Agents' Mental Models

by

Andreas Duus Pape

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2008

Doctoral Committee:

Professor Emre Ozdenoren, Co-Chair
Professor Scott E. Page, Co-Chair
Professor James M. Joyce
Professor John E. DiNardo

To my greatest teachers: My parents; my wife; Emre, Scott, and John; my friends; and all my students.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

Three essays investigating the construction and implications of economic agents' internal representations of problems they face.

In the second chapter, "Optimal Auctions under Ambiguity," we investigate the construction of an optimal auction mechanism when agents are ambiguity averse over the valuation of the other bidder. A crucial assumption in the optimal auction literature is that each bidder's valuation is known to be drawn from a unique distribution. In this paper we study the optimal auction problem allowing for ambiguity about the distribution of valuations. Agents may be ambiguity averse (modeled using the maxmin expected utility model of Gilboa and Schmeidler [14].) When the bidders face more ambiguity than the seller we show that (i) an auction that provides full insurance to all types of bidders is always in the set of optimal auctions, and in certain cases the seller can strictly increase his revenue by switching to a full insurance mechanism; (ii) if the seller is ambiguity neutral and any prior that is close enough to the seller's prior is included in the bidders' set of priors then the optimal auction must be a full insurance auction; (iii) in general, neither the first nor the second price auction is optimal (even with suitably chosen reserve prices). When the seller is ambiguity averse and the bidders are ambiguity neutral an auction that fully

insures the seller must in the set of optimal mechanisms.

In the third chapter, "Causal Coherence," I investigate agents with differing mental models of the same phenomenon. 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.

In the fourth chapter, "Of Wolves and Sheep," I place the agents developed in the previous chapter into an economy. In this simple dynamic economy, agents with different theories of how ideas develop into firms leads them to choose different optimal take-up of these ideas. Their different behaviors yields a predator/prey relationship among these agents, which causes natural population cycles of theories and behavior to emerge endogenously. The agents are identical but for their theories (identical data, actions, preferences) so the predator/prey relationship emerges only from their different interpretations of common data. Since the system does not collapse, it shows

that agents with differing theories may persist in a long-run, dynamic equilibrium.

CHAPTER II

Optimal Auctions with Ambiguity

2.1 Introduction

Optimal auctions for an indivisible object with risk neutral bidders and independently distributed valuations have been studied by, among others, Vickrey [51], Myerson [36], Harris and Raviv [17], and Riley and Samuelson [43]. These papers show that the set of optimal mechanisms or auctions is quite large, and that the set contains both the first and second price auctions with reserve prices. One of the assumptions in this literature is that each bidder's valuation is known to be drawn from a unique distribution. In this paper we relax this assumption and study how the design of the optimal auction is affected by the presence of ambiguity about the distribution from which the bidders' valuations are drawn.

The unique prior assumption is based on the subjective expected utility model, which has been criticized among others by Ellsberg [8]. Ellsberg shows that lack of knowledge about the distribution over states can affect choices in a fundamental way that can not be captured within the subjective expected utility framework. In one version of Ellsberg's experiment, a decision maker is offered two urns, one that has 50 black and 50 red balls, and one that has 100 black and red balls in unknown proportions. Faced with these two urns, the decision maker is offered a bet on black

but can decide from which urn to draw the ball. Most decision makers prefer the first urn. The same is true when the decision maker is offered the same bet on red. This behavior is inconsistent with the expected utility model. Intuitively, decision makers do not like betting on the second urn because they do not have enough information or, put differently, there is too much ambiguity. Being averse to ambiguity, they prefer to bet on the first urn. Ellsberg and many subsequent studies have demonstrated that ambiguity aversion is common.

Following Gilboa and Schmeidler [14], we model ambiguity aversion using the maxmin expected utility (MMEU) model. The MMEU model is a generalization of the subjective expected utility model, and provides a natural and tractable framework to study ambiguity aversion. In MMEU agents have a set of priors (instead of a single prior), on the underlying state space, and their payoff is the minimum expected utility over the set of priors. Specifically, when an MMEU bidder is confronted with an auction, he evaluates each bid on the basis of the minimum expected utility over the set of priors, and then chooses the best bid. An MMEU seller, on the other hand, evaluates each auction on the basis of its minimum expected revenue over the set of priors and chooses the best auction. In order to better contrast our results with the risk case, we assume that the bidders and the seller are risk neutral (i.e. have linear utility functions).

Our main result, Proposition II.1, is that when the bidders face more ambiguity than the seller an auction that provides full insurance to the bidders[1] is always in the set of optimal mechanisms. Moreover, given any incentive compatible and individually rational mechanism, the seller can strictly increase his revenue by switching to a full insurance mechanism if the minimum expected utility of a bidder over the

[1]A full insurance auction keeps the bidders' payoffs constant for all reports of the other bidders and consequently keeps them indifferent between winning or losing the object.

seller's set of priors is strictly larger than the one over the bidders' set of priors for a positive measure of types.

This result can explain some auction mechanisms that are observed in real life. In particular Goeree and Offerman [15] observe that: "In Europe, sellers of houses, land, boats, machinery and equipment regularly offer a premium to the highest losing bidder to promote competitive bidding. Many Dutch and Belgian towns have their own variant of premium auctions, some of which date back to the Middle Ages." Goeree and Offerman explain the existence of such auctions by asymmetries among bidders. They argue that even though premium auctions are not optimal, in environments with asymmetries among bidders they may be second best. In this paper we provide an alternative explanation by showing that even with symmetric bidders when there is ambiguity premium auctions may outperform standard auctions[2].

To obtain some intuition for the main result, consider the special case where the seller is ambiguity neutral, i.e., his set of priors is a singleton. In this case the main result says that if an incentive compatible and individually rational mechanism is optimal for the seller then the minimizing set of distributions for all types of the bidders must include the seller's prior. Suppose this is not true for a positive measure of types and consider some such type θ. In this case, the seller and type θ of the bidder will be willing to bet against each other. The seller would recognize that they have different beliefs about the underlying state space and would offer "side bets" using transfers. The crucial issue is that the modified mechanism will have to maintain overall incentive compatibility. In our proof we address this issue by explicitly

[2]Goeree and Offerman [15] study the Amsterdam auction which proceeds in two stages. In the first stage all but two bidders are eliminated, and in the second stage a premium auction is conducted. Our comments are only relevant for the second stage of their auction. Also, premium auctions are rich in many institutional details and we do not claim that any of these particular auctions exactly implement the optimal selling mechanism described in this paper. Rather, our objective is to point out that presence of ambiguity might be an alternative explanation as to why the seller in these auctions find it profitable to offer a payment to a losing bidder.

constructing the additional transfers that continue to satisfy incentive compatibility constraints while making the seller better off. Essentially, we show that these additional transfers (to the seller) can be chosen so that in the new mechanism, under truth telling type θ gets the minimum expected utility that he gets in the original mechanism in *every* state, and thus is fully insured against the ambiguity. Obviously then, under truth telling, type θ is indifferent between the original mechanism and the new mechanism since he gets the same minimum expected utility under both. More interestingly, no other type wants to imitate type θ in the new mechanism. This is because the additional transfers in the new mechanism are constructed so as to have zero expected value under the minimizing set of distributions for type θ in the original mechanism, but to have strictly positive expected value under any other distribution. Therefore, if type θ' imitates type θ in the new mechanism, he gets at best what he would get by imitating type θ in the original mechanism. Hence, since the original mechanism is incentive compatible, the new mechanism must also be incentive compatible. Moreover, since by assumption, the seller's distribution is not in the minimizing set for type θ in the original mechanism, the additional transfers (to the seller) must have strictly positive expected value under the seller's distribution. Since the original mechanism can be modified this way for a positive measure of types, the seller strictly increases his revenue. In fact, for any incentive compatible and individually rational mechanism, by modifying the mechanism for all types as described above we can obtain a full insurance auction that is weakly preferred by the seller. Therefore, a full insurance auction must always be in the set of optimal mechanisms.

There may be optimal selling mechanisms in addition to the full insurance mechanism, but in some cases the full insurance mechanism is the unique optimal mecha-

nism. In Proposition II.4 we show that if the seller is ambiguity neutral and any prior that is close enough to the seller's prior is included in the bidders' set of priors then the optimal auction must be a full insurance auction. We also show in Proposition II.7 that, in general, the first and the second price auctions are not optimal.

To highlight some of the economic implications of the above analysis, in section 2.4, we explicitly derive the optimal mechanism when the seller is ambiguity neutral and bidders' set of priors is the ε-contamination of the seller's prior. We show that the seller's revenue and efficiency both increase as ambiguity increases. We also describe an auction that implements the optimal mechanism.

When the seller is ambiguity averse and the bidders are ambiguity neutral we show that for every incentive compatible and individually rational selling mechanism there exists an incentive compatible and individually rational mechanism which provides deterministically the same payoff to the seller. From this it follows that when an optimal mechanism exists, an auction that fully insures the seller must in the set of optimal mechanisms. A similar result was first shown by Eso and Futo [10] for auctions (in independent private value environments) with a risk averse seller and risk (and ambiguity) neutral bidders. Hence, as long as bidders are risk and ambiguity neutral, ambiguity aversion on the part of the seller plays a similar role to that of risk aversion.

There is a small but growing literature on auction theory with non-expected utility starting with a series of papers by Karni and Safra ([24], [26], [25]) and Karni [22]. The papers that look at auctions with ambiguity averse bidders, and thus are closer to this paper are by Salo and Weber [45], Lo [28], Volij [52] and Ozdenoren [37]. These papers look at specific auction mechanisms, such as the first and second price auctions, and not the optimal auction problem.

Billot, Chateauneuf, Gilboa and Tallon [1] analyze the question of when it is optimal to take bets for agents with MMEU preferences in a pure exchange economy. They show that if the intersection of the set of priors for all agents is non-empty, then any Pareto optimal allocation is a full insurance allocation. This result is in the same spirit as our results. Furthermore, even though a direct comparison of the two models are difficult, a possible implication of our result could be that Billot et. al. [1] result may be robust to the introduction of incentive constraints. Another related paper is Mukerji [33] that shows that in the investment hold-up model ambiguity aversion can explain the existence of incomplete contracts. The incomplete or null contract is where the ex post surplus is split equally between the two parties and they thus agree on the ranking of the states. To implement more efficient investments, a contract has to introduce more variation in ex-post payoffs which would also result in disagreement among the two parties; and when ambiguity is sufficiently large any such contract would be dominated by the null contract.

Matthews [32] and Maskin and Riley [30] study auctions with risk averse bidders. A more detailed comparison of our paper with Maskin and Riley [30] is given in section 2.7.

Finally, there is also a strand of literature that studies robust mechanism design. (See for example Bergemann and Morris [4], Ely and Chung [9], and Heifetz and Neeman [19]). Even though there is some similarity between that literature and our work here (as we, just like them, relax certain assumptions of the standard mechanism design framework), it is important to point out that we differ significantly from this literature. Standard mechanism design - in particular Bayesian implementation - relies crucially on the underlying model being common knowledge. The focus of those papers is to study mechanism design while relaxing (some of) the common knowl-

edge assumptions. In contrast, we maintain throughout the standard methodological assumption of considering the underlying *model* - that includes the modelling of the ambiguity - to be common knowledge and we relax assumptions on the preferences of the agents, in particular, we allow the agents to exhibit ambiguity aversion.

2.2 The Optimal Auction Problem

In this section we generalize the optimal auction problem by allowing the bidders and the seller to have MMEU preferences (Gilboa and Schmeidler [14].) There are two bidders and a seller. We assume that both the bidders and the seller have linear utility functions. Bidders have one of a continuum of valuations $\theta \in \Theta = [0, 1]$. Let Σ be the Borel algebra on Θ. Each bidder knows his true valuation but not that of the other. The set Δ_B^m is a set of probability measures on (Θ, Σ) with a corresponding set Δ_B of distribution functions. This set represents each bidder's belief about the other bidder's valuation. Bidders believe that valuations are generated independently, but they may not be confident about the probabilistic process that generates the valuations. This possible vagueness in the bidders' information is captured by allowing for a set of priors rather than a single prior in this model.

The seller is also allowed to be ambiguity averse. The set Δ_S^m is a set of probability measures on (Θ, Σ) with a corresponding set Δ_S of distribution functions. This set represents the seller's belief about the bidders' valuations. That is, the seller believes that bidders' valuations are generated independently from some distribution in Δ_S.[3] Each bidder's reservation utility is 0. As is standard, we assume that all of the above is common knowledge.

[3]Formally the seller's belief is the set of product measures $\mu \times \mu$ on the product space $(\Theta \times \Theta, \Sigma \times \Sigma)$ where $\mu \in \Delta_S^m$, even though for notational simplicity, throughout the paper we continue to refer to the seller's belief simply as as $\mu \in \Delta_S^m$. It is important, however, to keep this in mind, especially for a later result (Proposition 2) which talks about the seller's belief being in the interior of the set of the bidders' set of beliefs.

We consider symmetric mechanisms where bidders make simultaneous reports. As we show in section .1.1 of the appendix the revelation principle holds in this setting so we restrict attention to direct revelation mechanisms. In the direct revelation mechanism, each bidder is asked to report his type, where a report is some $\theta \in \Theta$. The mechanism stipulates a probability for assigning the item and a transfer rule as a function of reported types. Let $x(\theta, \theta')$ be the item assignment probability function and $t(\theta, \theta')$ the transfer rule. The convention is that the first entry is one's own report, the second entry is the report of the other bidder.[4]

The seller's problem is to find a mechanism (x, t) that solves

$$\sup_{(x,t)} \left[\inf_{F \in \Delta_S} \iint \left[t\left(\theta, \theta'\right) + t\left(\theta', \theta\right) \right] dF\left(\theta\right) dF\left(\theta'\right) \right] \tag{2.1}$$

subject to

$$\text{(IC)} \qquad \inf_{G \in \Delta_B} \int \left(x\left(\theta, \theta'\right) \theta - t\left(\theta, \theta'\right) \right) dG\left(\theta'\right) \tag{2.2}$$

$$\geq \inf_{G \in \Delta_B} \int \left(x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta') \right) dG(\theta') \text{ for all } \theta, \tilde{\theta} \in \Theta \tag{2.3}$$

and,

$$\text{(IR)} \quad \inf_{G \in \Delta_B} \int \left(x\left(\theta, \theta'\right) \theta - t\left(\theta, \theta'\right) \right) dG\left(\theta'\right) \geq 0 \text{ for all } \theta \in \Theta. \tag{2.4}$$

The first inequality gives the incentive compatibility (IC) constraints, and the second inequality gives the individual rationality (IR) or participation constraint. These are the usual constraints except that the bidders compute their utility in the mechanism using the MMEU rule. For example, the IC constraint requires that the infimum expected utility a bidder of type θ gets reporting his type truthfully is at

[4] A word about notation before we proceed. Formally, one should have separate notation for reports, say $\widehat{\theta}$ as opposed to θ, and define the mechanism in terms of reports, not types. Since the optimal mechanisms we describe will all be incentive compatible, here and in several other places, we save on notation by describing the mechanisms directly in terms of θ. We hope this departure from convention, however, will cause no confusion.

least as much as the infimum expected utility that he gets under reporting any other type $\tilde{\theta}$.

One way to think about the set of priors in the above formulation is a "subjective" interpretation where preferences of the players are common knowledge, and the sets are subjective representations of the uncertainty (as well as the aversion to this uncertainty) players face about the stochastic process that generates the valuations. Alternatively, one can think of an "objective" interpretation of the set of priors, in which, players learn everything that they can learn about the stochastic process that generates the types, but there are hard to describe factors that prevent them from learning the process completely. The objective interpretation is more restrictive then the subjective one for two reasons. First, when the set of priors is objectively fixed, bidders ambiguity attitude is represented by the minimum functional *only*, which may be viewed as extreme. Second, the objective interpretation makes sense when both the seller and the buyers have the same set of priors (which is covered in our framework), since the set of priors is assumed to be common knowledge.

Note that in our formulation, we differ slightly from Gilboa and Schmeidler since we use infimum (supremum) instead of minimum (maximum); however, we continue to refer to these preferences as maxmin since this is the standard terminology. At the end of the next section we provide conditions on preferences and mechanisms that will guarantee that the minimum over the sets of priors and an optimal auction exist.

2.3 Full Insurance auction

In this section we show that, when $\Delta_S^m \subseteq \Delta_B^m$,[5] a full insurance auction is always in the set of optimal auctions and discuss when the seller can make strict gains by switching to a full insurance auction. In what follows, for a given mechanism (x,t), it will be convenient to define

$$q_{(x,t)}(\theta,\theta') \equiv x(\theta,\theta')\theta - t\,(\theta,\theta')$$

for all $\theta,\theta' \in \Theta$. So $q_{(x,t)}(\theta,\theta')$ is the payoff to type θ from truth telling in the mechanism (x,t) when the other bidder reports θ'. When it is clear from the context which mechanism we are referring to, we drop the subscript (x,t) and use q instead of $q_{(x,t)}$.

We say that an event $\widetilde{\Theta} \subseteq \Theta$ has positive measure if $\inf_{\mu\in\Delta_S^m} \mu\left(\tilde{\Theta}\right) > 0$ and zero measure otherwise. Next, we formally define a full insurance auction.

Definition 1. *A* full insurance mechanism *is one where the payoff of any bidder is constant for any report of the competing bidder. That is (x,t) is a full insurance mechanism if, for almost all $\theta \in \Theta$, $q(\theta,\theta')$ is constant as a function of $\theta' \in \Theta$.*

Next, we give the formal statement of the main proposition. All proofs are in the appendix.

Proposition II.1. *Suppose that the seller and the bidders are ambiguity averse, the seller's set of priors is Δ_S and the bidders' set of priors is Δ_B with $\Delta_S \subseteq \Delta_B$. Let (x,t) be an arbitrary incentive compatible and individually rational mechanism. There is always a full insurance mechanism, also satisfying incentive compatibility and individual rationality, that generates at least as much minimum expected revenue*

[5] In particular, this covers two interesting cases. If Δ_S^m is a singleton set, then the seller is ambiguity neutral and the bidders are (weakly) ambiguity averse. On the other hand, if $\Delta_S^m = \Delta_B^m$, then both the seller and the bidders are (weakly) ambiguity averse with a common set of priors.

over the set of priors Δ_S for the seller. Moreover if there exists some positive measure event $\widetilde{\Theta} \subseteq \Theta$ such that for all $\theta \in \widetilde{\Theta}$,

$$\inf_{G \in \Delta_S} \int_{\Theta} q(\theta, \theta') dG(\theta') > \inf_{H \in \Delta_B} \int_{\Theta} q(\theta, \theta') dH(\theta') \tag{2.5}$$

then (x, t) is not optimal. In fact, the seller can strictly *increase his minimum expected revenue over the set of priors Δ_S using a full insurance mechanism.*

To understand this result consider the case where the seller is ambiguity neutral, i.e., $\Delta_S = \{F\}$. Let

$$\Delta(\theta) = \arg\min_{H \in \Delta_B} \int_{\Theta} q(\theta, \theta') dH(\theta')$$

where for ease of exposition we assume that the minimum exists so that we write min instead of inf[6]. In this case Proposition II.1 says that if a mechanism (x, t) is optimal then F must be in $\Delta(\theta)$ for almost all $\theta \in \Theta$. Suppose to the contrary that there exists a positive measure of types for which this is not true. Consider some such type $\tilde{\theta}$ for which $F \notin \Delta\left(\tilde{\theta}\right)$. The seller can always adjust transfers of type $\tilde{\theta}$, so that type $\tilde{\theta}$ under truth telling gets the *same* minimum expected utility that he gets in the original mechanism in every state, and thus is fully insured against ambiguity in the new mechanism. Furthermore, by construction, the difference between the transfers in the new mechanism and the original mechanism has weakly positive expected value[7] for any distribution in Δ_B. This is true because this difference has zero expected value under $\Delta\left(\tilde{\theta}\right)$, the minimizing set of distributions in the original mechanism, and strictly positive expected value under any other distribution, i.e., for distributions in $\Delta_B - \Delta\left(\tilde{\theta}\right)$. Obviously, under truth telling, type $\tilde{\theta}$ is indifferent between the original mechanism and the new mechanism since he gets the same minimum expected utility under both. More interestingly, no other type wants to

[6] See proposition II.2 below for conditions that guarantee that this assumption holds.

[7] Recall that these are transfers *to* the seller.

imitate type $\tilde{\theta}$ in the new mechanism. This is true since the original mechanism is incentive compatible and imitation in the new mechanism is even worse given that the difference in transfers has weakly positive expected value under any distribution in Δ_B. Moreover, by assumption, the seller's distribution is not in the minimizing set for the original mechanism, which means the additional transfers (to the seller) must have strictly positive expected value under the seller's distribution. Thus the seller is strictly better off in the new mechanism, contradicting the optimality of the original mechanism.

When the infimums and supremums in equations (2.1), (2.3) and (2.4) are replaced with minimums and maximums, we can prove a stronger version of Proposition II.1. The next proposition provides sufficient conditions for this.

Proposition II.2. *Suppose the seller can only use mechanisms such that transfers are uniformly bounded and suppose that $\Delta_B^m \cup \Delta_S^m$ is weakly compact and convex and its elements are countably additive probability measures. Then the sets of minimizing priors in equations (2.1), (2.3), (2.4) and the set of optimal mechanisms are nonempty.*

The following corollary strengthens Proposition II.1 when the hypothesis of Proposition II.2 holds.

Corollary II.3. *Suppose that the hypothesis of Proposition II.2 holds. Let*

$$\Delta_S^{\min} = \arg\min_{G \in \Delta_S} \int_\Theta \int_\Theta \left[t(\theta, \theta') + t\left(\theta', \theta\right) \right] dG\left(\theta\right) dG\left(\theta'\right).$$

For any mechanism (x, t), if there exists some positive measure event $\widetilde{\Theta} \subseteq \Theta$ such that for all $\theta \in \widetilde{\Theta}$ and for all $G \in \Delta_S^{\min}$,

$$\int_\Theta q(\theta, \theta') dG(\theta') > \min_{H \in \Delta_B} \int_\Theta q(\theta, \theta') dH(\theta') \tag{2.6}$$

then the seller can strictly *increase his minimum expected revenue over the set of priors Δ_S using a full insurance mechanism. Moreover, there is a full insurance mechanism that is optimal for the seller.*

Corollary II.3 is stronger than Proposition II.1 in two ways. First inequality in (2.6) is checked only for the *minimizing* distributions for the seller (not all the distributions in Δ_S). Second, since an optimal mechanism exists a full insurance auction is always optimal for the seller[8].

In general there may be optimal selling mechanisms that are different from the full insurance mechanism. On the other hand, if the seller's belief F has strictly positive density and if any prior that is close enough to F is in Δ_B, then the set of distributions that give the minimum expected utility will not include F unless the ex post payoffs are constant. In this case the optimal auction must be a full insurance auction. The next proposition states this observation.

Proposition II.4. *Suppose that the seller is ambiguity neutral with $\Delta_S = \{F\}$ where F has strictly positive density. If there exists $\varepsilon > 0$ such that for any distribution H on Θ, $(1-\varepsilon)\,F + \varepsilon H \in \Delta_B$, then the unique optimal auction is a full insurance auction.*

In the next two sections we provide some applications of the results in this section.

2.4 Full insurance under ε-contamination

In this section we explicitly derive the optimal mechanism in the case of ε-contamination when the seller is ambiguity neutral with $\Delta_S = \{F\}$. In ε-contamination we assume that the seller's distribution F is a focal point, and bidders allow for an

[8]In contrast, Proposition II.1 says that the seller can get arbitrarily close the supremum in equation (2.1) using a full insurance mechanism.

ε-order amount of noise around this focal distribution. We make the common assumptions that F has a strictly positive density f and,

$$L(\theta) = \theta - \frac{1 - F(\theta)}{f(\theta)}$$

is strictly increasing in θ. We construct Δ_B as follows:

$$\Delta_B = \{G : G = (1-\varepsilon) F + \varepsilon H \text{ for any distribution } H \text{ on } \Theta\}$$

where $\varepsilon \in (0, 1]$. By Proposition II.4 we know that the unique optimal mechanism for the ε-contamination case is a full insurance mechanism. This implies that we can restrict ourselves to full insurance mechanisms in our search for the optimal mechanism.

Let (x, t) be a full insurance mechanism i.e., $q(\theta, \theta')$ is constant for all θ'. Let $u(\theta) = q(\theta, \theta')$. Next we define some useful notation. Let

$$\begin{aligned} X(\theta) &= \int x(\theta, \theta') dF(\theta'), \\ X^{\min}(\theta) &= \inf_{G \in \Delta_B} \int x(\theta, \theta') dG(\theta'), \\ X^{\max}(\theta) &= \sup_{G \in \Delta_B} \int x(\theta, \theta') dG(\theta'). \end{aligned}$$

Using the IC constraint we obtain,

$$u(\theta) = \inf_{G \in \Delta_B} \int (x(\theta, \theta')\theta - t(\theta, \theta')) dG(\theta') \geq \inf_{G \in \Delta_B} \int \left(x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta') \right) dG(\theta') \tag{2.7}$$

$$= u\left(\tilde{\theta}\right) + \inf_{G \in \Delta_B} \int \left(\theta - \tilde{\theta}\right) x(\tilde{\theta}, \theta') dG(\theta').$$

If $\theta > \tilde{\theta}$ then

$$u(\theta) \geq u\left(\tilde{\theta}\right) + \left(\theta - \tilde{\theta}\right) X^{\min}\left(\tilde{\theta}\right). \tag{2.8}$$

Exchanging the roles of θ and $\tilde{\theta}$ in (2.7) we obtain

$$u\left(\tilde{\theta}\right) \geq u\left(\theta\right) + \inf_{G \in \Delta_B} \int \left(\tilde{\theta} - \theta\right) x(\theta, \theta') dG(\theta')$$

Again if $\theta > \tilde{\theta}$ then

$$u\left(\tilde{\theta}\right) \geq u\left(\theta\right) + \left(\tilde{\theta} - \theta\right) X^{\max}\left(\theta\right). \tag{2.9}$$

Now observe that u is non-decreasing since, for $\theta > \tilde{\theta}$ by the IC constraint we have,

$$u\left(\theta\right) \geq u\left(\tilde{\theta}\right) + \left(\theta - \tilde{\theta}\right) X^{\min}\left(\tilde{\theta}\right) \geq u\left(\tilde{\theta}\right).$$

The next lemma is useful in characterizing the optimal auction.

Lemma II.5. *The function u is Lipschitz.*

Since u is Lipschitz, it is absolutely continuous and therefore is differentiable almost everywhere. For $\theta > \tilde{\theta}$ we use (2.8) and (2.9) to obtain,

$$X^{\max}\left(\theta\right) \geq \frac{u\left(\theta\right) - u\left(\tilde{\theta}\right)}{\theta - \tilde{\theta}} \geq X^{\min}\left(\tilde{\theta}\right).$$

We take the limit as $\tilde{\theta}$ goes to θ to obtain for almost all θ that,

$$X^{\max}\left(\theta\right) \geq \frac{\partial u}{\partial \theta} \geq X^{\min}\left(\theta\right).$$

Since an absolutely continuous function is the definite integral of its derivative,

$$\int_0^\theta X^{\max}\left(y\right) dy \geq u\left(\theta\right) - u\left(0\right) \geq \int_0^\theta X^{\min}\left(y\right) dy. \tag{2.10}$$

Equation (2.10) suggests that the auctioneer may set,

$$u\left(\theta\right) = \int_0^\theta X^{\min}\left(y\right) dy \tag{2.11}$$

and

$$t\left(\theta, \theta'\right) = x\left(\theta, \theta'\right)\theta - \int_0^\theta X^{\min}\left(y\right) dy, \tag{2.12}$$

since for a given allocation rule x, transfers as in (2.12) are the highest transfers the auctioneer can set without violating (2.10). Of course, (2.10) is only a necessary condition and for a given allocation rule x, the resulting mechanism (x,t) may not be incentive compatible. Fortunately, this difficulty does not arise if the allocation rule x is chosen optimally for transfers given as in (2.12). In other words, our strategy is to find the optimal allocation rule x, assuming that the transfers are given by (2.12), and then show that the resulting mechanism, (x,t) is incentive compatible.

For transfer function given by (2.12), we can rewrite the seller's revenue as,

$$R = 2\int_0^1 \int_0^1 \left(\theta x\left(\theta,\theta'\right) - \int_0^\theta X^{\min}\left(y\right) dy\right) dF\left(\theta'\right) dF\left(\theta\right).$$

Using integration by parts we obtain,

$$R = 2\int_0^1 \theta X\left(\theta\right) f\left(\theta\right) d\theta - \int_0^1 \left(1 - F\left(\theta\right)\right) X^{\min}\left(\theta\right) d\theta. \tag{2.13}$$

Define,

$$L^{\varepsilon}\left(\theta\right) = \theta - \left(1-\varepsilon\right)\frac{1-F\left(\theta\right)}{f\left(\theta\right)},$$

and let $r \in (0,1)$ be such that $L^{\varepsilon}\left(r\right) = 0$.

The following proposition characterizes the optimal allocation when transfer function is given by (2.12).

Proposition II.6. *For any θ and θ', let*

$$x\left(\theta,\theta'\right) = \begin{cases} 1 & \text{if } \theta > \theta' \text{ and } \theta \geq r \\ 1/2 & \text{if } \theta = \theta' \text{ and } \theta \geq r \\ 0 & \text{otherwise} \end{cases}$$

and let t be given by (2.12). The mechanism (x,t) defined this way is the unique optimal mechanism for the seller.

It is interesting to note economic implications of the above analysis for revenue and efficiency. First, the seller's revenue increases as ambiguity increases. To see this note that under the above allocation rule $X^{\min}(\theta) = (1-\varepsilon) X(\theta)$ for all $\theta < 1$. Plugging this into the revenue expression (2.13) we see that the revenue increases as ε increases. In fact, when ambiguity becomes extreme, i.e., as ε approaches to one, the seller can extract all the surplus.

Second, an increase in ambiguity helps efficiency. To see this note that L^{ϵ} shifts up as ε increases and since $L^{\varepsilon}(\theta)$ is an increasing function of θ, the cutoff type r decreases as ε increases. Again in the case of extreme ambiguity the seller does not exclude any types, and full efficiency is achieved.

Finally, a natural question to ask at this stage is how to implement the optimal mechanism described above. There are several auctions that implement the mechanism, and we will describe one such auction here. Consider an auction where bidders submit bids for the object and the allocation rule is the usual one, namely, the highest bidder who bids above the reservation value r obtains the object. The payment scheme is as follows: the winning bidder pays to the auctioneer an amount equal to his bid, and all bidders (regardless of having won or lost) who have bid above the reservation price receives a gift from the seller. For a bidder who bids, say, b, (where b is greater than r), the amount of the gift is given by $S(b) = (1-\varepsilon)\int_r^b F(y)\,dy$. In this auction, the equilibrium strategy of a bidder with valuation θ is to bid his valuation. To see this note that the allocation rule is the same as the one in Proposition II.6. Moreover, a bidder who bids θ pays $\theta - (1-\varepsilon)\int_r^{\theta} F(y)\,dy$ if he wins the auction and $-(1-\varepsilon)\int_r^{\theta} F(y)\,dy$ if he loses the auction, and these transfers are also the ones in Proposition II.6. Since reporting one's true value is incentive compatible in the optimal mechanism, it is also optimal to bid one's true value in this auction as well.

2.5 The First and Second Price Auctions

Lo [28] showed that the revenue equivalence result does not hold when bidders are ambiguity averse. In particular, the first price auction may generate more revenue than the second price auction. In this section we show that the first price auction is in general not optimal either[9]. In fact under rather general conditions, the first and second price auctions, as well as many other standard auctions are not optimal in this setting. The following proposition gives a weak condition on Δ_B that is sufficient for the non-optimality of a large class of auctions including the first and second price auctions.

Proposition II.7. *Suppose that Δ_S and Δ_B are weakly compact and convex with elements that are countably additive probability measures. Suppose that for any $G \in \Delta_S$ there exists some distribution $H \in \Delta_B$ such that H first-order stochastically dominates G. Now, if under some mechanism (x,t) with uniformly bounded transfers, there exists a positive measure subset $\widetilde{\Theta} \subseteq \Theta$ such that for all $\tilde{\theta} \in \widetilde{\Theta}$, $q(\tilde{\theta},\theta)$ is weakly decreasing in θ and $q(\tilde{\theta},\theta') < q(\tilde{\theta},\theta'')$ for some $\theta',\theta'' \in \Theta$ then (x,t) is not optimal.*

To apply the above proposition to the first and second price auctions, we need to show that in the direct mechanisms that correspond to these auction forms $q(\tilde{\theta},\theta)$ is weakly decreasing in θ and $q(\tilde{\theta},\theta') < q(\tilde{\theta},\theta'')$ for some $\theta',\theta'' \in \Theta$ for a positive measure subset $\widetilde{\Theta} \subseteq \Theta$. First note that the payoff $q(\tilde{\theta},\theta)$ of all types of a bidder is weakly decreasing in the report of the other bidder. Next consider a type $\tilde{\theta}$ that is greater than the reserve price and less than one which is the highest possible valuation. The payoff of $\tilde{\theta}$ is strictly larger if the other bidder reports a type θ' that is less than $\tilde{\theta}$ as opposed to a type θ'' more than $\tilde{\theta}$. This is because in both of these

[9]When the type space is discrete, neither the first nor the second price auction is the optimal auction for reasons completely unrelated to the issues being studied in this paper.

auctions if the other bidder reports more than $\tilde{\theta}$ the payoff of $\tilde{\theta}$ is zero, but if the other bidder reports less than $\tilde{\theta}$ the payoff of $\tilde{\theta}$ is strictly positive. This shows that $q(\tilde{\theta}, \theta') < q(\tilde{\theta}, \theta'')$. Therefore under the hypothesis of Proposition II.7 the first and second price auctions are not optimal.

2.6 Ambiguity Averse Seller

In this section we first provide a result that is in some sense a counterpart of Proposition II.1.

Proposition II.8. *Suppose that the seller is ambiguity averse, with a set of priors Δ_S and the bidders are ambiguity neutral with a prior $F \in \Delta_S$. For every incentive compatible and individually rational selling mechanism (x, t) there exists an incentive compatible and individually rational mechanism $\left(x, \tilde{t}\right)$ which provides deterministically the same revenue to the seller, i.e. $\tilde{t}(\theta, \theta') + \tilde{t}(\theta', \theta)$ is constant for all $\theta, \theta' \in \Theta$. Moreover if,*

$$\inf_{G \in \Delta_S} \iint \left[t(\theta, \theta') + t(\theta', \theta)\right] dG(\theta) dG(\theta') < \iint \left[t(\theta, \theta') + t(\theta', \theta)\right] dF(\theta) dF(\theta')$$

then $\left(x, \tilde{t}\right)$ strictly increases the minimum expected revenue of the seller over the set of priors Δ_S.

When an optimal mechanism exists, Proposition II.8 implies that an auction that fully insures the seller must in the set of optimal mechanisms. Eso and Futo [10] prove a similar result for auctions with a risk averse seller in independent private values environments with risk (and ambiguity) neutral bidders.

The basic idea of the proof is simple. For any individually rational and incentive compatible mechanism (x, t), one can define a new mechanism $\left(\tilde{x}, \tilde{t}\right)$ where the

allocation rule $\tilde{x}$ is the same as x, but with the following transfers:

$$\tilde{t}(\theta, \theta') = T(\theta) - T(\theta') + \int T(i)dF(i)$$

where

$$T(\theta) = \int t(\theta, \theta')dF(\theta').$$

Note that in the new mechanism $\tilde{t}(\theta, \theta') + \tilde{t}(\theta', \theta)$ is always $2\int T(i)dF(i)$ which is constant. It is straightforward to check that this mechanism is incentive compatible and individually rational as well. The reason this mechanism works in both risk and ambiguity settings is that, since the bidders are risk and ambiguity neutral $\left(\tilde{x}, \tilde{t}\right)$ is incentive compatible in either setting (risk or ambiguity) and provides full insurance to the seller against both.

2.7 Comparison of Optimal Auctions with Risk Averse vs. Ambiguity Averse Bidders

Matthews [32] and Maskin and Riley [30], henceforth, MR, relax the assumption that bidders are risk neutral and replace it with risk aversion. Even though there is some similarity between risk aversion and ambiguity aversion, the two are distinct phenomena. In particular, an environment with risk-averse bidders gives rise to optimal auctions that are different from the optimal auctions when bidders are ambiguity averse. In this section we contrast our results with those in MR, to highlight this distinction. To facilitate comparison, we assume, like MR, that the seller is risk and ambiguity neutral. Bidders, on the other hand, are risk averse and ambiguity neutral in MR and risk neutral and ambiguity averse in this paper.

MR define $u(-t, \theta)$ as the utility of a bidder of type θ when he wins and pays t, and $w(-t)$ as the utility when the bidder loses the auction (and pays t). Assuming

$u(\cdot)$ and $w(\cdot)$ to be concave functions, they note that if the auction mechanism is such that the marginal utility u_1 is different from w_1, then keeping other things constant, a seller can gain by rearranging the payments in such a way that the bidder's expected utility remains the same while the expected value of the revenue increases. They note however, that providing this insurance can change the incentives of the bidders; in particular when the marginal utility, u_1 varies with θ, the seller can exploit this to earn higher revenue by exposing all but the highest type to some risk, thus, in effect, screening types better. MR define a mechanism called *perfect insurance auction* where the marginal utility u_1 is equal to marginal utility w_1 for all types. Their results show that in general the optimal auction is not perfect insurance, the exception being the situation when bidders' preferences satisfy the condition $u_{12} = 0$, i.e. when the marginal utility u_1 does not vary with θ. (See their discussion following Theorem 11).

To contrast their result with ours, notice first that in our model, (using their notation) $u(-t, \theta) = \theta - t$, and $w(-t) = -t$, so that $u_{12} = 0$, and more importantly, the marginal utilities, when a bidder wins and when he loses are equal to each other in all situations. (This is just restating the fact that we assume risk-neutral bidders in our model). With ambiguity averse bidders, our results show that a full insurance auction is always within the set of optimal auctions and in some situations it is the uniquely optimal one. With risk-averse bidders, MR show that for the special case when $u(-t, \theta) = \theta - v(t)$ and $w(-t) = -v(t)$, so that $u_{12} = 0$, the optimal auction is a perfect insurance auction (given that $v(.)$ is a convex function)[10]. Notice however, that our full insurance auction is *different* from their perfect insurance auction, since in a full insurance auction $x(\theta, \theta')\theta - t(\theta, \theta')$ is a function of θ only (i.e., does not vary

[10]Put differently, letting a and b be the payments when a bidder wins and loses the auction respectively, MR show that convexity of $v(.)$ implies that $a = b$ under the optimal mechanism.

with θ'), which means that the realized payoff when the bidder wins, $\theta - t(\theta, \theta')$ is the same as the realized payoff when he loses, $-t(\theta, \theta')$. Hence, the optimal auctions under the two situations are *different* mechanisms even when preferences in their model satisfy the restriction $u_{12} = 0$.

Finally, note that in their framework, perfect insurance auctions do become full insurance auctions when preferences satisfy what they call Case 1. This is when $u(-t, \theta) = U(\theta - t)$ and $w(-t) = U(-t)$, (with U a concave function) so that equating marginal utilities implies equating utilities. However, in this situation, the perfect insurance auction (and hence the full insurance auction) is revenue equivalent to the second price auction (MR, Theorem 6). When U is strictly concave, both full insurance and second price auctions generate expected revenue that is strictly less than the expected revenue from the high bid auction (MR, Theorem 4 and Theorem 6; see in particular, the discussion at the bottom of page 1491). Hence, the full insurance auction, which is the optimal mechanism under ambiguity aversion (in some cases, as mentioned above, is the uniquely optimal mechanism) is not the optimal mechanism in the risk aversion framework.

2.8 Conclusion

We analyzed auctions a seller designs to maximize profit when agents might not know the distribution from which bidders' valuations are drawn. We have shown that when bidders face more ambiguity than the seller, an auction that provides full insurance to the bidders is optimal and sometimes it is uniquely optimal. We have also shown that standard auctions such as the first and the second price auctions with reserve prices, are not optimal in this setting. We have also shown that when the bidders are ambiguity neutral, but the seller ambiguity averse, it is the seller who

is perfectly insured.

These methods developed here maybe be used in other mechanism design problems with incomplete information in which agents are ambiguity averse. We believe that the results in this paper will naturally extend to these situations, especially in environments where the payoffs are quasilinear. For example in a bargaining problem (see Myerson [35]) we conjecture that the most efficient (from the mechanism designer's point of view) mechanism will require that some agent be fully insured against the ambiguity. In any case, and unlike the standard unique prior environment, the transfer and not just the allocation rule will play a crucial role in the design of the optimal mechanism in the presence of ambiguity. We hope to explore these extensions in future research.

CHAPTER III

Causal Coherence

3.1 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 running a company is the difference between prediction and intervention. To make money on the stock of a company, on only needs to predict what will happen to the company. If Sam and Quincy make the different causal inferences from the same data, they may disagree on counterfactuals, though they are equally good at predicting the future of a company that they both observe.

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 3.2, I describe causal bayesian networks. Causal Bayesian networks are mathematical objects that can represent probabilistic and causal information. Causal Bayesian Networks represent variables as nodes on a network connected with causal arrows, and have a associated family of conditional distributions. They are used extensively in artificial intelligence and statistics [40]. In statistics, they are used by model makers to estimate causal effects. In artificial intelligence, they are used to represent a mental model of a problem. I describe this framework, which allows me to 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.

Causal Bayesian Networks emerge in the utility representation provided in section

3.3, 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 Karni05 [23], 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 3.4, 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 3.5, 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 El61 [8], who treat causal ambiguity in a manner similar to GiSc89's [14] Maxmin expected utility agents. The third would use this framework to construct agents who act in accordance with QuattroneTversky84's [42] empirical finding that people attribute causation to correlation.

I conclude in section 3.6.

3.2 Background in Causal Bayesian Networks

Causal Bayesian Networks are a way of representing causal models. 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 observes a cross-section of firms in the world, some of which were successes and some failures, and may observe other characteristics of these firms. Given these observations, she constructs a theory (a causal model) of how characteristics of the firms determine success.

Suppose the agent has the opportunity, after constructing her model, to manipulate the firm by changing one or more characteristics. (One such manipulation would be replacing the CEO.) The agent's causal model of the phenomenon provides, given the agent's observations about the firm, a forecast for the outcome of each of possible manipulations. Given these predictions, the agent can decide how to optimally manipulate the 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 SlomanLagnado04 [48]: "A formal framework has recently been developed based on Bayesian graphical probability models to reason about causal systems (Spirtes93 [49]; reviewed in Pearl00 [40]). 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 Pearl00 [40]. 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 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[18] 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 Spirtes93 [49] and Pearl00 [40].

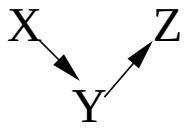


Figure 3.1: A causal structure

3.2.1 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 3.1 depicts a directed acyclic graph.

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 [40]. Figure 3.1 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 3.1, 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 [46], 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.

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

[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).

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 2. *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.

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

[3] Adapted from Pearl00 [40].

One causal structure $\boldsymbol{\mathcal{S}}$ over those variables is represented by Figure 3.2, 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.

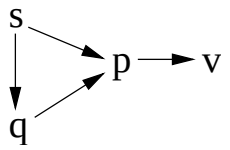


Figure 3.2: Causal structure $\boldsymbol{\mathcal{S}}$

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.

$$S = \beta_{S,0} + \epsilon_S$$

$$Q = \beta_{Q,0} + \beta_{Q,1}S + \epsilon_Q$$

$$P = \beta_{P,0} + \beta_{P,1}S + \beta_{P,2}Q + \epsilon_P$$

$$V = \beta_{V,0} + \beta_{V,1}P + \epsilon_V$$

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:

$$S \sim \Phi_S^{\boldsymbol{\mathcal{C}}}$$

$$Q \sim \Phi_Q^{\boldsymbol{\mathcal{C}}}(S)$$

$$P \sim \Phi_P^{\boldsymbol{\mathcal{C}}}(S, Q)$$

$$V \sim \Phi_V^{\boldsymbol{\mathcal{C}}}(P)$$

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.

3.2.2 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 3. *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.*

Pearl00 [40] 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 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

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

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 III.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

[5] The weather/barometer example is in both DruzelSimon93 [6] and Pearl00 [40].

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

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

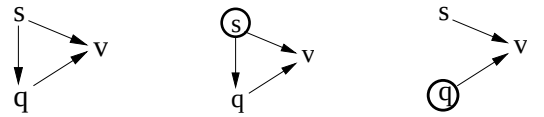
I will illustrate the all the objects discussed so far in this framework with the investor example: suppose a CEO is taking over a firm and investors are trying to forecast what will happen to the value of this company when she takes over.

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");
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 3.3(a).



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

Figure 3.3: 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 3.3(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):

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

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 3.3(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:

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

3.2.3 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 III.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.

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

□

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 [40]. 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 [40]."

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:

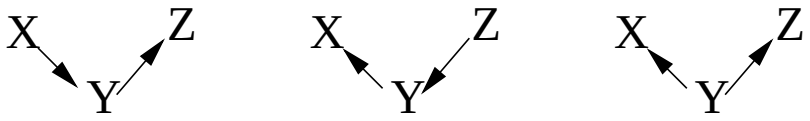


Figure 3.4: Directed Acyclic Graphs that are compatible with F

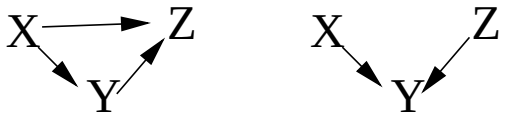


Figure 3.5: Directed Acyclic Graphs that are incompatible with F

Figures 3.4(a)-3.4(c) are all compatible with F. Figure 3.5(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 3.5(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 3.4(a)-3.4(c) form the three reasonable models given F.

The set of reasonable models is characterized by this theorem, by PearlVerma90 [50]:

Theorem III.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.*

3.2.4 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 3.3(a) and discussed in the previous section). Taking her causal model to be correct, she acts rationally. This is defined as causal coherence.

Definition 4. *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 5. *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.

3.3 A Utility Representation of the Causally Coherent Agent

In this section, I adapt Karni05's [23] representation theorem to the causal model setting to gain a utility representation for the causally coherent agent. Karni05 [23] 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. Jeffrey64 [20], Joyce99 [21]) 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.

3.3.1 Axioms

Here are the axioms. Axioms 0-4 are from Karni05 [23]. 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*

$(do(z'), (b''_{-\theta}, r'')) \succsim (do(z'), (b'''_{-\theta}, r'''))$

Karni05 [23] 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.

[6] Modified to have consistent notation.

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 $\Theta_{Y=y} = \{(x, y) | x \in \text{supp}(X)\}$.

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 $do(y')$.

3.3.2 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 III.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 $(do(W=w),b)$, $(do(Z=z),b') \in \mathbb{C}$,

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

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 3.6: No effects are elementarily linked

As Figure 3.6 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.

3.4 Applications

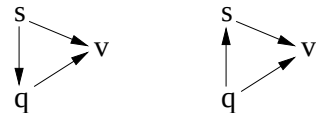
I have described the parts of the causal Bayesian network framework to represent agent's models of phenomena. I have proposed an axiomatic framework by which one can, after observing an agents' choices of interventions and bets, deduce her utility function (a utility function which represents her behavior) and the unique stochastic effect she believes her interventions will result in. This representation in an application of a result by Karni05 [23] 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.

3.4.1 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



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

Figure 3.7: Directed Acyclic Graphs $\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{Q}}$

F. This calibration generates a unique $\widehat{\mathbf{\Phi}}_{\boldsymbol{\mathcal{S}}}$.

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

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 $\widehat{\mathbf{\Phi}}_{\boldsymbol{\mathcal{Q}}}$:

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

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:

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

3.4.2 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 Fudenberg93's [13] "Self-Confirming Equilibrium," exemplified in EysterRabin05 [12] and Esponda05 [11], 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 6. *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 7. 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 Fudenberg93's [13] "Self-Confirming Equilibrium," exemplified in EysterRabin05 [12] and Esponda05 [11]. 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.

3.4.3 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.

[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 3.4.4.

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.

3.4.3.1 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]

[8]See appendix section .2.1.1 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 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

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

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.

3.4.4 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]

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.

[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.

3.5 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.

SlomanLagnado04 [48] 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 AndersonLepperRoss80 [2], 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 [34]. 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 [41] 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 [27]." 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 (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.

3.5.1 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 El61 [8], who treat causal ambiguity in a manner similar to GiSc89's [14] Maxmin expected utility agents. The third would use this framework to construct agents who act in accordance with QuattroneTversky84's [42] empirical finding that people attribute causation to correlation. These agents could be used to derive economic

[11] See Hacking67 [16] for some related issues regarding construction of reasonable beliefs.

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.[3]

Causal ambiguity is a form of ambiguity or Knightian uncertainty. El61 [8] 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. GiSc89 [14] 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 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. QuattroneTversky84 [42] 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.

3.6 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

[12] This is an informal treatment of GiSc89's [14] work. GiSc89's [14] 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.

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.

CHAPTER IV

Of Wolves and Sheep

4.1 Introduction

Rational expectations equilibria in macro- and microeconomic models are generally premised on the idea that agents not only forecasting by using information rationally, but that they use the information correctly; i.e. that they have the correct theory. In those models, the predictions the agents make are the same that the model-maker would make, given the same information. An alternative is proposed in this paper: agents have theories that are internally consistent and are consistent with data; however, like econometricians faced with an under-identified model, these theories may have the wrong functional form. In particular, these agents have data about a positive correlation between CEO skill and firm performance. The agents' theories differ on the direction of causality between CEO skill and firm performance. Some agents believe that the correlation of skill and performance is explained by high CEO skill causing high firm performance, while others believe that high-skilled CEOs are better at finding firms that would, regardless of their CEO, be high performers.

These two theories are consistent with the data they see, but, like econometricians running different regressions on the same data, imply different predictions for out-of-equilibrium (or out-of-sample) behavior. This leads to differing equilibrium behavior

on the part of these agents, and, when these agents are set in a dynamic economy with many agents with differing theories, this leads to competition in theories that is external to the agents (in some sense, a population learning process.)

The dynamic economy in this paper is a series of subgames. These subgames consist of agents who must choose between developing new ideas into firms versus letting others develop firms and buying firms that appear to be successful. Agents who are otherwise identical, but who differ on theories of firm performance, make different choices. Namely, agents who believe firm performance is CEO-lead find it optimal to develop ideas into firms, because the success of the firm is only dependent on their own skill, and therefore information about prior firm performance is irrelevant. On the other hand, agents who believe firm performance is not caused by CEO skill find it worthwhile to let other agents develop firms, and purchase high-performing firms. This creates a self-contained *causally coherent equilibrium* [39] for each subgame.

In this dynamic economy, the initial parameter values of each subgame are determined by the subgame previous; the theory that generates more profits attracts more followers in the next subgame, and the final distribution of CEO skill and firm performance gives the agents of the next subgame a dataset with which to calibrate their theories. The predator/prey aspects of subgame behavior (in which 'waiters' prey on 'developers') causes population cycles to emerge in some of these dynamic economies. This causes waves of corporate takeovers, which is an empirical phenomenon that has been observed in the world. Moreover, the population popularity of theories, in the long run, is a predictor of the 'true' model (the extent to which one theory versus the other is, in fact, true) which suggests that an opinion poll of these agents, in which none have the true theory, could be used to predict the true theory (as said above, a kind of population learning process). Finally, the dynamic system

does not, in general, converge to a state in which one theory prevails, suggesting that when agents have incomplete, but rational theories, continuing disagreement can be a long-run phenomenon. Given that theories in the world inevitably have identifying assumptions, this model suggests that long-run disagreement might be the state of nature.

This paper proceeds in five sections. In the Background and Literature section (section 4.2) which follows, I discuss two main models in the literature related to this paper. I discuss the causal coherence model, which describes differing agents' theories, and the predator/prey model, which is the basis of the population dynamics (some literature related to the conclusions of this model appears in the discussion, in section 4.5). In the Model section (section 4.3), I describe the core of the model, both each subgame and the dynamic system, which is a series of linked subgames. Subsection 4.3.1 describes the two-period subgame of agents who develop ideas, buy each others' ideas, and bring those developed ideas to market. Subsection 4.3.2 describes how the subgames are linked into an overall dynamic system: namely the relative success of agents holding particular theories induces how well those theories are passed to the next subgame of agents, and the outgoing distribution of owners and firms provides the raw data for the next subgame of agents. The Results section (4.4) contains the explicit description of a causally coherent equilibrium of the subgame and simulations of the dynamic system. Subsection 4.4.1 shows that both theories are consistent with the incoming data and what the agents observe in each subgame. Subsection 4.4.2 shows simulated results for the dynamic system with regard to the value of the fixed population ratio and the 'true' model (namely, that one can be used to predict the other) and how and when population cycles emerge, and provides evidence for the claims made above. The Discussion section (4.5) describes in greater

detail the relationship of the results with applications; and the Conclusion section (4.6) provides an overview of the work.

4.2 Background and Literature

This paper builds on two models distinct models. The first is a model of how individuals model problems that they face; it is a decision theory model, about individual choice. The second is a model of predator/prey interaction, and hence is a model of social dynamics.

The first model is the belief model outlined in Causal Coherence [39]. In Causal Coherence, agents confront a phenomenon and generate causal theories about how the phenomenon works. In the ongoing example, the phenomenon is what makes a firm profitable. Agents see the same data about which firms are successful and the characteristics of successful and unsuccessful firms, but they have different theories to explain those data. In particular, they differ about the direction of causal relationships among the characteristics of these firms. Given their different theories, they make different predictions about what will happen when they take over a firm and manipulate those characteristics. In an equilibrium in an auction, these different theories alone gives rise to a phenomenon similar to the winner's curse, where some agents lose out in the auction by overvaluing the firm.

The central component of Causal Coherence is the causal model (or causal theory.) The firms are described by three variables: *Skill* of the CEO of the firm (s), *quality* of the firm (q), and *value* of the firm (v). Some agents believe that *skill* causes *quality* (i.e. high skilled CEOs bring about high quality firms). The typical agent who believes this is called Sam ("S" for skill, since that variable is the root in his mental causal model). Other agents believe that *quality* causes *skill* (i.e. that high

skilled CEOs are able to find firms that are high quality.) The typical agent who believes this theory is called Quincy ("Q" for quality). Sam and Quincy can both explain the correlation of skill and quality that is observed in the data, but with differing theories. When Sam is forecasting the effects of changing the CEO (and therefore changing s), he predicts a change in quality q. When Quincy is forecasting the effects of changing the CEO, he believes there will be no change in quality q. And, since quality affects value, they have different predictions of the change in value of the firm if this change in CEO skill were to occur. Page07 [38] discusses multiple implications for behavior of agents maintaining different forms of models.

The second model in this paper is the predator/prey model from population biology (or, as referred to here, Wolf/Sheep[47]). In this model, briefly, there are three characters: grass, sheep, and wolves. The sheep eat the grass and the wolves eat the sheep, and the grass grows at a constant rate. When wolves or sheep eat enough of their food, they reproduce, and if they eat too little, they die. The equilibrium of this system is expressed in the populations of wolves, sheep, and grass. There are a few uninteresting equilibria: for example, if there are too few sheep relative to wolves, the sheep all get eaten, and then the wolves, without a food source, also die, leading to a world of only grass. A more interesting equilibrium is an unchanging one, in which there are stable population ratios of wolves to sheep to grass, and they eat at a rate which exactly replaces their population (or allows all populations to grow at the same constant rate.) This equilibrium is unstable. The stable equilibrium is a cyclic, dynamic equilibrium. To whit, consider the case when there are many sheep and few wolves. For wolves, there is a bonanza of food, so their population grows. As their population grows, sheep get eaten more quickly. As this happens, the population growth rate of sheep begins to fall, eventually becoming negative.

Since there is now a decreasing supply of sheep, the wolves' population growth rate begins to fall as well, with a lag. The lag is key: the populations of wolves and sheep fall into offset waves or cycles. (In particular, the growth rate of sheep is determined by the population level of wolves and vice versa. In fact, the system can be described by those simple differential equations.) (Note: The population of grass also falls into this cycle, but it was not necessary for this explanation.)

4.3 Model

In this section, the explicit model is introduced. This model connects the causal coherence model discussed above with the wolf/sheep model. In the Sam/Quincy model in Causal Coherence, agents have theories about established firms. In this paper, the agents instead have theories about *product ideas* which can be developed into firms. These product ideas are like grass: they simply appear, ready to be developed by CEOs. Sam believes that the responsibility of the quality of the idea rests with him. That is, he believes that ideas do not have inherent quality; instead, he believes the quality of the product is determined by the *s*kill of the CEO. Since Sam believes every idea is as good as any other, he is like a sheep: he "grazes" indiscriminately. Quincy, on the other hand, believes that the quality is inherent to the idea, and therefore finds it optimal to wait, allow "the Sams" to develop some ideas. Then Quincy will buy one of those ideas that he believes are high quality (because the development process reveals a signal of the inherent quality of the idea). Quincy is like a wolf: he preys on the Sams' ideas, and "eats" a Sam by buying a firm for less than it (a *posteriori*) is worth, thus cutting into Sam's profits. Since Sams are like sheep, the Quincies like wolves, and the firms like grass, the same population dynamics should emerge.

The model consists of two parts: a two-period subgame, and a larger game which consists of an (infinite) series of these subgames. Each subgame begins with a new set of players, and some data is transferred from the previous subgame.

4.3.1 Subgame

A subgame G_t consists of two periods, so there are four phases to the game: before play begins, period one, period two, and after play ends. Before the game, the product ideas are created, players are endowed with skill s and causal models, and a data set enters the subgame exogenously. Players observe the data and calibrate their theories to those data before play begins. In period one, players are given the opportunity to claim product ideas and develop them into firms. (In equilibrium, only Sams will take advantage of this opportunity.) In period two, players who have not developed ideas can choose to buy an existing, developed idea (to be precise, potential sellers post a bid price and potential buyers post an ask price. If the prices are compatible, trade occurs at a price between them). (In equilibrium, all Quincies will choose to do this). After play ends, developed ideas have become firms. Each firm has one owner of some skill s and the firm is assigned a quality q. Firms' value are its quality q. This value accumulates to the CEO of that firm.

Before play begins, three objects come into existence: a set I of product ideas, a set P of players of type S (for Sam) or type Q (for Quincy), and a data set consisting of a distribution F_t. (The distribution will be explained below.)

The set I of product ideas, with component product ideas i, have one characteristic: **original quality** o_i, which is either zero or one (or low and high). o_i can be thought of as the inherent quality of idea i. All o_i are independently distributed, with a fixed probability p_o of equalling one. The original quality o has a role in creation of the firm's final quality q. Quincy's theory is that the skill of the CEO is

irrelevant to final firm quality: in this language, it can be said that Quincy believes that, for all ideas i, $q_i = f(o_i)$, for some function f, while Sam believes that q_i is not caused by o_i, and instead $q_i = g(s_j)$, for some function g, where j is the index of the player who owns firm i.

Players, indexed by j, of both types, have a skill s_j, which is also zero or one (low or high). There is a continuum of agents along the index interval $[0, 1]$. All players know their own skill and not that of any other. All players' skill values are drawn independently, with a fixed probability p_s of equaling one. Furthermore, λ_t^S of the players are Sams, and $\left(1 - \lambda_t^S\right)$ of the players are Quincies, where λ_t^S is an exogenous value which varies for each subgame. λ_t^S is determined by the outcome of subgame G_{t-1}.

The data set is public and observed by all players. The public data consists of the distribution F_t of CEO skill s and firm quality q from the pervious subgame. To whit,

$F_t(s,q)$	q_L	q_H
s_L	a	b
s_H	c	d

where a, b, c, d all correspond to the fraction of firm/CEO pairs that fit that criteria. For example, b designates the fraction of all firm/CEO pairs which were high quality and run by a low-skill CEO.

F_t plays a central role in the calibration of agents' theories. Sam, who believes that skill causes quality, uses F to determine how likely he believes a CEO of each skill level will make his product high quality. Quincy uses F_t to determine how accurate the signals are about true quality.

Play begins in period one. All players may choose to take a product idea and, for

a cost κ, develop the idea. Sams will choose to do this, and Quincies will not. In the beginning of period two, each player j receives a signal σ_i^j about the original quality o_i of idea i, for all ideas i which have been developed (i.e. selected in period one). Signals σ also take on the value of zero or one. The accuracy of these signals will be discussed below.

Players who have not chosen to develop an idea in period one can choose to bid on any existing, developed idea. Quincies will choose to do this: in particular, Quincy will choose to bid on a developed idea at random from the set of ideas about which he received a high signal. (Note that the ideas are otherwise indistinguishable.)

After play ends, each player makes his developed idea into a firm, and firm quality is determined. Firm quality is a stochastic function of both the agent's skill and the original quality of the idea. As mentioned before, Sams believe it is a function only of agents' skill and Quincy of original quality.

The data set F_{t+1} is constructed from the outcome of subgame G_t in a natural way: the final distribution of skill and quality forms F_{t+1}.

4.3.1.1 The Players

All players are risk neutral and value final profits:

$$u_j = v(j) - \text{costs} \tag{4.1}$$

where costs are κ, a random variable, discussed below, if the agent has developed an idea in the period one, and zero otherwise.

$v(j)$ is the value of the final firm that the agent has. If the agent chooses no firm, then $v(j)$ is zero.

All agents have the following actions available to them:

1. In period one, agents can choose to develop an idea.

2. In period two, agents who have not developed ideas can post bid prices for particular firms.

All players also have a theory. This theory consists of three components. The primary component is the direction of causality: either that skill causes quality or quality causes skill. The second component of the subgame is a theory about other agents' preferences: agents believe (counterfactually) that other agents might have different risk preferences: namely that other agents might be risk-loving or risk-averse. The third component is a belief about the distribution of κ, the cost of developing ideas. The theory is allocated to agents upon their inception–however, the calibrate the theory to the data available. That is, the theory provides functional relations among visible variables, and agents take these theories to available data and measure the parameters of the functions that their theory provides.

4.3.1.2 The determination of final quality

Final quality of the firm is determined jointly and stochastically from both the owner's skill and the original quality of the firm. This means that, in fact, the causal models that Sam and Quincy believe are both wrong; or, alternatively, that they are both partially right. The extent to which skill causes quality will be denoted by α (and therefore, the extent to which quality is caused by original quality is $(1-\alpha)$.) In particular, the determination of true final quality is determined in the following way:

Suppose the firm of player j is based on idea i. The quality q_j of that firm is created as a function of both the skill s_j of player j and the original quality o_i of idea i. In particular:

$$Prob(q_j = 1) = \alpha s_j + (1-\alpha)\, o_i \tag{4.2}$$

for some $\alpha \in [0, 1]$.

α is an exogenous parameter which describes the *true causal model*. If $\alpha = 1$, then quality is completely determined by the skill of the CEO, and Sam is correct (recall, s_j is either zero or one). If $\alpha = 0$, then quality is completely determined by original quality, and Quincy is correct. For values of α between zero and one, they are each in part correct.

4.3.1.3 The Outcome of the Subgame

After all players have moved in period one and period two, the profit of each player is determined according to the true causal model (i.e. the function of α above.) At this point, all players have received their payoffs, and there is a new distribution of skill of owners and quality of firm. This new distribution is determined in the obvious way: the fraction of *all* agents (of type both Sam and Quincy) who are low skill and have a high quality firm comprises the 'new' value of b in the new distribution, for example. All players then cease to exist, and the subgame is finished.

4.3.2 The Dynamic Game

The dynamic game (or, more accurately, the dynamic series of subgames) is a connected series of subgames, where the outcomes of subgame G_t determines some initial parameters of subgame G_{t+1}.

In the description of the subgame above, λ_t^S and F_t were taken to be exogenous. In the dynamic series of subgames, those two variables are determined in subgame G_{t-1}.

F_t is determined in the obvious way from the outcome of subgame G_{t-1}. Namely, the new distribution F_t is calculated according to the skill and quality of final owners; i.e. intermediate owners (Sams after the first period) play no role. If, for example, a

fraction a_{t-1} of all players were both low-skill and ran low-quality firms at the end of subgame G_{t-1}, and a fraction b_{t-1} of all players were low-skill and ran high-quality firms, and so on, then F_{t+1} would be constructed as:

$F_{t+1}(s,q)$	q_L	q_H
s_L	a_t	b_t
s_H	c_t	d_t

This means the players in subgame G_{t+1} are able to observe the final distribution of the previous subgame and base their theories on that subgame.

The other parameter of subgame G_{t+1} which is determined by the outcome of subgame G_t is λ^S_{t+1}. λ^S_{t+1} is constructed from the fraction of average profits made by Sam. In particular:

$$\lambda^S_{t+1} = \frac{\text{Average Profit of Sams}}{\text{Average Profit of Quincies}} \tag{4.3}$$

The total profit of Sams is calculated according to:

1. For the fraction of Sams j with low original quality ($(1-p_o)$ of them), they receive αs_j minus the true cost of development;

2. For those Sams with high original quality, they receive either the sale price if their firm was sold, or the true average q given above, minus the true cost of development.

The total profit of Quincies is calculated according to:

1. For low-skilled Quincies, they receive zero,

2. For high-skilled Quincies, they either receive true average q above minus the sale price, or zero, if they were unable to buy a firm.

4.3.3 Conclusion to the Model Section

The above fully describes the subgames, the creation of agents, and how subgames form a dynamic system. If there is an equilibrium for each possible subgame (i.e. for every F and λ^S) then, each subgame can act as a mapping from F and λ^S to a new F and λ^S. This mapping constructs a difference model, the properties of which can be studied.

In the Results section 4.4 below, I propose a complete equilibrium of the subgame, and then show the implications for the dynamic system, given those equilibria.

4.4 Results

4.4.1 The Causally Coherent Equilibrium of the Subgame

When agents have distinct theories that can both be made to describe the data, it can be possible to construct a *causally coherent equilibrium*[39]. In this equilibrium, (1) all agents believe the theories specified in equilibrium, (2) all agents are best responding to their opponents, given their beliefs about their opponents' theories, (3) they believe their opponents' theories match their own (that is, they are not aware of alternative theories), and (4) they see no outcomes in equilibrium which are *impossible* given their theories.

The causally coherent equilibrium must specify actions for all players and theories for all players, such that those conditions are met. The following theorem describes such an equilibrium of the subgame:

Theorem IV.1. *There exists a causally coherent equilibrium in each subgame (for all λ and F), where F*

$F(s,q)$	q_L	q_H
s_L	a	b
s_H	c	d

In which all Sams:

1. *Choose to develop a firm (i.e. move in period one.)*

2. *Have the theory that skill causes quality ($q_i = g(s_j)$) and that next period movers are more risk-averse*

3. *Which they calibrate such that, if p_s^s is the probability that an agent of skill s will make an arbitrary idea into a high-quality firm. Sam's estimates of these values will be denoted by:*

$$\hat{p}_L^s = \frac{\#\textit{ high quality firms run by low-skill CEOS}}{\textit{total }\#\textit{ firms run by low-skill CEOS}} = \frac{b}{a+b} \tag{4.4}$$

$$\hat{p}_H^s = \frac{\#\textit{ high quality firms run by high-skill CEOS}}{\textit{total }\#\textit{ firms run by high-skill CEOS}} = \frac{d}{c+d} \tag{4.5}$$

And costs κ are distributed according to an arbitrary distribution with mean $\frac{b}{a+b}$.

And in which all Quincies:

1. *Choose to not develop a firm, and instead buy one, if they are high skill, or do nothing, if they are low skill,*

2. *Have a theory in which quality is inherent to the idea ($q_i = f(o_i)$) and that first period movers are more risk-loving,*

3. *Which they calibrate such that they believe, if γ_s^q is the accuracy of a signal σ_i^j that Quincy j receives about developed idea i, when Quincy j has skill s, then:*

$$\gamma_L^q = \frac{1}{2} \tag{4.6}$$

$$\gamma_H^q = \frac{(a+c)d}{bc + (a+2c)d} \tag{4.7}$$

To demonstrate that this is a causally coherent equilibrium, three things are necessary. First, the true actions and outcomes for all players are established. Second, it must be shown that this outcome is consistent with Sam's theory. Third, it must be shown that this outcome is consistent with Quincy's theory.

In the true outcome of the subgame, Sams all move in the first period, and offer an ask price $p < \frac{d}{c+d}$. All Quincies move in the second period, offering a bid price p above $b + d$. There are prices, therefore, in the range $\left(b + d, \frac{d}{c+d}\right)$.[1] The outcome of the subgame, moreover, is that get a random κ cost from the true distribution and a value (final quality) of zero or one. All these facts must be compatible with each of Sam and Quincy's theory and information for this to be a causally coherent equilibrium.

4.4.1.1 The Outcome is Consistent with Sam's Theory

Agents of type Sam have the theory that skill determines quality of the firm. If that is the case, then the process by which ideas were attached to owners was irrelevant to final quality. Sam reviews the data from the previous subgame to evaluate how effective each skill level is at producing quality. The data he has available is F_t. He uses F_t to estimate (p^s_L, p^s_H).[2] p^s_s is the probability that an agent of skill s will make an arbitrary idea into a high-quality firm. Sam's estimates of these values will be denoted by:

$$\hat{p}^s_L = \frac{\text{\# high quality firms run by low-skill CEOS}}{\text{total \# firms run by low-skill CEOS}} = \frac{b}{a+b} \tag{4.8}$$

$$\hat{p}^s_H = \frac{\text{\# high quality firms run by high-skill CEOS}}{\text{total \# firms run by high-skill CEOS}} = \frac{d}{c+d} \tag{4.9}$$

This allows one to construct Sam's belief of the value of a particular idea, when developed into a firm.

[1] Note that, since skill and quality are correlated, $b + d < \frac{d}{c+d}$.

[2] The superscript S is to suggest that this probability corresponds to Sam.

Sam believes there is be a cost associated with developing an idea. Sam believes the cost κ is a random variable with distribution μ and expected value $\bar{\kappa}$, where $\bar{\kappa} = \frac{b}{a+b}$.

Suppose Sam believes there is a price p at which firms will be bought and sold in the second period. Therefore, Sam must choose between developing a firm in the first period or waiting to buy a developed firm in the second period. His payoffs are dependent on his skill. The payoffs are:

If Sam is high skill, then

- The payoff to developing a firm in the first period is:

$$\left(\frac{d}{c+d} - \bar{\kappa}\right)$$

  or $(p - \bar{\kappa})$ if he decides to sell it, and

- The payoff to waiting until the second period is:

$$\left(\frac{d}{c+d} - p\right)$$

  assuming he can buy a firm at that price.

If Sam is low skill, then

- The payoff to developing a firm in the first period is:

$$\left(\frac{b}{a+b} - \bar{\kappa}\right)$$

  or $(p - \bar{\kappa})$ if he decides to sell it, and

- The payoff to waiting until the second period is:

$$\left(\frac{b}{a+b} - p\right)$$

  assuming he can buy a firm at that price.

Given these choices, Sam might be better acting in the first period or waiting; it depends on the value of p.

Sam's theory must both induce Sam to act in the first period and provide an explanation as to why other agents with presumably the same theory would choose to act in the second period. His explanation is that there are some agents who are risk-averse, but otherwise the same. These risk-averse agents are hoping to get rid of the risk associated with the cost random variable. The choices they face are:

If this risk-averse Sam is high skill, then

- The payoff to developing a firm in the first period is:

$$\left(\frac{d}{c+d}\int\left(1-\bar{\kappa}\right)d\mu(\kappa)+\left(1-\frac{d}{c+d}\right)\int u\left(-\bar{\kappa}\right)d\mu(\kappa)\right)$$

  or $\int u\left(p-\kappa\right)d\mu(\kappa)$ if he decides to sell it, and

- The payoff to waiting until the second period is:

$$\left(\frac{d}{c+d}u(1-p)+\left(1-\frac{d}{c+d}\right)u(-p)\right)$$

  assuming he can buy a firm at that price.

If this risk-averse Sam is low skill, then

- The payoff to developing a firm in the first period is:

$$\left(\frac{b}{a+b}\int\left(1-\bar{\kappa}\right)d\mu(\kappa)+\left(1-\frac{b}{a+b}\right)\int u\left(-\bar{\kappa}\right)d\mu(\kappa)\right)$$

  or $\int u\left(p-\kappa\right)d\mu(\kappa)$ if he decides to sell it, and

- The payoff to waiting until the second period is:

$$\left(\frac{b}{a+b}u(1-p)+\left(1-\frac{b}{a+b}\right)u(-p)\right)$$

  assuming he can buy a firm at that price.

Sam believes that risk-averse, high-skilled Sams will wait until second period and offer some bid price. The low-skill Sam must consider what price to set as an ask price and what price to expect as a bid price. This is a complex problem for the low-skill Sam; he would like to ask $\frac{d}{c+d}$, which is the expected value of the firm to the second-period Sam. However, the second-period Sam would only accept such a high price if he were actually risk-neutral (since he is accepting the risk associated with whether the firm will be high or low quality.) So the low-skilled Sam offers an ask price below $\frac{d}{c+d}$, and expects some bid price above $\frac{b}{a+b}$.

4.4.1.2 The Outcome is Consistent with Quincy's Theory

Quincy has a theory that the skill of CEOs determines their ability to identify high-quality, developed ideas (or, developed ideas that will make high-quality firms.) The variable that Quincy is interested in estimating with the data F_t are γ_H^q and $\bar{\kappa}$, the mean cost of developing an idea. γ_s^q is the accuracy of a signal σ_i^j that Quincy j receives about developed idea i, when Quincy j has skill s. In other words, suppose that Quincy j is of high skill, and gets signal $\sigma_i^j = 1$ about developed idea i. This signal equals the true original quality with probability γ_H^q. Quincy believes that the low-quality signal γ_L^q is $\frac{1}{2}$; i.e. that it contains no information.

Quincy believes that some risk-loving, low-skilled agents found it worthwhile to develop ideas in the first period while high-skilled agents found it worthwhile to wait until the second period and bid on an existing idea. Applying this theory to the observed data allows Quincy to estimate γ_H^q.

Suppose this was the case, and the final distribution F_t was, as before,

$F_t(s,q)$	q_L	q_H
s_L	a	b
s_H	c	d

If, in period one, only low-skilled agents moved, then they were allocated high-quality ideas/firms in proportion to the amount of high-quality in the system overall. The total number of high- and low- quality firms will not change between periods one and two (by Quincy's theory.) Therefore, the allocation after period one would be:

$F_{t:1}(s,q)$	q_L	q_H
s_L	$(a+c)$	$(b+d)$
s_H	0	0

Then, high skill agents, who enter in period two, successfully bid on firms for which they have received a high signal. The ratio of d (high skill, high quality) to c (high skill, low quality) must equal the original likelihood, times the rates given by the accuracy of the signal:

$$\frac{d}{c} = \frac{\gamma_H^q}{(1-\gamma_H^q)} \frac{b+d}{a+c} \tag{4.10}$$

$$\implies \gamma_H^q = \frac{(a+c)d}{bc+(a+2c)d} \tag{4.11}$$

Therefore, F_t allows Quincy to calculate γ_H^q, and he calibrates accordingly.

For this theory to dictate this behavior, Low-skilled Quincies and High-skilled Quincies must both believe that $\bar{\kappa}$ is sufficiently high as to make investing in period one unprofitable on average; in other words, that $\bar{\kappa}$ is at least $(b+d)$ (which is the probability, Quincy believes, of getting a high-quality firm on average.) On the other hand, Quincy sees that there are agents who have entered in the first period. How does Quincy explain this phenomenon? It would only make sense, for sufficiently

small costs κ (that is, close to $(b+d)$), if there were risk-loving, low-skilled agents. This belief allows Quincy to rationalize the outcome of the previous subgame.

Given these estimates, high-skilled Quincies find it profitable to wait until the second period to bid on existing firms, and low-skilled Quincies find themselves better off by opting out of the process altogether.

Given the price mechanism, high-skilled Quincies would like to set the lowest bid price to capture what he believes are a monolithic set of risk-loving agents who moved in the first turn. That price must certainly be above $b+d$, but how far above depends on Quincy's beliefs about the risk-lovingness of these agents.

Thus, any price in the range $\left(b+d, \frac{d}{c+d}\right)$ is consistent with Sam and Quincy's expectations.

This finishes the proof that this is a causally coherent equilibrium.

4.4.2 Results of the Dynamic Game

The dynamic game is constructed by setting an initial ratio of Sams and Quincies and an initial distribution F, and allowing the repeated subgames to unfold until some kind of convergence in λ^S emerges. In some cases, the dynamic game does not settle on a particular value, but, instead, goes into a cycle. The hypotheses, which are confirmed by simulation but not yet formal proof, are:

Hypothesis One

The average, long-run value of λ^S is a positive monotonic function of α. This function is a function of parameters of the model (the fraction of high skilled agents, the fraction of high original quality ideas.) This reveals that the average λ^S in fact carries the same information as α. That is to say, that, although individual agents are incapable of learning the 'truth,' the population ratio contains the truth.

However, it is not the case, in general, that the average value of λ^S is equal to α, which one might suspect. This arises from the fact that the true distribution of κ allows for arbitrary re-distribution of profits between Sams (who are first-period movers, and therefore bear the cost κ) and Quincies. In particular, however, there is a value of the mean of the distribution of true κ which allows for average λ^S to equal α. However, there is no mechanism within the model which would make this equality hold.

Consider the following graph, which is of convergent λ^S as a function of α, holding all other parameters fixed at:

$F(s,q)$	q_L	q_H
s_L	0.4	0.1
s_H	0.2	0.3

$$\lambda^S = .4 \tag{4.12}$$

$$p_o = 0.5 \qquad p_s = 0.5 \qquad \kappa = 0 \tag{4.13}$$

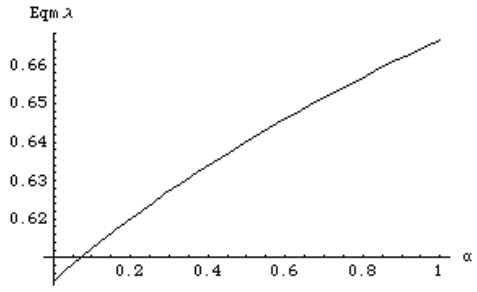


Figure 4.1: Lambda is a positive function of alpha

Hypothesis Two

There are parameter values of this model which cause population cycles to emerge. The population cycle of Sams and Quincies arises from the following two forces at work:

When there are few Quincies, the high-skill Quincies can all be served, and they get high profits per Quincy; moreover, as their population grows, they cut into the profits of the Sams. When there are many Quincies, they crowd each other out; each additional Quincy does not generate profit (hence lowering the average profit of Quincies) but does not lower the profits of Sams any further.

The following simulations of the dynamic game are constructed in the following way. The causally coherent behaviors, described in theorem IV.1, are programmed in on an individual level. Then, the behavior and outcomes for each subgroup of agents is calculated. For example, high-skill Sams or low-skill Sams whose firms are purchased. The mass of each of these subgroups are also calculated; this allows for the calculation of average profit for Sams and Quincies, and the final allocation of skill level with quality level. Those pieces of information are used to construct the outgoing λ^S and F, and then the simulation is repeated with the new values.

Below are two example runs; the first shows an example in which population cycles do not emerge. In the second, the population cycles emerge and converge to a long-run λ^S.

4.4.2.1 Example 1: A stable, low level of Quincies

Initial Values and parameters:

$F(s,q)$	q_L	q_H
s_L	0.4	0.1
s_H	0.2	0.3

$$\lambda^S = .4 \tag{4.14}$$

$$\alpha = 0.5 \qquad p_o = 0.5 \qquad p_s = 0.5 \qquad \kappa = 0 \tag{4.15}$$

With these initial values and parameter values, the system converges to a steady state of a fixed ratio of Sams to Quincies. This is analogous to the constant population ratios of wolves, sheep, and grass mentioned in the background.

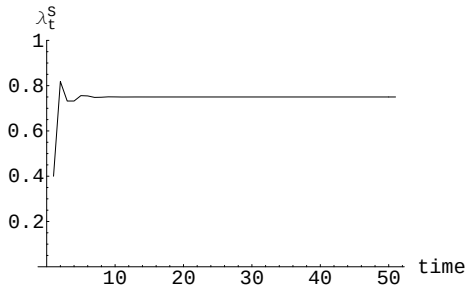


Figure 4.2: No cycles emerge

4.4.2.2 Example 2: Rapid Population Cycles

Example two uses the same initial values and parameters as the previous example, but reduces p_o; that is, reduces the prevalence of original quality in the population.

$$p_o = 0.1 \tag{4.16}$$

This is enough to cause the Quincies to bump into their crowding-out limit early enough to induce cycles:

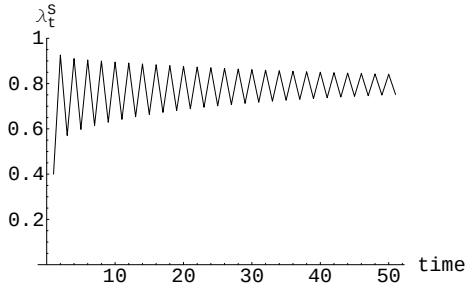


Figure 4.3: Rapid Population Cycles

The population cycles that have been induced are severe; that is, they cycle back and forth between (roughly) two values. Population cycles that we would expect to observe would be more gradual, and this suggests a weakness of this modeling

framework. However, given the binary nature of all variables in this model, this outcome is not surprising: if there were a distribution of skills and perceived signal strengths, one would expect the Quincies to be more gradually crowded out, which, I hypothesize, would yield a smoother set of population curves.

Note that the true α in both these settings is .5, well below the convergent λ^S.

4.5 Discussion

There are several key points to this combined model.

First, the predator and prey roles emerge only because Sam and Quincy have different theories about how firms work. They see the same data (the outcomes of the previous subgame of Sams and Quincies bringing ideas to the market), they have the same preferences, and they have the same actions available. Sam *chooses* to move on the ideas instead of waiting, and Quincy *chooses* to wait instead of move. Other than their theories, the Sams and Quincies are identical. This differentiates from the classic Wolf/Sheep model, in which wolves are simply biologically incapable of eating grass. There are existing models in the macro literature that involve endogenous cycles, most recently Matsuyama99 [31], based on RiveraBatizRomer91's [44] model of endogenous innovation. In Matsuyama's paper the macro economy follows a cyclic growth path that alternates between periods of innovation and periods of asset accumulation. These papers differ significantly from this paper: they are driven by the decisions of a representative consumer. The representative consumer has no role in this paper, in which agents do not have fundamentally compatible beliefs.

Second, the cyclicality is emergent. That is, there are stochastic inputs into the system from stationary distributions, and cycles emerge endogenously. This is in contrast with Macro models in which there must be exogenous shocks (magnified by

the system) for there to be business cycles within the model. Moreover, this emergent cyclicality may explain waves of firm takeovers or waves of increased competition that seem heretofore inadequately explained.

Third, this model demonstrates that in a reasonable and interesting setting, agents with incompatible but rationalizable theories can co-exist (albeit, not peacefully.) The differing theories of Sam and Quincy do not get learned away in the long run: on the contrary, they continue to believe their incompatible theories forever.

Fourth, in this model, the long run ratio of Sams to Quincies is a monotonic function of α, which describes the true model. Hence, the ratio can be interpreted as a way that the population learns the truth after a fashion, even though no individual agent is capable of learning the truth.

4.5.1 Discussion of the Subgame

The following table describes agents' believed average quality versus the true average quality ex-post:

Expected Quality for Players

type	skill	expected q	true s component	true o component	true avg q
Sam	s_L	$\frac{b}{a+b}$	0	p_o	$(1-\alpha)(p_o)$
Sam	s_H	$\frac{d}{c+d}$	1	p_o	$\alpha\,(1)+(1-\alpha)\,(p_o)$
Quincy	s_L	n/a	n/a	n/a	n/a
Quincy	s_H	$\gamma_H^q(b+d)$	1	1	1

Note that the true average q for the Sams and low-skill Quincies excludes the fact that Sams who have a high original quality are more likely to be purchased by a Quincy. The true value for the Quincy, on the other hand, corresponds to the value of the firm he has purchased.

Expected value for Quincy is: $\gamma_H^q\left((b+d)1\right)+(1-\gamma_H^q)\left((a+c)0\right)=\gamma_H^q(b+d)$

High-skilled Quincies are able to buy at a bargain those firms with high original quality, and therefore make both high- and low-skilled Sams worse off for accepting their offers. There is no reason, in general, to believe that Quincies will purchase from the high-skill Sams; there would be no particular reason for Quincy to choose to offer a bid price as high as $\frac{d}{c+d}$, which would be required. Instead, Quincies may only buy from low-skill Sams. Nevertheless, the offer agreed upon can be below the true value that the low-skilled Sam would receive.

4.6 Conclusion

This model serves four purposes: it shows that agents with differing theories can co-exist, without particular theories dying out; it shows how cycles of takeovers can emerge endogenously in a model with no external cyclic shocks; it shows that this behavior can arise from different theories with regard to the same data alone; and it shows that the long-run population ratio of theories can provide a measure of the true model, which allows for 'population learning' which transcends what individuals can learn.

CHAPTER V

Conclusion

Three essays investigating the construction and implications of economic agents' internal representations of problems they face. In the second chapter, "Optimal Auctions under Ambiguity," we investigate the construction of an optimal auction mechanism when agents are ambiguity averse over the valuation of the other bidder. In the third chapter, "Causal Coherence," I investigate agents with differing mental models of the same phenomenon. 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 the fourth chapter, "Of Wolves and Sheep," I place the agents developed in the previous chapter into an economy. In this simple dynamic economy, agents with different theories of how ideas develop into firms leads them to choose different optimal take-up of these ideas. Their different behaviors yields a predator/prey relationship among these agents, which causes natural population

cycles of theories and behavior to emerge endogenously. The agents are identical but for their theories (identical data, actions, preferences) so the predator/prey relationship emerges only from their different interpretations of common data. Since the system does not collapse, it shows that agents with differing theories may persist in a long-run, dynamic equilibrium.

APPENDICES

.1 Chapter 2 (Appendix)

.1.1 Revelation Principle

An (indirect) selling mechanism is a set of possible bids $\mathcal{B}_i$ for each bidder, an allocation rule $a_i : \mathcal{B}_i \times \mathcal{B}_{-i} \to [0,1]$ for each bidder such that $a_i \geq 0$ and $a_1 + a_2 \leq 1$ and a payment rule $m_i : \mathcal{B}_i \times \mathcal{B}_{-i} \to \mathbb{R}$ for each bidder. Each mechanism defines a game of incomplete information where strategies $\beta_i : \Theta \to \mathcal{B}_i$ is an equilibrium if for each bidder and for all θ given β_{-i}, $\beta_i(\theta)$ maximizes bidder i's maxmin expected payoff:

$$\inf_{G \in \Delta_B} \int_\Theta \left(a_i \left(\beta_i(\theta), \beta_{-i}(\theta') \right) \theta - m \left(\beta_i(\theta), \beta_{-i}(\theta') \right) \right) dG(\theta').$$

A direct revelation mechanism is a mechanism where $\mathcal{B}_i = \Theta$ for both bidders. The revelation principle says that given a mechanism and an equilibrium of that mechanism, there exists a direct mechanism in which it is an equilibrium for each bidder to bid his value truthfully and the outcomes of the truthful equilibrium in the direct mechanism are the same as the given equilibrium in the original mechanism.

Next we show that the revelation principle holds in our setting. To see this suppose $(\mathcal{B}_i, a_i, m_i)$, $i = 1, 2$ describes the original mechanism and (β_1, β_2) is an equilibrium of that mechanism. Let $x_i(\theta_i, \theta_{-i}) = a_i \left(\beta_i(\theta), \beta_{-i}(\theta_{-i}) \right)$ and $t_i(\theta_i, \theta_{-i}) = m_i \left(\beta_i(\theta), \beta_{-i}(\theta_{-i}) \right)$. To see that bidding truthfully for each bidder is an equilibrium in the direct mechanism note that,

$$\begin{aligned}
&\inf_{G \in \Delta_B} \int_\Theta (a_i \left(\beta_i(\theta), \beta_{-i}(\theta') \right) \theta \\
&- m \left(\beta_i(\theta), \beta_{-i}(\theta') \right)) dG(\theta') \\
&\geq \inf_{G \in \Delta_B} \int_\Theta \left(a_i \left(\beta_i\left(\hat{\theta}\right), \beta_{-i}(\theta') \right) \theta - m \left(\beta_i\left(\hat{\theta}\right), \beta_{-i}(\theta') \right) \right) dG(\theta')
\end{aligned}$$

implies that

$$\inf_{G\in\Delta_B}\int_\Theta \left(x_i\left(\theta,\theta'\right)\theta - t_i\left(\theta,\theta'\right)\right)dG\left(\theta'\right) \geq \inf_{G\in\Delta_B}\int_\Theta \left(\left(x_i\left(\hat{\theta},\theta'\right)\theta - t_i\left(\hat{\theta},\theta'\right)\right)\right)dG\left(\theta'\right).$$

Finally, by construction the outcomes of the truthful equilibrium in the direct mechanism are the same as the given equilibrium in the original mechanism.

.1.2 Proof of Proposition II.1

Fix a mechanism (x,t). Let

$$K(\theta) = \inf_{G\in\Delta_B}\int_\Theta q(\theta,\theta')dG(\theta')$$

so that $K(\theta)$ is bidder θ's maxmin expected payoff. For any $\theta \in \Theta$, define the function $\delta\left(\theta,\cdot\right):\Theta\to\mathbb{R}$ by

$$\delta(\theta,\theta') = q(\theta,\theta') - K(\theta) \text{ for all } \theta'\in\Theta.$$

Let $t'(\theta,\theta') = t(\theta,\theta') + \delta(\theta,\theta')$ and consider the mechanism (x,t').

We prove the proposition in several steps. In the first step we show that (x,t') is a full insurance mechanism. Furthermore, it leaves the bidders' payoffs unchanged under truth-telling and therefore it is individually rational.

To see that (x,t') is a full insurance mechanism consider an arbitrary bidder $\theta\in\Theta$ and note that,

$$\begin{aligned} x(\theta,\theta')\theta - t'\left(\theta,\theta'\right) &= x(\theta,\theta')\theta - t(\theta,\theta') - \delta(\theta,\theta') \\ &= q(\theta,\theta') - q(\theta,\theta') + K(\theta) = K(\theta). \end{aligned}$$

Thus bidders' payoffs under truth telling are unchanged since

$$\inf_{G\in\Delta_B}\int_\Theta \left[x(\theta,\theta')\theta - t'(\theta,\theta')\right]dG(\theta') = K(\theta).$$

In the second step of the proof we show that (x, t') is incentive compatible. The payoff for $\theta \in \Theta$ to deviate to an arbitrary $\tilde{\theta} \in \Theta$, $\theta \neq \tilde{\theta}$, is:

$$\begin{aligned}
&\inf_{G\in\Delta_B} \int_\Theta \left[x(\tilde{\theta}, \theta')\theta - t'(\tilde{\theta}, \theta')\right] dG(\theta') \\
&= \inf_{G\in\Delta_B} \int_\Theta \left[x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta') - \delta(\tilde{\theta}, \theta')\right] dG(\theta') \\
&\leq \inf_{G\in\Delta_B} \int_\Theta \left(x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta')\right) dG(\theta') - \inf_{G\in\Delta_B} \int_\Theta \delta(\tilde{\theta}, \theta') dG(\theta').
\end{aligned}$$

The inequality above follows since the sum of the infimum of two functions is (weakly) less than the infimum of the sum of the functions. But note that

$$\inf_{G\in\Delta_B} \int_\Theta \delta(\tilde{\theta}, \theta') dG(\theta') = \inf_{G\in\Delta_B} \int_\Theta \left[q(\tilde{\theta}, \theta') - K(\tilde{\theta})\right] dG(\theta') = 0.$$

Combining this with (**??**) implies that

$$\inf_{G\in\Delta_B} \int_\Theta \left[x(\tilde{\theta}, \theta')\theta - t'(\tilde{\theta}, \theta')\right] dG(\theta') \leq \inf_{G\in\Delta_B} \int_\Theta \left[x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta')\right] dG(\theta').$$

Now the payoff for type θ to truth-telling in (x, t') must be weakly larger than the last expression, because the mechanism (x, t) was assumed to be incentive compatible, and by the first step the truth telling payoffs are unchanged. Thus (x, t') is incentive compatible.

In the third step we show that the seller is weakly better off using (x, t'). To see this first note

$$\begin{aligned}
&\inf_{G\in\Delta_S} \int_\Theta \int_\Theta \left[t'(\theta, \theta') + t'(\theta', \theta)\right] dG(\theta) dG(\theta') \\
&= \inf_{G\in\Delta_S} \int_\Theta \int_\Theta 2t'(\theta, \theta') dG(\theta) dG(\theta') \\
&= \inf_{G\in\Delta_S} \left[\int_\Theta \int_\Theta 2t(\theta, \theta') dG(\theta) dG(\theta') + \int_\Theta \int_\Theta 2\delta(\theta, \theta') dG(\theta) dG(\theta')\right] \\
&\geq \inf_{G\in\Delta_S} \int_\Theta \int_\Theta 2t(\theta, \theta') dG(\theta) dG(\theta') + \inf_{G\in\Delta_S} \int_\Theta \int_\Theta 2\delta(\tilde{\theta}, \theta') dG(\theta') dG(\tilde{\theta}).
\end{aligned}$$

Moreover for any $G \in \Delta_S$,

$$\begin{aligned}
&\int_\Theta \int_\Theta \delta(\theta, \theta') dG(\theta') dG(\theta) \\
&= \int_\Theta \int_\Theta \left(q(\theta, \theta') dG(\theta') - K(\theta) \right) dG(\theta) \\
&= \int_\Theta \left[\int_\Theta q(\theta, \theta') dG(\theta') - \inf_{G' \in \Delta_B} \int_\Theta q(\theta, \theta') dG'(\theta') \right] dG(\theta) \geq 0.
\end{aligned}$$

Combining equations (**??**) and (**??**) we see that the seller is weakly better off using (x, t').

Finally we show that if there exists some positive measure $\widetilde{\Theta} \subseteq \Theta$ such that for any $\tilde{\theta} \in \widetilde{\Theta}$

$$\inf_{G \in \Delta_S} \int_\Theta q(\tilde{\theta}, \theta') dG(\theta') > \inf_{H \in \Delta_B} \int_\Theta q(\tilde{\theta}, \theta') dH(\theta')$$

then the seller strictly prefers (x, t') to (x, t). To see this note that,

$$\begin{aligned}
&\inf_{G \in \Delta_S} \int_\Theta \int_\Theta \delta(\tilde{\theta}, \theta') dG(\theta') dG(\tilde{\theta}) \\
&= \inf_{G \in \Delta_S} \int_\Theta \left[\int_\Theta q(\tilde{\theta}, \theta') dG(\theta') - \inf_{H \in \Delta_B} \int q(\tilde{\theta}, \theta') dH(\theta') \right] dG(\tilde{\theta}) > 0.
\end{aligned}$$

The strict inequality follows because for all $G \in \Delta_S$ the expression inside the integral is greater than zero for all $\tilde{\theta} \in \widetilde{\Theta}$ and, by assumption, the event $\widetilde{\Theta}$ gets strictly positive weight for all distributions in Δ_S. Combining equations (**??**) and (**??**) we conclude that the seller strictly prefers the mechanism (x, t').

This completes the proof.

.1.3 Proof of Proposition II.2

Note that since in this paper we deal in environments where the bidders' valuations are drawn independently, restricting attention to mechanisms where the transfers are uniformly bounded is without any loss of generality as far as search for optimal selling mechanism is concerned.

In our proof we will use the following definitions and results. Suppose that p and q are conjugate indices, i.e. $1/p + 1/q = 1$. If $p = 1$ then the conjugate is $q = \infty$. Suppose that $f^n \in L_p(\Theta, \Sigma, \tilde{\mu})$ for $n \in \{1, 2, ...\}$. (From now on we will write L_p instead of $L_p(\Theta, \Sigma, \tilde{\mu})$ for notational simplicity.) We say that f^n converges weakly to $f \in L_p$ if $\int g f^n d\tilde{\mu}$ converges to $\int g f d\tilde{\mu}$ for all $g \in L_q$.

Let $\text{ca}(\Sigma)$ be the set of countably additive probability measures on (Θ, Σ).

Chateauneuf, Maccheroni, Marinacci and Tallon [29] prove that when $\Delta \subset \text{ca}(\Sigma)$ is weakly compact and convex then there is a measure $\tilde{\mu} \in \Delta$ such that all measures in Δ are absolutely continuous with respect to $\tilde{\mu}$. Using this result we fix $\tilde{\mu}$ to be a measure such that $\mu << \tilde{\mu}$ for all $\mu \in \Delta_B^m \cup \Delta_S^m$.

For each $\mu \in \Delta_B^m \cup \Delta_S^m$ there exists a Radon-Nikodym derivative $f \in L_1(\tilde{\mu})$. By the Radon-Nikodym Theorem, there is an isometric isomorphism between $\text{ca}(\tilde{\mu})$ and $L_1(\tilde{\mu})$ determined by the formula $\mu(A) = \int_A f d\tilde{\mu}$ (see Dunford and Schwartz [7], p. 306). Hence, a subset is weakly compact in $\text{ca}(\tilde{\mu})$ if and only if it is in $L_1(\tilde{\mu})$ as well.

Let $\tilde{\Delta}_B$ and $\tilde{\Delta}_S$ be the set of Radon-Nikodym derivatives of measures in Δ_B^m and Δ_S^m with respect to $\tilde{\mu}$ respectively.

Finally, let

$$\mathcal{B}_\infty^r = \{g \in L_\infty : \|g\|_\infty \leq r\}.$$

By theorem 19.4 in Billingsley [5], $\mathcal{B}_\infty^r$ is weakly compact.

Now we turn to the proof.

Proof of Proposition II.2 First we show that the minimizing set of priors is nonempty in (2.1). Let

$$g_{\tilde{\theta}\theta}(\theta') = x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta').$$

Recall that we assume $|t(\theta, \theta')| \leq K$ for some $K > 0$. In other words transfers are

uniformly bounded. Therefore by assumption $g_{\tilde{\theta}\theta} \in L_\infty$.

Now suppose that $f^n \in \tilde{\Delta}_B$ is such that

$$\int g_{\tilde{\theta}\theta} f^n d\tilde{\mu}$$

converges to

$$\inf_{f \in \tilde{\Delta}_B} \int g_{\tilde{\theta}\theta} f d\tilde{\mu}.$$

Since $\tilde{\Delta}_B$ is weak compact, by passing to a subsequence we can find $\bar{f} \in \tilde{\Delta}_B$ such that f^n weakly converges to $\bar{f}$. Thus,

$$\bar{f} \in \arg\min_{f \in \tilde{\Delta}_B} \int g_{\tilde{\theta}\theta} f d\tilde{\mu}.$$

This proves that the minimizing set of priors is nonempty in the IC and IR constraints.

Now, we show that the minimizing set of priors in the seller's objective function is nonempty. Suppose $f^n \in \tilde{\Delta}_S$ is such that

$$\iint t(\theta, \theta') f^n(\theta) f^n(\theta') d\tilde{\mu}(\theta) d\tilde{\mu}(\theta')$$

approaches to

$$\inf_{f \in \tilde{\Delta}_S} \iint t(\theta, \theta') f(\theta) f(\theta') d\tilde{\mu}(\theta) d\tilde{\mu}(\theta').$$

Since $\tilde{\Delta}_S$ is weak compact, by passing to a subsequence we can find $\bar{f} \in \tilde{\Delta}_S$ such that f^n weakly converges to $\bar{f}$. Thus $\int t(\theta, \theta') f^n(\theta) d\tilde{\mu}(\theta)$ converges to $\int t(\theta, \theta') \bar{f}(\theta) d\tilde{\mu}(\theta)$. Let

$$g^n(\theta') = \int t(\theta, \theta') f^n(\theta) d\tilde{\mu}(\theta)$$

and let

$$g(\theta') = \int t(\theta, \theta') \bar{f}(\theta) d\tilde{\mu}(\theta).$$

Consider $\int g^n(\theta') f^n(\theta') d\tilde{\mu}(\theta')$. Note that,

$$\begin{aligned}
&\left|\int g^n(\theta') f^n(\theta') d\tilde{\mu}(\theta') - \int g(\theta') \bar{f}(\theta') d\tilde{\mu}(\theta')\right| \\
&\leq \left|\int g^n(\theta') f^n(\theta') d\tilde{\mu}(\theta') - \int g^n(\theta') \bar{f}(\theta') d\tilde{\mu}(\theta')\right| \\
&+ \left|\int g^n(\theta') \bar{f}(\theta') d\tilde{\mu}(\theta') - \int g(\theta') \bar{f}(\theta') d\tilde{\mu}(\theta')\right| \\
&\leq |K+1| \left|\int \left(f^n(\theta') - \bar{f}(\theta')\right) d\tilde{\mu}(\theta')\right| \\
&+ \left|\int \left(g^n(\theta') - g(\theta')\right) \bar{f}(\theta') d\tilde{\mu}(\theta')\right|.
\end{aligned}$$

The first term goes to zero. To see that the second term also goes to zero note

$$\begin{aligned}
&\left|\int \left(g^n(\theta') - g(\theta')\right) \bar{f}(\theta') d\tilde{\mu}(\theta')\right| \\
&= \left|\int \left(\int t(\theta,\theta') f^n(\theta) d\tilde{\mu}(\theta) - \int t(\theta,\theta') \bar{f}(\theta) d\tilde{\mu}(\theta)\right) \bar{f}(\theta') d\tilde{\mu}(\theta')\right| \\
&= \left|\int \left(\int t(\theta,\theta') \bar{f}(\theta') d\tilde{\mu}(\theta')\right) f^n(\theta) d\tilde{\mu}(\theta) - \int \left(\int t(\theta,\theta') \bar{f}(\theta') d\tilde{\mu}(\theta')\right) \bar{f}(\theta) d\tilde{\mu}(\theta)\right|.
\end{aligned}$$

Thus,

$$\bar{f} \in \arg\min_{f \in \tilde{\Delta}_S} \iint t(\theta,\theta') f(\theta) f(\theta') d\tilde{\mu}(\theta) d\tilde{\mu}(\theta').$$

This proves that the minimizing set of priors in the seller's objective function is nonempty.

Next, we show that there exists a mechanism (x,t) that satisfies the IC and IR constraints and achieves the optimal revenue for the seller. Since transfers are bounded, the seller's revenue is bounded. Suppose that the value of the seller's problem (2.1) is R. This means that there exist a sequence of mechanisms $\{(x^n,t^n)\}$ such that (x^n,t^n) satisfies IC and IR constraints for each n, and if we let,

$$R^n = \min_{\mu \in \Delta_S^m} \iint \left[t^n(\theta,\theta') + t^n(\theta',\theta)\right] d\mu(\theta) d\mu(\theta'),$$

then $R^n \to R$.

Note that $x^n \in \mathcal{B}_\infty^1$ and $t^n \in \mathcal{B}_\infty^K$. Therefore passing to subsequences x^n converges weakly to x and and t^n converges weakly to t. Clearly $x(\theta, \theta') + x(\theta', \theta) \leq 1$ for all $\theta, \theta' \in \Theta$.

Next, we will show that (x, t) satisfies IC and IR constraints. Note that it is sufficient to show that for any $\theta, \tilde{\theta} \in \Theta$,

$$\lim_{n\to\infty} \min_{\mu\in\Delta_B^m} \int \left(x^n(\tilde{\theta}, \theta')\theta - t^n(\tilde{\theta}, \theta') \right) d\mu(\theta') = \min_{\mu\in\Delta_B^m} \int \left(x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta') \right) d\mu(\theta').$$

To simplify notation let

$$g_{\tilde{\theta}\theta}^n(\theta') = x^n(\tilde{\theta}, \theta')\theta - t^n(\tilde{\theta}, \theta')$$

and

$$g_{\tilde{\theta}\theta}(\theta') = x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta')$$

for all $\theta, \tilde{\theta} \in \Theta$. Observe that $g_{\tilde{\theta}\theta}^n$ and $g_{\tilde{\theta}\theta}$ are both bounded by $K + 1$ and thus they are both in L_∞. Moreover since x^n and t^n converge weakly to x and t, $g_{\tilde{\theta}\theta}^n$ converges weakly to $g_{\tilde{\theta}\theta}$.

Now note that for all $\hat{\mu} \in \Delta_B^m$,,

$$\lim_{n\to\infty} \min_{\mu\in\Delta_B^m} \int g_{\tilde{\theta}\theta}^n(\theta') d\mu(\theta') \leq \lim_{n\to\infty} \int g_{\tilde{\theta}\theta}^n(\theta') d\hat{\mu}(\theta') = \int g_{\tilde{\theta}\theta}(\theta') d\hat{\mu}(\theta'),$$

where the equality follows since $g_{\tilde{\theta}\theta}^n$ converges weakly to $g_{\tilde{\theta}\theta}(\theta')$. Thus,

$$\lim_{n\to\infty} \min_{\mu\in\Delta_B^m} \int g_{\tilde{\theta}\theta}^n(\theta') d\mu(\theta') \leq \min_{\mu\in\Delta_B^m} \int g_{\tilde{\theta}\theta}(\theta') d\mu(\theta').$$

On the other hand, for each n let $f^n \in \tilde{\Delta}_B$ be such that,

$$\int g_{\tilde{\theta}\theta}^n(\theta') f^n(\theta') \, d\tilde{\mu}(\theta') = \min_{f\in\tilde{\Delta}_B} \int g_{\tilde{\theta}\theta}^n(\theta') f(\theta') \, d\tilde{\mu}(\theta').$$

(We know such f^n exists since minimizing set of priors is nonempty.) Since $\tilde{\Delta}_B$ is weakly compact again by passing to a subsequence, f^n converges weakly to $\bar{f} \in \tilde{\Delta}_B$.

Note that,

$$\begin{aligned}
&\left|\int g^n_{\tilde{\theta}\theta}(\theta')f^n\left(\theta'\right)d\tilde{\mu}(\theta') - \int g_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta')\right| \\
&\leq \left|\int g^n_{\tilde{\theta}\theta}(\theta')f^n\left(\theta'\right)d\tilde{\mu}(\theta') - \int g^n_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta')\right| \\
&+ \left|\int g^n_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta') - \int g_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta')\right| \\
&\leq |K+1|\left|\int \left(f^n\left(\theta'\right) - \bar{f}\left(\theta'\right)\right)d\tilde{\mu}(\theta')\right| \\
&+ \left|\int g^n_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta') - \int g_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta')\right|.
\end{aligned}$$

The last inequality follows from the fact that $\left|g^n_{\tilde{\theta}\theta}(\theta')\right| \leq K+1$. Since f^n weakly converges to $\bar{f} \in \tilde{\Delta}_B$ and $g^n_{\tilde{\theta}\theta}$ converges weakly to $g_{\tilde{\theta}\theta}$ both terms on the right hand side of the last inequality approach to 0. This implies by taking limits in equation (.1.3) that,

$$\int g_{\tilde{\theta}\theta}(\theta')\bar{f}\left(\theta'\right)d\tilde{\mu}(\theta') = \lim_{n\to\infty}\min_{f\in\tilde{\Delta}_B}\int\left(x^n(\tilde{\theta},\theta')\theta - t^n(\tilde{\theta},\theta')\right)f\left(\theta'\right)d\tilde{\mu}(\theta'),$$

which in turn implies that

$$\min_{f\in\tilde{\Delta}_B}\int g_{\tilde{\theta}\theta}(\theta')f\left(\theta'\right)d\tilde{\mu}(\theta') \leq \lim_{n\to\infty}\min_{f\in\tilde{\Delta}_B}\int\left(x^n(\tilde{\theta},\theta')\theta - t^n(\tilde{\theta},\theta')\right)f\left(\theta'\right)d\tilde{\mu}(\theta').$$

The previous inequality together with (.1.3) implies (.1.3) which concludes the proof.

.1.4 Proof of Corollary II.3

Suppose that for some mechanism (x,t), there exists some positive measure event $\widetilde{\Theta} \subseteq \Theta$ such that for all $\theta \in \widetilde{\Theta}$ and for all $G \in \Delta^{\min}_S$,

$$\int_\Theta q(\theta,\theta')dG(\theta') > \min_{H\in\Delta_B}\int_\Theta q(\theta,\theta')dH(\theta')$$

We need to show that there exists a full insurance that is strictly preferred by the seller. Let (x, t') be defined as in the proof of Proposition II.1. We know that

$$\begin{aligned}
&\inf_{G\in\Delta_S} \int_\Theta \int_\Theta \left[t'(\theta,\theta') + t'(\theta',\theta)\right] dG(\theta) dG(\theta') \\
&\geq \inf_{G\in\Delta_S} \int_\Theta \int_\Theta 2t(\theta,\theta') dG(\theta) dG(\theta') + \inf_{G\in\Delta_S} \int_\Theta \int_\Theta 2\delta(\tilde{\theta},\theta') dG(\theta') dG(\tilde{\theta}).
\end{aligned}$$

Let $\tilde{G} \in \arg\min_{G\in\Delta_S^m} \iint \left[t'(\theta,\theta') + t'(\theta',\theta)\right] dG(\theta) dG(\theta')$. We will show the claim by considering two cases.

The first case is $\tilde{G} \in \Delta_S^{\min}$. In this case for all $\theta \in \widetilde{\Theta}$ equation (.1.4) holds. Therefore,

$$\begin{aligned}
&\int_\Theta \int_\Theta \delta(\theta,\theta') d\tilde{G}(\theta') d\tilde{G}(\theta) \\
&= \int_\Theta \left[\int_\Theta q(\theta,\theta') d\tilde{G}(\theta') - \min_{H\in\Delta_B} \int_\Theta q(\theta,\theta') dH(\theta')\right] d\tilde{G}(\theta) > 0.
\end{aligned}$$

Using equation (**??**) we conclude that the seller strictly prefers the mechanism (x, t').

The second case is $\tilde{G} \notin \Delta_S^{\min}$. In this case by definition of $\Delta_s^{\min}$,

$$\int_\Theta \int_\Theta 2t(\theta,\theta') d\tilde{G}(\theta) d\tilde{G}(\theta') > \min_{H\in\Delta_S} \int_\Theta \int_\Theta 2t(\theta,\theta') dH(\theta) dH(\theta').$$

Again from equation (**??**) we observe that,

$$\begin{aligned}
&\min_{G\in\Delta_S} \int_\Theta \int_\Theta \left[t'(\theta,\theta') + t'(\theta',\theta)\right] dG(\theta) dG(\theta') \\
&= \int_\Theta \int_\Theta 2t(\theta,\theta') d\tilde{G}(\theta) d\tilde{G}(\theta') + \int_\Theta \int_\Theta 2\delta(\tilde{\theta},\theta') d\tilde{G}(\theta') d\tilde{G}(\tilde{\theta}) \\
&> \min_{H\in\Delta_S} \int_\Theta \int_\Theta 2t(\theta,\theta') dH(\theta) dH(\theta').
\end{aligned}$$

and the seller strictly prefers (x, t').

.1.5 Proof of Proposition II.4

Towards a contradiction suppose that (x, t) is optimal but for a positive measure set of $\bar{\theta}$, $q(\bar{\theta}, \theta)$ is not constant. Since F has strictly positive density we have

$\int q(\bar{\theta},\theta')dF(\theta') > \inf_{\theta'\in\Theta} q(\bar{\theta},\theta')$. So

$$\begin{aligned}\int q(\bar{\theta},\theta')dF(\theta') &> (1-\epsilon)\int_{\Theta} q(\bar{\theta},\theta')dF(\theta') + \epsilon \inf_{\theta'\in\Theta} q(\bar{\theta},\theta') \\ &\geq \inf_{H\in\Delta_B}\int_{\Theta} q(\theta,\theta')dH(\theta').\end{aligned}$$

By Proposition II.1 (x,t) can not be optimal.

.1.6 Proof of Lemma II.5

We need to show that there exists $M > 0$ such that

$$\left|u(\theta) - u\left(\tilde{\theta}\right)\right| \leq M\left|\theta - \tilde{\theta}\right|.$$

We know that,

$$\left(\theta - \tilde{\theta}\right) X^{\min}\left(\tilde{\theta}\right) \leq u(\theta) - u\left(\tilde{\theta}\right) \leq \left(\theta - \tilde{\theta}\right) X^{\max}\left(\tilde{\theta}\right).$$

So if $\theta > \tilde{\theta}$, using the fact that u is increasing we can conclude that,

$$u(\theta) - u\left(\tilde{\theta}\right) \leq \left(\theta - \tilde{\theta}\right) X^{\max}\left(\tilde{\theta}\right) \leq \left|\theta - \tilde{\theta}\right|.$$

Similarly if $\theta < \tilde{\theta}$, then

$$-\left(u(\theta) - u\left(\tilde{\theta}\right)\right) \leq -\left(\theta - \tilde{\theta}\right) X^{\min}\left(\tilde{\theta}\right) \leq \left|\theta - \tilde{\theta}\right|.$$

Together these imply that Lipschitz condition holds with $M = 1$.

.1.7 Proof of Proposition II.6

First note that L^{ε} is increasing in θ, if L is increasing in θ. To see this note that,

$$\begin{aligned}\theta - \frac{1-F(\theta)}{f(\theta)} &> \theta' - \frac{1-F(\theta')}{f(\theta')} \\ &\Rightarrow \theta - \theta' > \frac{1-F(\theta)}{f(\theta)} - \frac{1-F(\theta')}{f(\theta')} \\ &\Rightarrow \theta - \theta' > (1-\varepsilon)\left(\frac{1-F(\theta)}{f(\theta)} - \frac{1-F(\theta')}{f(\theta')}\right) \\ &\Rightarrow \theta - (1-\varepsilon)\frac{1-F(\theta)}{f(\theta)} > \theta' - (1-\varepsilon)\frac{1-F(\theta')}{f(\theta')}.\end{aligned}$$

Note that $X^{\min}(\theta) \leq X(\theta)$. Therefore if $X(\theta) = 0$, $X^{\min}(\theta) = 0$ as well. Letting $\frac{X^{\min}(\theta)}{X(\theta)} = 1$ whenever $X(\theta) = 0$, we define $M(\theta) = \theta - \frac{X^{\min}(\theta)}{X(\theta)} \frac{1-F(\theta)}{f(\theta)}$. We can rewrite R as,

$$R = 2 \int_{\Theta} \int_{\Theta} M(\theta)\, x(\theta, \theta')\, f(\theta')\, f(\theta)\, d\theta' d\theta.$$

Now we can show that the optimal allocation rule is given by setting $x(\theta, \theta') = 1$ if $\theta > \theta'$ and $\theta \geq r$, $x(\theta, \theta') = \frac{1}{2}$ if $\theta = \theta'$ and $\theta \geq r$, and $x(\theta, \theta') = 0$ otherwise. First note that, in the ε-contamination case, $X^{\min}(\theta) \geq (1-\varepsilon) X(\theta)$ for all θ such that $X(\theta) < 1$[1].

Under the above allocation rule $X^{\min}(\theta) = (1-\varepsilon) X(\theta)$ for all θ such that $X(\theta) < 1$. Therefore this allocation rule maximizes $M(\theta)$. By construction $x(\theta, \theta') = 1$ if and only if $M(\theta) > M(\theta')$ and $M(\theta) \geq 0$ therefore maximizing (.1.7).

Finally we show that (x, t) is incentive compatible. To this end first we show that if $X^{\min}$ is non-decreasing selecting u as in (2.11) satisfies IC. We check two cases.

If $\theta > \tilde{\theta}$,

$$u(\theta) - u\left(\tilde{\theta}\right) = \int_{\tilde{\theta}}^{\theta} X^{\min}(y)\, dy \geq X^{\min}\left(\tilde{\theta}\right)\left(\theta - \tilde{\theta}\right)$$

and if $\theta < \tilde{\theta}$,

$$u\left(\tilde{\theta}\right) - u(\theta) = \int_{\theta}^{\tilde{\theta}} X^{\min}(y)\, dy \leq X^{\min}\left(\tilde{\theta}\right)\left(\tilde{\theta} - \theta\right).$$

So in either case,

$$\begin{aligned} u(\theta) &\geq u\left(\tilde{\theta}\right) + \inf_{G \in \Delta_B} \int_{\Theta} \left(\theta - \tilde{\theta}\right) x(\tilde{\theta}, \theta') dG(\theta') \\ &= \inf_{G \in \Delta_B} \int_{\Theta} \left(x(\tilde{\theta}, \theta')\theta - t(\tilde{\theta}, \theta')\right) dG(\theta'). \end{aligned}$$

[1]This is true since:

$$\begin{aligned} X^{\min}(\theta) &= \inf_{G \in \Delta_b} \int_{\Theta} x(\theta, \theta') dG(\theta') = (1-\varepsilon) \int_{\Theta} x(\theta, \theta') d\tilde{\mu}(\theta') + \varepsilon \inf_{\hat{\mu} << \mu} \int_{\Theta} x(\theta, \theta') d\hat{\mu}(\theta') \\ &\geq (1-\varepsilon) \int_{\Theta} x(\theta, \theta') dF(\theta'). \end{aligned}$$

which is the IC constraint.

Now, note that for the allocation rule in the statement of Proposition II.6, $X^{\min}$ is non-decreasing, and thus the mechanism (x,t) is incentive compatible.

.1.8 Proof of Proposition II.7

Suppose that given a mechanism (x,t), there exists a positive measure subset $\widetilde{\Theta} \subseteq \Theta$ such that for all $\tilde{\theta} \in \widetilde{\Theta}$, $q(\tilde{\theta},\theta)$ is weakly decreasing in θ and $q(\tilde{\theta},\theta') < q(\tilde{\theta},\theta'')$ for some $\theta',\theta'' \in \Theta$. First note that if $H \in \Delta_B$ first-order stochastically dominates $G \in \Delta_S$ then

$$\int_\Theta q(\tilde{\theta},\theta')dH(\theta') < \int_\Theta q(\tilde{\theta},\theta')dG(\theta').$$

So there exists $H \in \Delta_B$ such that

$$\int_\Theta q(\theta,\theta')dH(\theta') < \min_{G\in\Delta_S} \int_\Theta q(\theta,\theta')dG(\theta')$$

which in turn implies that

$$\min_{H\in\Delta_B} \int_\Theta q(\theta,\theta')dH(\theta') < \min_{G\in\Delta_S} \int_\Theta q(\theta,\theta')dG(\theta').$$

Note that since Δ_S and Δ_B are weakly compact and convex with elements that are countably additive probability measures and the transfers are uniformly bounded, the minimums above exist by Proposition II.2. Finally, by Proposition II.1, (x,t) is not optimal.

.1.9 Proof of Proposition II.8

Let (x,t) be an arbitrary incentive compatible and individually rational mechanism. Define $T(\theta)$ as bidder θ's expected transfer under F, that is,

$$T(\theta) = \int_\Theta t(\theta,\theta')dF(\theta')$$

Now let

$$\tilde{t}(\theta,\theta') = T(\theta) - T(\theta') + \int_{\Theta} T(i)dF(i).$$

First, we show that the mechanism $(x,\tilde{t})$ makes the seller (weakly) better off, leaves the bidders' payoffs unchanged under truthtelling, and is incentive compatible.

To see that the seller is (weakly) better off under $(x,\tilde{t})$,note that the seller's payoff in the mechanism $(x,\tilde{t})$ is:

$$\begin{aligned}
&\inf_{G\in\Delta_S} \int_{\Theta}\int_{\Theta} \left[\tilde{t}(\theta,\theta') + \tilde{t}(\theta',\theta)\right] dG(\theta)dG(\theta') \\
&= \inf_{G\in\Delta_S} \int_{\Theta}\int_{\Theta} [T(\theta) - T(\theta') + \int T(i)dF(i) + T(\theta') - T(\theta) \\
&+ \int_{\Theta} T(j)dF(j)]dG(\theta)dG(\theta') \\
&= \inf_{G\in\Delta_S} \int_{\Theta}\int_{\Theta} \left[2\int T(i)dF(i)\right] dG(\theta)dG(\theta') \\
&= 2\int_{\Theta} T(i)dF(i) = \int_{\Theta}\int_{\Theta} \left[t(\theta,\theta') + t(\theta',\theta)\right] dF(\theta)dF(\theta') \\
&\geq \inf_{G\in\Delta_S} \int_{\Theta}\int_{\Theta} \left[t(\theta,\theta') + t(\theta',\theta)\right] dG(\theta)dG(\theta')
\end{aligned}$$

where the last inequality follows since $F \in \Delta_S$. Hence the seller weakly prefers $(x,\tilde{t})$.

Next we show that $(x,\tilde{t})$ leaves the bidders' payoffs unchanged under truth-telling. By construction:

$$\begin{aligned}
\int_{\Theta} \tilde{t}(\theta,\theta')dF(\theta') &= \int_{\Theta} \left[T(\theta) - T(\theta') + \int_{\Theta} T(i)dF(i)\right] dF(\theta') \\
&= T(\theta) - \int_{\Theta} T(\theta')dF(\theta') + \int_{\Theta} T(i)dF(i) = T(\theta) = \int_{\Theta} t(\theta,\theta')dF(\theta').
\end{aligned}$$

Finally we show that $(x,\tilde{t})$ is incentive compatible. Note that,

$$\int \left[\theta x(\tilde{\theta}, \theta') - t(\tilde{\theta}, \theta')\right] dF(\theta') = \int \theta x(\tilde{\theta}, \theta') dF(\theta') - \int t(\tilde{\theta}, \theta') dF(\theta')$$
$$= \int \theta x(\tilde{\theta}, \theta') dF(\theta') - T(\tilde{\theta}) = \int \theta x(\tilde{\theta}, \theta') dF(\theta') - \int \tilde{t}(\tilde{\theta}, \theta') dF(\theta').$$

So the payoff for type θ to pretend to be $\tilde{\theta}$ is the same in both mechanisms (x, t) and $(x, \tilde{t})$ and since (x, t) is incentive compatible, $(x, \tilde{t})$ must be as well. Since, by construction, $\tilde{t}(\theta, \theta') + \tilde{t}(\theta', \theta)$ is constant for all $\theta, \theta' \in \Theta$, the first part of the proof is completed. Next suppose

$$\inf_{G \in \Delta_S} \int_\Theta \int_\Theta \left[t(\theta, \theta') + t(\theta', \theta)\right] dG(\theta) dG(\theta') < \int_\Theta \int_\Theta \left[t(\theta, \theta') + t(\theta', \theta)\right] dF(\theta) dF(\theta').$$

Then the weak inequality becomes strict, and the seller becomes strictly better off.

.2 Chapter 3 (Appendix)

.2.1 Notes Concerning the Two-Price Auction

.2.1.1 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:

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

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

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

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

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

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

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

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

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

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

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

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

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] I. G. A. Billot, A. Chateauneuf and J. Sharing beliefs: Between agreeing and disagreeing. *Econometrica*, 68:685–694, 2000.

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

[3] N. Anderson and A. Pape. An insurance model of property tax limitations. Working Paper, 2006.

[4] D. Bergemann and S. Morris. Robust mechanism design. *Econometrica*, 73:1771–1813, 2005.

[5] P. Billingsley. *Convex Probability and Measure.* Wiley Series in Probability and Mathematical Statistics, New York, NY, 1995.

[6] M. Druzel and H. Simon. Causality in bayesian belief networks. *Proceedings of Ninth Conference on Uncertainty in Artificial Intelligence*, 1993.

[7] N. Dunford and J. T. Schwartz. *Linear Operators, Part I.* Wiley Interscience, New York, 1957.

[8] D. Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.

[9] J. Ely and K. Chung. Foundations of dominant strategy mechanisms. Northwestern University, mimeo, 2005.

[10] P. Eso and G. Futo. Auction design with a risk averse seller. *Economics Letters*, 61:71–74, 1999.

[11] I. Esponda. Behavioral equilibrium in economies with adverse selection. *unpublished*, 2005.

[12] E. Eyster and M. Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, 09 2005.

[13] D. Fudenberg and D. K. Levine. Self-confirming equilibrium. *Econometrica*, 61(3):523–45, May 1993.

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

[15] J. K. Goeree and T. O§erman. The amsterdam auction. *Econometrica*, 72:281–294, 2004.

[16] I. Hacking. Slightly more realistic personal probability. *Philosophy of Science*, 34(4):311–325, December 1967.

[17] M. Harris and A. Raviv. Allocation mechanisms and the design of auctions. *Econometrica*, 49:1477–1500, 1981.

[18] J. Heckman. The scientific view of causality. University of Chicago, University of College London, and the American Bar Association, April 2005.

[19] A. Heifetz and Z. Neeman. On the generic (im)possibility of full surplus extraction in mechanism design. *Econometrica*, 74:213–233, 2006.

[20] R. Jeffrey. *The Logic of Decision*. University of Chicago Press, 1964.

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

[22] E. Karni. On the equivalence between descending bid auctions and örst price sealed bid auctions. *Theory and Decision*, 25:211–217, 1988.

[23] E. Karni. Subjective expected utility theory without states of the world. JHU WP523, August 2005.

[24] E. Karni and Z. Safra. Vickrey auctions in the theory of expected utility with rank-dependent probabilities. *Economics Letters*, 20:15–18, 1986.

[25] E. Karni and Z. Safra. Ascending bid auctions with behaviorally consistent bidders. *Annals of Operations Research*, 19:435–446, 1989.

[26] E. Karni and Z. Safra. Dynamic consistency, revelations in auctions and the structure of preferences. *Review of Economic Studies*, 56:421–433, 1989.

[27] R. Langlois. *International Encyclopedia of Business & Management, 2nd edition.*, chapter Entry on "Rationality in Economics". London: Thompson International Publishers, 2001.

[28] K. Lo. Sealed bid auctions with uncertainty bidders. *Economics Theory*, 12:1–20, 1998.

[29] A. C. M. Marinacci, F. Maccheroni and J. Tallon. Monotone continuous multiple priors. *Economic Theory*, forthcoming.

[30] E. Maskin and J. Riley. Optimal auctions with risk averse buyers. *Econometrica*, 52:1473–1518, 1984.

[31] K. Matsuyama. Growing through cycles. *Econometrica*, 67(2):335–348, March 1999.

[32] S. Matthews. Selling to risk averse buyers with unobservable tastes. *Journal of Economic Theory*, 30:370–400, 1983.

[33] S. Mukerji. Ambiguity aversion and incompleteness of contractual form. *American Economic Review*, 88:1207–1231, 1998.

[34] J. Muth. Rational expectations and the theory of price movements. *Econometrica*, 29(3):315–335, July 1961.

[35] R. Myerson. Incentive-compatibility and the bargaining problem. *Econometrica*, 47:61–73, 1979.

[36] R. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6:58–73, 1981.

[37] E. Ozdenoren. Auctions and bargaining with a set of priors. University of Michigan, mimeo, 2001.

[38] S. Page. *The Difference, How The Power of Diversity Creates Better Groups, Teams, Schools, and Societies*. Princeton University Press, 2007.

[39] A. D. Pape. Causal coherence. University of Michigan, 2006.

[40] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2000.

[41] K. Popper. *The Open Society and Its Enemies*, volume II. Princeton: Princeton University Press., 2nd edition, 1966.

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

[43] J. Riley and W. Samuelson. Optimal auctions. *American Economic Review*, 71:381–392, 1981.

[44] L. A. Rivera-Batiz and P. M. Romer. Economic integration and endogenous growth. *Quarterly Journal of Economics*, 106(2):531–555, 1991.

[45] A. Salo and M. Weber. Ambiguity aversion in first-price sealed-bid auctions. *Journal of Risk and Uncertainty*, 11:123–137, 1995.

[46] L. J. Savage. *The Foundations of Statistics*. Wiley, 1954.

[47] C. Simon and L. Blume. *Mathematics For Economists*. W.W. Norton, 1994.

[48] S. Sloman and D. Lagnado. *The Psychology of Learning and Motivation*, volume 44, chapter Causal Invariance in Reasoning and Learning. San Diego: Academic Press, 2004.

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

[50] T. Verma and J. Pearl. Equivalence and synthesis of causal models. *Proceedings of the 6th Conference on the Uncertainty in Artificial Intelligence*, pages 220–227, July 1990.

[51] W. Vickrey. Counterspeculation, auctions and competitive sealed tender. *Journal of Finance*, 16:8–37, 1961.

[52] O. Volij. Payoff equivalence in sealed bid auctions and the dual theory of choice under risk. *Economics Letters*, 76:231–237, 2002.