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.

Causally coherent agents believe their own model and are unaware of alternative models. They are unaware in the sense of Dekel, Lipman, and Rustichini (1998). That is to say, these agents are not aware of other models, and are not aware that they are unaware of them. This unawareness also means that they are unaware of the possibility that others may believe these other models. Dekel, Lipman, and Rustichini show that non-trivial unawareness is incompatible with the agent having a well-defined partition of a state space (which includes events of which he is unaware) and is therefore incompatible with the standard model of knowledge.

Since the state space is incompatible with this kind of unawareness, it seems appropriate to consider a representation of the agent without reference to a state space. Karni (2005) provides a framework in which a subjective expected utility representation emerges without reference to a state space. Suppose a homeowner is worried about protecting his property. There is a set A of *actions*, for example, the homeowner may choose to buy an alarm system or install smoke detectors. There is a set Θ of *effects*; the possible states that the house may be in after some possible calamity (note that the states of the house are not states of the world—the probability of these states is affected by the action.) Bets are defined on effects, in this case, homeowner's insurance that pays off different amounts depending on the state of the house.

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 (a) interventions $do(S = s), do(Q = q)$ and (b) 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 axioms from Karni's (2005) representation, plus one additional axiom which imposes the intervention structure, then a representation emerges. This representation specifies

1. a utility for each (s, q) variable realization pair and money (the bets), unique up to a positive, affine transformation, and
2. A unique set of probability distribution that each intervention induces on the other variable.

In this section, I describe the setting and introduce the axioms of behavior on the choices. I describe the implications for my additional axiom. Then I introduce the representation adapted from Karni. Then I describe how to evaluate whether the agent's behavior is causally coherent.

The phenomenon $\mathbb{V} = \{S, Q\}$. Support of S and Q are finite. X and Y will be used to refer to arbitrary variables. The set of effects, that is, possible outcomes for the intervention acts, is

$$\Theta = (\text{supp}(S) \times \text{supp}(Q))$$

with arbitrary elements (s, q).

As usual, $do(X = x)$ is the intervention action which sets variable X to some value x. When it will not cause confusion, $do(x)$ will represent $do(X = x)$. The intervention induces, per a causal probability function, a distribution on the other variable. For example,

$$do(S = s) : S \to \Delta(Q)$$

where $\Delta(Q)$ is a set of distributions over Q. Each choice s of S induces a distribution on Q. $\mathbb{I}_X$ is the set of intervention acts on the variable X.

$$\mathbb{I}_X = \{do(X = x) | x \in \text{supp}(X)\}$$

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. The set of all intervention acts will be denoted $\mathbb{I} = \mathbb{I}_S \cup \mathbb{I}_Q \cup do(\emptyset)$.

Note that, for any intervention, some effects are impossible. For example, under $do(S = s_1)$, (s_2, q_1) is impossible, since the act $do(S = s_1)$ sets S to s_1 instead of s_2. This will be addressed under axiom A5 below, where such impossible states will be rendered as null events. I will demonstrate that this systematic application of null events does not upset the structure required for the representation.

The decision maker will be called to make choices among pairs of intervention acts and bets over outcomes. Bets over outcomes will be defined as $b(s, q)$, which is a real-valued payoff whenever the variable $S = s$ and $Q = q$. That is,

$$b : S \times Q \to \mathbb{R}$$

Let $\mathbb{B}$ be the set of all such bets. Following Karni, define $(b_{-(s,q)}, r)$ as the bet which awards b in all effects $(s', q') \neq (s, q)$, and awards r in effect (s, q). In other words, it is the bet b, with the (s, q)th entry replaced with r.

The decision maker is faced with (intervention, bet) pair choices. Therefore, her choice set is

$$\mathbb{C} = \mathbb{I} \times \mathbb{B}$$

with arbitrary elements $(do(X = x), b)$.

Here is an example of an intervention act which results in a bet. Suppose the intervention act $do(Q = q')$ is performed, and a distribution is assigned to S. That distribution yields a specific realization $S = s'$. Hence the effect is is the vector (s', q'). Under bet b', the prize $b'(s', q')$ will be awarded, and the agent will experience the prize $b'(s', q')$ and the resulting effect (s', q').

The outcomes $\mathbb{O}$ over which agents have final utility is $\mathbb{O} = \Theta \times \mathbb{R}$, with arbitrary elements $((s, q), b(s, q))$. The decision maker expresses preferences over *choices* from the choice set $\mathbb{C}$, and these preferences are observable by the modeler. Under the behavioral axioms below, a utility representation emerges over $\mathbb{O}$, that is, (effect, money) pairs $((s, q), b(s, q))$.

Before introducing the behavioral axioms, some more terms must be defined. Karni defines a constant-valuation bet, a null event, the set of effects which are non-null for a given action, and linked events. This allows us to define Karni's Axiom A0, a background assumption in his model.

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 give the same final utility 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.)

An effect (s,q) is null given $do(x)$ if $(do(x),(b_{-(s,q)},r)) \sim (do(x),(b_{-(s,q)},r')) \quad \forall r,r' \in \mathbb{R}$. (This is the standard definition.) $\Theta(do(x),\succsim) \subseteq \Theta$ are the set of effects that are nonnull given $do(x)$. Two effects $(s,q),(s',q')$ are then *elementarily linked* if there exists actions $do(x),do(x')$ such that $(s,q),(s',q') \in \Theta(do(x),\succsim) \cap \Theta(do(x'),\succsim)$. And two events $(s,q),(s',q')$ are *linked* if there are a sequence of events, such that each is linked to its neighbor, and the first is linked to (s,q) and the last linked to (s',q').[8] 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.

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 $(do(x),b) \sim \mathbb{C}$.*

Here are the behavioral axioms. Axioms 1-3 are from Karni (2005). The first two are standard. The third is discussed at length in Karni and introduced there. Karni has an additional axiom 4 which provides a special case which I do not consider here. Additionally, I introduce an axiom A5, which induces the causal structure. In addition, there is Axiom 0.[9]

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.*

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

Karni (2005) explains:[10] "To grasp the meaning of action-independent betting preferences, think of the preferences $(do(x),(b_{-\theta},r)) \succsim (do(x),(b'_{-\theta},r'))$ and $(do(x),(b'_{-\theta},r'')) \succsim (do(x),(b_{-\theta},r'''))$ as indictating that, given action $do(x)$ 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 $do(x')$ instead of $do(x)$."

Define $\Theta_{X=x}$ as those effects that are consistent with variable X having value x. Namely, $\Theta_{S=s} = \{(s,q)|q \in \text{supp}(Q)\}$ and $\Theta_{Q=q} = \{(s,q)|s \in \text{supp}(S)\}$.

[8] See (Karni 2005) for details on the meaning of these concepts. I only define them to demonstrate how the requirements of Karni's theorem are met in this case.

[9] The assumptions are modified to have notation consistent with my model.

[10] 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 $do(S = s_1)$. Then this axiom requires that it doesn't matter what the rewards are in any effect $(s_2, q) \in \Theta - \Theta_{S=s_1}$, where $s_2 \neq s_1$. The agent knows with certainty that those effects (s_2, q) will never occur. Hence changing the rewards on those effects should do nothing to change preference.

Then the following is true. It is the Karni representation theorem, with the additional causal structure, which appears in statement 1.

Theorem 1. *(Karni) 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-A3 and A5*
    (b) *There exists*
        i. *a continuous function* $u : \mathbb{O} \to \mathbb{R}$*,*
        ii. *a family of probability measures* $\{\rho(s, q|do(Y = y))\}$ *on supp*$(S) \times$*supp*(Q)*, for* $Y = S, Q$*, and* $\rho(s, q|do(\emptyset))$*, and*
        iii. *a family of continuous, increasing functions* $\left\{f_{do(x)}\right\}_{do(x)\in\mathbb{I}}$*,*
        such that, for all $(do(X = x), b)$*,* $(do(Z = z), b') \in \mathbb{C}$*,*

$$(do(x), b) \succsim (do(y'), b') \iff f_{do(x)}\left(\sum_{\theta_{w,x}\in\Theta} u(\theta_{w,x}; b(\theta_{w,x}))\rho(\theta_{w,x}|do(x))\right) \geq f_{do(y')}\left(\sum_{\theta_{y,z}\in\Theta} u(\theta_{y,z}; b'(\theta_{y,z}))\rho(\theta_{y,z}|do(z))\right) \tag{1}$$

where

$$\theta_{\mathbb{X}=x,\mathbb{Y}=y} = \begin{cases} (x, y) & \textit{if } \mathbb{X} = \mathbb{S} \\ (y, x) & \textit{if } \mathbb{X} = \mathbb{Q} \end{cases} \tag{2}$$

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

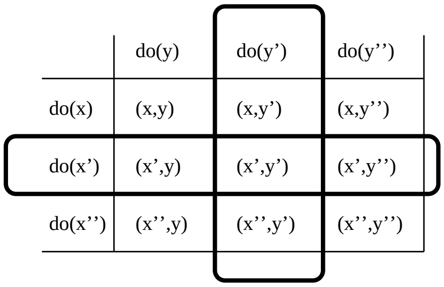
3. *For each* $do(x) \in \mathbb{I}$*,* $\rho(\theta_{x,y}|do(x))$ *is unique and* $\rho(\theta_{x,y}|do(x)) = 0$ *if and only if* (x, y) *is null given* $do(x)$*, so* $\rho(\theta_{x',y}|do(x)) = 0, \forall x' \neq x$*.*

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 $\rho(s', q; do(s))$ is null for all $s' \neq s$. This means that $\rho(s, q; do(s))$ can be interpreted as the causal probability function on q of $do(s)$).

Here I explain why axiom A5 is not inconsistent with axioms A0-A3:

Axiom A5 is trivially not inconsistent with A1 and A2 (weak order and continuity.) Since the implication of A3 is $\succsim$, and A5 only requires some indifference, then A5 is also trivially not inconsistent with A3. There is also trivially no inconsistency between A5 and the constant-valuation bets requirements of Axiom A0.

The only point of potential inconsistency is between A5 and A0's requirement that all effects are linked. Recall, effects are elementarily linked when they are non-null for the same two actions, and two effects (s, q) and (s', q') are linked if there is a sequence of linked events connecting (s, q) to (s', q'). 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 3: No effects are elementarily linked

As figure 3 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 events consistent with the axiom A5, there are too many null events to have consistency.

However, with the non-intervention act, then many effects might be elementarily linked. For example, (x', y') and (x', y'') are elementarily linked:$(x', y'), (x', y'') \in \Theta(do(x'), \succsim) \cap \Theta(do(\emptyset), \succsim)$. With this method, every two events which vary in only one coordinate can be elementarily linked (and therefore linked). Since any two events which vary in only one coordinate can be elementarily linked, than any two events $(x, y), (x'y')$ can be linked: (x, y) is elementarily linked to (x, y') is elementarily linked to (x', y').

Now I can state what it means for an agent to be causally coherent from observed choices.

Claim 1. *An agent is causally coherent with a model $\boldsymbol{\mathcal{C}}$: $X\boldsymbol{\mathcal{C}}Y$ and data $\rho(x, y|do(\emptyset))$ iff:*

1. $\rho(\theta_{x,y}|do(y)) = \rho(\theta_{x,y'}|do(y')) \quad \forall y \in supp(Y), and$

2. $\rho(x, y|do(\emptyset)) = \rho(\theta_{x,y}|do(x))\rho(x), \quad \forall x \in supp(X), y \in supp(Y)$

The first item is the requirement of *causal irreversibility*, that if X causes Y, then changing Y does not change X. The second requirement is of *causal model consistency*, that the distribution under the non-intervention act, $\rho(x, y|do(\emptyset))$, is consistent with the causal probability functions which form that distribution of the data, $\rho(y|do(x))\rho(x)$.

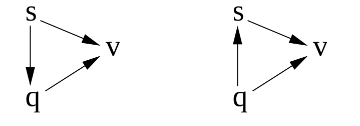
4 Applications

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 additional axiom 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.



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

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

Figure 4: Directed Acyclic Graphs $\boldsymbol{\mathcal{S}}$, $\boldsymbol{\mathcal{Q}}$

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 $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{\mathcal{S}}}$.

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

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:

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

As we observe, then, different causal models are sufficient to generate different behavior, given the same data and preferences.

4.2 Describing the interaction of agents with different models

For an agent to play in a causally coherent manner, I assume she plays according to a Bayes-Nash equilibrium of the game she believes she is playing.

Causally coherent agents with incorrect models run into a problem. It might be the case that an agent encounters an event for which she has no explanation: that is, an event that cannot be explained by any element in her set of plausible explanations. This is called a knowledge violation.

Definition 13. *A coherent agent i's* **knowledge is violated** *when the agent observes an event that is impossible given K_i, where "impossible" means an event that occurs with a non-positive probability (Pr=0 or zero density, as appropriate.)*

I would like to impose that no agent has her knowledge violated in equilibrium. I call this equilibrium a causally coherent equilibrium:

Definition 14. 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, and has knowledge set K_i such that $s\left\{\boldsymbol{\mathcal{C}}_i, F, \ldots\right\} \subseteq K_i$*
3. *If G specifies common knowledge K, then $K \subseteq K_i$ for all agents i.*
4. *Each agent i plays an action consistent with some Bayes-Nash equilibrium E_i of the game implied by K_i.*
5. *Each augmented knowledge set $\tilde{K}_i = K_i \bigcup E_i$ is never 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.

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$\boldsymbol{S}$ and investor I$\boldsymbol{Q}$, 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.[11] 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

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

[11] This example has discrete types to be consistent with the representation theorem, which is for finite spaces supp(S) × 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.

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 $0<M\leq\frac{1}{2}\min\{\alpha,\beta\}$[12]

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 $S=1$.

[12] See appendix section 7.1.1 for the details.

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.[13]

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 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 $(0, 1)$;

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

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.[14]

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

[14] 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.

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 expecions 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.[15] 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:

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

a similar set would be all possible values for (α, β). Standard Bayesian updating will eliminate possibilities (in the long run) as the $\widehat{\boldsymbol{\Phi}}$ 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

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

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.)[16] If the representation 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$\mathcal{S}$) and Quincy (investor I$\mathcal{Q}$): 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

[16] 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.

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.

Dekel, E., B. L. Lipman, and A. Rustichini (1998): "Standard State-Space Models Preclude Unawareness," *Econometrica*, 66(1), 159–174.

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.

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.

Pearl, J., and T. Verma (1991): “A Theory of Inferred Causation,” *Second International Conference on the Principles of Knowledge Representation and Reasoning.*

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.

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:

$$PO_{\boldsymbol{\mathcal{S}}}(M) \geq PO_{\boldsymbol{\mathcal{S}}}(0) \tag{3}$$

$$Prob_{\boldsymbol{\mathcal{S}}}(win|M)\alpha - M \geq Prob_{\boldsymbol{\mathcal{S}}}(win|0)\alpha \tag{4}$$

$$\left(\frac{1}{2}\gamma_s + (1-\gamma_s)\right)\alpha - M \geq (1-\gamma_s)\frac{1}{2}\alpha \tag{5}$$

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

$$\tag{6}$$

$$\frac{1}{2}\alpha \geq M \tag{7}$$

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

$$PO_{\boldsymbol{\mathcal{Q}}}(M) \geq PO_{\boldsymbol{\mathcal{Q}}}(0) \tag{8}$$

$$Prob_{\boldsymbol{\mathcal{Q}}}(win|M)\beta - M \geq Prob_{\boldsymbol{\mathcal{Q}}}(win|0)\beta \tag{9}$$

$$\left(\frac{1}{2}\gamma_q + (1-\gamma_q)\right)\beta - M \geq (1-\gamma_q)\frac{1}{2}\beta \tag{10}$$

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

$$\tag{11}$$

$$\frac{1}{2}\beta \geq M \tag{12}$$

7.1.2 That investor I$\boldsymbol{\mathcal{S}}$ 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 $\frac{2}{3} < \beta$.