

Can Agents with Causal Misperceptions be Systematically Fooled?*

Ran Spiegler[†]

February 25, 2017

Abstract

The conventional rational-expectations postulate rules out the possibility that agents will form systematically biased estimates of economic variables. This paper revisits this question under the assumption that agents' expectations are based on a misspecified causal model. I present a model in which an agent forms estimates (or forecasts) of individual variables after observing a signal. His estimates are based on fitting a subjective causal model - formalized as a directed acyclic graph, following the "Bayesian networks" literature - to objective long-run data. I show that the agent's estimates are never systematically biased if and only if his graph is perfect - it links every two causes of some third variable. This result identifies a certain form of *correlation neglect* as the kind of misperception that generates systematic prediction errors. I demonstrate the relevance of this result for economic applications: speculative trade, manipulation of a firm's reputation and a stylized "monetary policy" example in which the inflation-output relation obeys an expectational Phillips Curve.

*Financial support by the Sapir Center and ERC Advanced Investigator grant no. 692995 is gratefully acknowledged. I thank Yair Antler, Kfir Eliaz, Raffaella Giacomini, Todd Sarver, Vasiliki Skreta, Dimitri Vayanos, Michael Woodford and especially Heidi Thysen, for helpful comments and discussions.

[†]Tel Aviv University, University College London and CFM. URL: <http://www.tau.ac.il/~rani>. E-mail: rani@post.tau.ac.il.

1 Introduction

The outcome of many real-life interactions depends on agents' ability to predict particular variables. In conventional models, an agent's predictions are constrained by the "rational expectations" postulate - i.e., the agent fully understands the statistical regularities in his environment and thus forms "optimal" forecasts of any variable conditional on his information. His predictions may miss the target, but the prediction errors will cancel out on average. In other words, the agent cannot be "systematically fooled". Indeed, economists sometimes identify the latter property with the rational-expectations principle itself:

"The concept of rational expectations asserts that outcomes do not differ systematically (i.e., regularly or predictably) from what people expected them to be. The concept is motivated by the same thinking that led Abraham Lincoln to assert, "You can fool some of the people all of the time, and all of the people some of the time, but you cannot fool all of the people all of the time." From the viewpoint of the rational expectations doctrine, Lincoln's statement gets things right. It does not deny that people often make forecasting errors, but it does suggest that errors will not persistently occur on one side or the other." (Sargent (2003))

However, rational expectations involve more than the requirement that the agent's predicted outcome is unbiased on average - they demand a correct perception of the *entire* joint distribution over all relevant variables. A priori, an agent's beliefs may satisfy the former while violating the latter. This paper is an attempt to get a better understanding of this distinction.

Of course, one can violate rational expectations in many ways; this paper focuses on the role of *causal misperceptions* in the formation of beliefs. As in Spiegler (2016a), I assume that the agent holds a subjective causal model that links some of the relevant variables. Following the Statistics and Artificial Intelligence literature on "Bayesian networks" (Cowell et al. (1999), Pearl

(2009), Koller and Friedman (2009)), a causal model is represented by a directed acyclic graph (DAG): nodes represent variables, and a link $x \rightarrow y$ signifies a perceived direct causal effect of x on y (without any pre-conception regarding the sign or magnitude of this effect). The agent fits his causal model to long-run data, thus quantifying the perceived causal relations. The agent then employs this quantified model to estimate the expected value of individual variables in his model, conditional on the observed realization of one of them.

This process has at least two interpretations. First, a DAG can represent a lay person's intuitive causal perceptions, or a *narrative* that he employs to focus his understanding of empirical regularities (for a summary of psychological research on the role of intuitive causal models in reasoning about uncertainty, see Sloman (2005)). The DAG can also represent a professional forecaster's explicit *formal* model, which consists of a recursive system of non-parametric structural equations. The forecaster's commitment to a particular model can result from theoretical preconceptions, or from its ability to "tell a story".¹

What is common to both interpretations is the idea that reasoning about joint probability distributions is cognitively demanding. One cannot perceive them directly as a whole; measuring correlations among any given set of variables (and communicating the measurements' outcomes to others) carries an implicit cost. Moreover, the number of potentially relevant correlations that one might need to digest grows exponentially with the number of relevant variables. Thinking in terms of a causal model - whether intuitive or formal - simplifies this task, especially if it is represented by a *sparse* DAG. The model alerts the agent to specific correlations. His overall perception of empirical regularities in his environment - which forms the basis of his individual estimates - is the result of putting these estimated correlations together.

¹Consider the following quote from an economic forecasting company (<http://www.macroadvisers.com/why-model-based-forecasting>): "A model-based forecast tells a story. The model allows us to identify the key forces that are driving the economy...We quickly found that most of our clients didn't want to sort through computer output for the hundreds of variables in our model over the next twelve quarters (or more). They wanted to understand why; they wanted stories...". For a critical discussion of theory-based forecasting, see Giacomini (2015).

The question that I analyze in this paper is whether an agent with a misspecified causal model will nevertheless produce conditional estimates that are correct on average. The following example illustrates the formalization of this question and its economic relevance.

Example 1.1: Exploiting a belief in monetary neutrality

Monetary theory offers what is arguably the most well-known economic example of the “systematic fooling” problem. In a textbook model that goes back to Kydland and Prescott (1977) and Barro and Gordon (1983), a central bank controls a policy variable that affects inflation. The private sector forms an inflation forecast, possibly after observing some signal regarding the central bank’s decision. Private-sector expectations are relevant because real output is determined by an “expectations-augmented” Phillips Curve, such that the real effect of inflation is at least partly offset when inflation is anticipated. Thus, if the central bank wants to raise expected output, it would like to be able to set inflation systematically above private-sector expectations.

This paper will make use of a running example that is based on a simple reformulation of the Barro-Gordon model studied by Sargent (2001), Athey et al. (2005) and others. Needless to say, this paper is purely theoretical and the example is too stylized to count as a “proper” macro application. Nevertheless, I find it appropriate, given the important historical role that it played in the development of the rational-expectations modeling approach - and the crucial role of the “systematic fooling” problem within that historical development.

In this model, the central bank chooses an action a after privately observing a collection of indicators summarized by the real-valued variable θ . Inflation π is a stochastic function of a alone. The private sector forms its inflation forecast e after observing a . Real output y is given by a “New Classical” Phillips Curve $y = \pi - e + \eta$, where η is independent Gaussian noise. Thus, only unanticipated inflation has real effects. The central bank’s utility function is $y - \theta\pi$ - i.e., it wants higher output and lower inflation, where θ measures the central bank’s trade-off between the two motives. If the private sector had rational expectations, e would be equal to the true expected

value of π conditional on a , such that ex-ante expected output would be zero, independently of the central bank's strategy. In this case, the central bank would choose an action that minimizes expected inflation, regardless of θ .

The private sector's causal model is represented by the following DAG, denoted R'' :

$$\begin{array}{ccc}
 \theta & \rightarrow & a \\
 \downarrow & & \downarrow \\
 y & \rightarrow & \pi
 \end{array} \tag{1}$$

According to this causal model, inflation is a consequence of two causes: output and the central bank's action (the model omits the private sector's expectations). These two causes are in turn consequences of the exogenous variable θ . This causal model is misspecified because it perceives output to be independent of monetary policy conditionally on θ - whereas according to the true process, output is a consequence of the central bank's action via the Phillips Curve, independently of θ . Thus, the private sector's causal model tells a "classical" story that postulates the absolute *neutrality of monetary policy*, allowing a and y to be correlated only via their dependence on θ - whereas the true model allows for non-neutrality via inflationary surprises.

How does the private sector employ its causal model to forecast inflation? It simply *fits* the model to the true steady-state joint distribution p over θ, a, π, y . If p were consistent with R , $p(\theta, a, \pi, y)$ could be written as

$$p_{R''}(\theta, a, \pi, y) = p(\theta)p(a | \theta)p(y | \theta)p(\pi | a, y) \tag{2}$$

The formula $p_{R''}(\theta, a, \pi, y)$ describes the private sector's subjective belief as a function of the true steady-state distribution p . It is an example of a "*Bayesian-network factorization formula*", which factorizes the steady-state distribution p over θ, a, π, y into a product of conditional-probability terms, *as if* p were indeed consistent with R'' . The four terms on the R.H.S of (2) can be viewed as outcomes of specific correlation measurements that a forecaster makes, for these are the measurements that are needed for quantifying his model. Because the forecaster perceives statistical regularities through the prism of an incorrect model, the subjective belief $p_{R''}$ may systematically

distort the correlation structure of the actual steady-state distribution p .

The private sector's inflation forecast after observing the central bank's action a is²

$$E_{R''}(\pi | a) = \sum_{\pi} p_{R''}(\pi | a)\pi = \sum_{\pi} \left(\sum_y \sum_{\theta} p(\theta | a)p(y | \theta) \right) p(\pi | a, y)\pi \quad (3)$$

This is in general different from the “rational” inflation forecast

$$E(\pi | a) = \sum_{\pi} p(\pi | a)\pi = \sum_{\pi} \left(\sum_y p(y | a) \right) p(\pi | a, y)\pi$$

The discrepancy arises because $p_{R''}(\pi | a)$ involves an implicit expectation over y *without* full conditioning on a - it only acknowledges the correlation between a and y through their correlation with θ . E.g., if the central bank mixed over actions independently of θ , we would have $p(y | a) \equiv p(y)$. (Note that because the steady-state distribution p is affected by private-sector expectations, it is an “equilibrium” distribution; the equilibrium requirement is that $e = E_{R''}(\pi | a)$ with probability one, for every a .)

How does the private sector's “non-rational” inflation forecast affect the central bank's considerations? Because the term $p(y | \theta)$ in (3) is not independent of a , a change in the central bank's strategy can lead to a change in $E_{R''}(\pi | a)$. This dependency is what makes the central bank's problem non-trivial. In Section 4.2, I present a simple specification of $p(\pi | a)$ for which the central bank has a strategy that leads the private sector to systematically underestimate inflation - i.e.,

$$\sum_a p(a)E_{R''}(\pi | a) < \sum_{\pi} p(\pi)\pi$$

Consequently, the central bank can use monetary policy to enhance expected output. Our task will be to find restrictions on R'' or the domain of p that would make such “systematic fooling” impossible.

²I abuse notation and use simple summations rather than integration, for expositional clarity.

Overview of the model and the main results

In Section 2, I present a general model in which an agent forms estimates of economic variables after observing the realization of one variable. The agent’s subjective causal model is represented by a DAG R over a set of nodes that correspond to some subset of the economic variables. He fits this model to an objective joint probability distribution p defined over all variables (including the agent’s own estimates) that satisfies the above “equilibrium” condition, and this produces a subjective distribution p_R over the variables that his model admits.

Can such an agent be systematically fooled, in the sense that his average estimate of some individual variable will be biased? The main result, given in Section 3, provides a simple answer: The agent’s estimates are correct on average for any possible p , if and only if his DAG is *perfect*. A DAG is perfect if any pair of direct causes of any third variable are directly linked themselves. The private sector’s DAG in Example 1.1 violates perfection, because it perceives a and y as direct causes of π , and yet it does not postulate a direct causal link between them. As a result, we can find *some* objective distribution for which the agent’s forecast of *some* variable (in this case, inflation) is biased on average. In contrast, the DAG $\theta \rightarrow a \rightarrow \pi \rightarrow y$ is perfect, and therefore cannot give rise to systematically biased forecasts.

Perfection is a familiar property in the Bayesian-networks literature. Its significance in the present context is that it highlights the role of (a certain form of) *correlation neglect* in generating systematically biased estimates. In an imperfect DAG, two perceived causes of a third variable are not considered to be direct causes of one another. This may lead to an underestimation of the extent to which they are correlated in reality. For instance, the DAG R'' in Example 1.1 accounts for the correlation between a and y only to the extent to which these two variables are correlated with θ . Depending on the central bank’s strategy, this can result in underestimation of their actual correlation, which in turn leads to systematically biased inflation forecasts. Any type of causal misperception that does not fall into this class of correlation-neglect errors does not generate systematically biased estimates.

Perfect DAGs are significant for another reason. In perfect DAGs - and only in such DAGs - the direction of any given causal link is unidentified from observational data (i.e., there is an observationally equivalent DAG that reverses that link). Thus, the agent’s misspecified causal model renders him vulnerable to biased estimates if and only if it postulates empirically meaningful direction of causation.

In Section 4 I apply the model to environments in which the possibility of systematically biased estimates is economically relevant. First, I present a simple example of a firm that considers the use of sponsored reviews to enhance its reputation among consumers. Second, I provide a thorough analysis of the “monetary policy” example.

Section 5 studies two extensions of the model. First, I explore the role of restrictions on the domain of permissible objective distributions. Specifically, I show that when p is a *multivariate normal distribution*, the agent’s estimates are unbiased on average, *regardless* of his DAG. Second, I examine what happens when the agent observes *multiple* variables before forming his estimates. I use this characterization to obtain a “no-trade theorem” in a simple model of speculative trade in which traders form beliefs according to (possibly heterogeneous) perfect DAGs.

2 The Model

Let x_0, x_1, \dots, x_n be a collection of real-valued economic variables. In this section and the next, I assume that every economic variable can take finitely many values (the extension to continuous variables is straightforward). For every $M \subseteq \{0, 1, \dots, n\}$, denote $x_M = (x_i)_{i \in M}$. An agent observes the realization of one variable, which I will take to be x_0 . He then forms a subjective estimate e_i of each of the economic variables x_i , $i = 1, \dots, n$. In some applications, I refer to e_i as the agent’s forecast of x_i . I use p to denote an objective joint distribution over all $2n + 1$ variables. This distribution represents long-run (or steady-state) statistical regularities in the agent’s environment. If the agent’s estimates are based on rational expectations, then p must satisfy the restriction that for every $i = 1, \dots, n$, $p(e_i | x_0)$ assigns probability one to

$E(x_i | x_0) = \sum_{x_i} p(x_i | x_0)x_i$. (Throughout the paper, E without a subscript means expectation w.r.t the objective distribution p .)

Our agent is characterized by a *directed acyclic graph* (DAG) (N, R) , where $N \subseteq \{0, \dots, n\}$ is the set of nodes and R is the set of directed links, such that there is no directed path from a node to itself. I use jRi or $j \rightarrow i$ interchangeably to denote a directed link from j into i . Abusing notation, let $R(i) = \{j \in N \mid jRi\}$ be the set of “parents” of node i . Following Pearl (2009), I interpret R as a *causal model* - i.e., the link $j \rightarrow i$ means that x_j is perceived as an immediate cause of x_i . The model embodies no preconception regarding the causal effect’s sign or magnitude. In particular, it allows the effect to be null. I assume throughout that $0 \in N$ - i.e., the agent’s model acknowledges the variable he gets to observe. In contrast, it does *not* acknowledge the estimate variables e_1, \dots, e_n - they are not represented by nodes in the DAG. I will provide a formal justification for the latter restriction in Section 5.3. From now on, I suppress N and refer to R itself as the agent’s DAG.

The agent perceives the steady-state statistical regularities through the prism of his subjective causal model. Specifically, for any objective distribution p , the agent’s subjective belief over x_N is

$$p_R(x_N) = \prod_{i \in N} p(x_i \mid x_{R(i)}) \quad (4)$$

Thus, R encodes a mapping that transforms every objective distribution p into a subjective belief p_R . A probability distribution p is *consistent* with R if $p_R(x_N) \equiv p(x_N)$. When p is inconsistent with R , the agent’s belief distorts the true correlation structure of p . When the DAG is fully connected, (4) is reduced to a textbook chain rule, such that every p is consistent with R - i.e., the agent has “rational expectations”. For every $M \subset N$, $p_R(x_M) = \sum_{x_{N-M}} p_R(x_M, x_{N-M})$, as usual. The agent’s subjective distribution over any variable x_i , $i = 1, \dots, n$, conditional on his observation of x_0 is $p_R(x_i \mid x_0) = p_R(x_0, x_i)/p_R(x_0)$. (In the general analysis, I will assume that p has full support over X_N , such that all the the conditional probabilities in (4) are well-defined. In applications, I will sometimes be able to relax the full-

support assumption.)

I impose the following restriction on the objective distribution p . It plays no technical role in the general analysis (except in Section 5.3); I introduce it mainly to maintain the equilibrium interpretation of p .

Condition 1 *The domain of permissible objective distributions is restricted as follows. For every x_0 and $i = 1, \dots, n$, $p(e_i | x_0)$ assigns probability one to*

$$E_R(x_i | x_0) = \sum_{x_i} p_R(x_i | x_0)x_i \quad (5)$$

The conditional expected value $E_R(x_i | x_0)$ is the agent’s estimate of x_i after observing x_0 . I have in mind two primary interpretations of this account of how the agent forms his estimates. First, following the work of psychologists on causal reasoning (e.g. Sloman (2005)), the DAG R may capture *intuitive causal perceptions* of an agent in his everyday decision making. These prior perceptions determine the correlations that the agent pays attention to. He learns these correlations, interprets them causally and puts them together in accordance with his causal model. The output of this activity is a subjective belief, given by (4). Then, when he receives the signal x_0 , he relies on his subjective belief to form a conditional estimate of specific variables.

A different interpretation is that the agent is a *professional forecaster*, who has a prior model of the economic environment; he fits the model to the long-run distribution, and uses this “estimated model” to form forecasts of specific variables upon request. In this case, the DAG R represents an explicit, formal model that consists of a system of structural equations having two crucial characteristics. First, the system is *recursive*: a dependent variable in any given equation cannot appear as an explanatory variable in some earlier equation. This feature is implied by the graph’s acyclicity. It may be introduced as a simplifying device (recursive systems are easier to estimate), or because the agent has an explicitly causal theory. Second, each individual equation is *non-parametric* - i.e., it does not commit to any specific functional form. This feature captures situations in which the forecaster keeps tweaking the equations’ precise functional form for the sake of empirical fit, without

changing their R.H.S variables. This means that estimating an equation in which x_i is a dependent variable produces the true conditional distribution $p(x_i | x_{R(i)})$. It is as if the forecaster tweaks with functional forms until he gets perfect empirical fit, but he does not tamper with the basic causal structure of his model - possibly due to fundamental theoretical pre-conceptions. This is probably not the way successful forecasting *should* be done, but I believe it approximates the way it is sometimes practiced.³

Both interpretations are based on the idea that perceiving joint probability distributions is difficult, especially when many variables are involved. Relying on a model, whether intuitive or formal, facilitates this complex task. The model focuses the agent's attention on a collection of key correlations; these are the only correlations that he ends up processing. The agent extrapolates a belief from these measurements, and he relies on this belief whenever he faces the need to form a conditional or unconditional prediction. This argument suggests that in applications, we should think of the agent's DAG as sparse - i.e., it omits many nodes and edges from a fully connected graph over the set of all potentially relevant variables.

Although I have fixed the variable that the agent gets to observe to be x_0 , this is partly an expositional device; the general analysis does not depend on this assumption. The agent may be using the same estimated model for a variety of purposes and situations, and in each situation he may observe a different variable. As a result, the number of conditional predictions that the agent may be called upon to make can be extremely large. Relying on the estimated model simplifies this task.

The following is the central definition in this paper.

Definition 1 *A DAG R induces unbiased estimates if*

$$\sum_{x_0} p(x_0) E_R(x_i | x_0) \equiv E(x_i)$$

³There is a third interpretation, according to which R does not describe an explicit subjective model, but rather represents the agent's *objective* data limitations, such that p_R is the agent's extrapolated belief from his limited data. This interpretation is elaborated in Spiegler (2015b), and I discuss it briefly in Section 6, but I do not pursue it elsewhere in this paper.

for every $i \in N$ and every objective distribution p that has full support on X_N and satisfies Condition 1.

This definition allows the agent to form estimates that depart from the rational-expectations benchmark - i.e., $E_R(x_i | x_0) \neq E(x_i | x_0)$ for some x_i, x_0 . However, it means that the errors are not systematic: they even out on average.

The simplest example of a misspecified DAG that induces unbiased estimates is the empty DAG (i.e., $R(i) = \emptyset$ for every $i \in N$). This DAG fails to capture any correlation that might actually exist among variables, because $p_R(x_i, x_j) = p(x_i)p(x_j)$ for any pair of nodes $i, j \in N$. However, it is easy to see from (4) that $p_R(x_i | x_0) = p(x_i)$, and therefore $\sum_{x_0} p(x_0)E_R(x_i | x_0) = E(x_i)$. This shows that even when a DAG is misspecified, it can induce unbiased estimates. My goal in the next section will be to characterize the class of DAGs for which this is the case.

3 The Basic Result

I begin the analysis with a few basic concepts and results from the Bayesian-networks literature. Let \tilde{R} be the *skeleton* (i.e., undirected version) of R - i.e., $i\tilde{R}j$ if and only if iRj or jRi . A subset $M \subseteq N$ is a *clique* R if $i\tilde{R}j$ for every $i, j \in M$. A clique M is *ancestral* if $R(i) \subset M$ for every $i \in M$. In particular, a node i is ancestral if $R(i)$ is empty. A node j is an *ancestor* of another node i if R contains a directed path from j into i . Two sets of nodes $A, B \subset N$ are *mutually disconnected* if for every pair of nodes $i \in A$ and $j \in B$, there is no path in \tilde{R} that connects i and j .

Equivalent DAGs

A DAG encodes a mapping from objective distributions to subjective beliefs, which is given by (4). Two DAGs can be equivalent in the sense that they encode the same mapping.

Definition 2 *Two DAGs R and Q over N are **equivalent** if $p_R(x_N) \equiv p_Q(x_N)$ for every $p \in \Delta(X)$.*

For instance, the DAGs $1 \rightarrow 2$ and $2 \rightarrow 1$ are equivalent, by the basic identity $p(x_1)p(x_2 | x_1) \equiv p(x_2)p(x_1 | x_2)$. A DAG that involves intuitive causal relations can be equivalent to a DAG that makes little sense as a causal model (e.g., it postulates that a player's action causes his information).

A *v-collider* in R is an ordered triple of nodes (i, j, k) such that iRk , jRk , $i \not R j$ and $j \not R i$ (i.e., R contains links from i and j into k , yet there is no link between i and j). We say in this case that there is a *v-collider into k* .

Proposition 1 (Verma and Pearl (1991)) *Two DAGs R and Q are equivalent if and only if they have the same skeleton and the same set of v-colliders.*

To illustrate this result, all fully connected DAGs have the same skeleton (every pair of nodes is linked) and an empty set of *v-colliders*, hence they are all equivalent. In contrast, the DAGs $1 \rightarrow 2 \rightarrow 3$ and $1 \rightarrow 2 \leftarrow 3$ are not equivalent: although their skeletons are identical, the former DAG has no *v-colliders* whereas $(1, 3, 2)$ is a *v-collider* in the latter.

Perfect DAGs

The following class of DAGs will play an important role in this paper.

Definition 3 *A DAG is **perfect** if it contains no v-colliders.*

A perfect DAG has the property that if x_i and x_j are perceived as direct causes of x_k , then there must be a perceived direct causal link between them. If we think of a DAG as a recursive system of structural equations, perfection means that if x_i and x_j appear as explanatory variables in the equation for x_k , then there must be an equation in which one of these two variables is explanatory and the other is dependent.

The following is an immediate implication of Proposition 1.

Corollary 1 *Two perfect DAGs are equivalent if and only if they have the same skeleton. In particular, if $M \subseteq N$ is a clique in a perfect DAG R , then M is an ancestral clique in some DAG in the equivalence class of R .*

This corollary means that the causal links postulated by a perfect DAG are unidentified: if iRj , there exists a DAG R' that is equivalent to R , such that $jR'i$. A DAG contains observationally meaningful causal links only when they are part of a v -collider.

The following lemma establishes that if C is an ancestral clique in some DAG in the equivalence class of R , then subjective marginal distribution over x_C is always correct. Otherwise, we can find an objective distribution for which it will be distorted.

Lemma 1 (Spiegler (2016b)) *Let R be a DAG and let $C \subseteq N$. Then, $p_R(x_C) \equiv p(x_C)$ for every p with full support on X_N if and only if C is an ancestral clique in some DAG in the equivalence class of R .*

Thanks to Corollary 1, the lemma implies that in a perfect DAG, $p_R(x_C)$ is always correct for *any* clique C .

We are now ready to state the paper's main result.

Proposition 2 *A DAG R induces unbiased forecasts if and only if it is perfect.*

Proof. **(If).** Assume that R is perfect. By Corollary 1, we can take 0 or i to be ancestral w.l.o.g. By Lemma 1, $p_R(x_0) \equiv p(x_0)$ and $p_R(x_i) \equiv p(x_i)$. Therefore, we can write

$$\sum_{x_0} p(x_0)p_R(x_i | x_0) \equiv \sum_{x_0} p_R(x_0)p_R(x_i | x_0) \equiv p_R(x_i) \equiv p(x_i)$$

which implies the claim.

(Only if). See Appendix. ■

Thus, as long as the agent's DAG is perfect, he cannot be systematically fooled. For instance, suppose that $R : 1 \rightarrow 2 \rightarrow 0$. This DAG is perfect, hence Proposition 2 implies that the agent's estimates of x_1 or x_2 are unbiased. The key for this result is the property that in a perfect DAG, every

node can be regarded as ancestral, which ensures that the perceived marginal distribution of the variable it represents is undistorted.

In contrast, when the agent’s DAG is imperfect (e.g., $R : 1 \rightarrow 0 \leftarrow 2$), we can find an objective distribution for which the agent’s estimate of some variable will be biased. The basic idea behind this result is that an imperfect DAG necessarily contains two variables, x_i and x_j , which are not perceived as direct causes of one another and yet both are perceived as direct causes of a third variable x_k . The failure to acknowledge the direct causal link between x_i and x_j may lead to an underestimation of the extent to which they are correlated in reality. If the actual realization of x_k depends on the joint realizations of x_i and x_j , the partial neglect of the correlation between x_i and x_j can lead to a biased average prediction of at least one of the three variables involved (whether or not x_0 belongs to this triple). Thus, Proposition 2 identifies *correlation neglect* - specifically, neglect of direct causation between two perceived direct causes of a third variable - as the source of systematically biased estimates.

As mentioned above, perfect DAGs have the property that the causal links they postulate are unidentified from observational long-run data. Proposition 2 thus implies that the agent’s causal misperceptions expose him to systematic fooling if and only if they are meaningful for observational data.

Selective estimates

The definition of unbiased estimates is demanding, because it requires the estimates of *all* variables to be unbiased. However, not all estimates or forecasts need to be economically relevant. For example, Example 1.1 focused on the private sector’s inflation forecast, because it had implications for the realization of output under the true process. For other purposes, the private sector’s *output* forecast could be relevant. One would like to know whether the estimate of a particular variable is unbiased, even when the DAG is imperfect.

The following result is a sufficient condition for the agent’s estimate of a *given* x_i to be unbiased. Fix a DAG (N, R) . Define the following induced binary relation: for every $j, k \in N$, jPk if $j = k$ or there exists a directed path in R from j to k . The meaning of jPk is that x_j is (weakly) “above”

x_k in the causal hierarchy of R .

Proposition 3 *Let $i \in N - \{0\}$. Suppose that the subgraph induced by R over $\{j \in N \mid jP0 \text{ or } jPi\}$ is perfect. Then,*

$$\sum_{x_0} p(x_0) E_R(x_i \mid x_0) = E(x_0)$$

for every objective distribution p .

Thus, as long as there are no violations of perfection “above” 0 or i in the causal hierarchy of R , the agent’s conditional estimate of x_i is unbiased. For instance, the private sector’s *output* forecast in Example 1.1 is unbiased.

4 Two Applications

In this section I examine two applications in which possibility of systematically biased estimates is of crucial economic importance.

4.1 Manipulating Reputation with Sponsored Reviews

A firm offers a product of exogenous quality θ , which is the firm’s private information. The agent is a consumer who receives a signal t , interpreted as a *review* of the firm’s product. Based on the signal, the consumer forms an estimate e of the product’s quality. Let $s \in \{0, 1\}$ indicate whether the review is *sponsored* by the firm ($s = 1$ means that it is). The firm’s strategy specifies the probability of sponsoring the review as a function of θ . The realized review is some probabilistic function of θ and s . This function, the exogenous distribution over θ and the firm’s strategy constitute the objective joint distribution p over θ, s, t . Let R be the consumer’s DAG defined over the three variables. As usual in this paper, for every t , $e = E_R(\theta \mid t)$ with probability one.

The firm’s payoff is $e - cs$, where $c \in (0, \frac{1}{2})$ is the cost of sponsoring a review. That is, the firm trades off its reputation and the cost of sponsoring

reviews. The firm’s ex-ante expected payoff is

$$\sum_{\theta} p(\theta) \sum_s p(s | \theta) \sum_t p(t | \theta, s) [E_R(\theta | t) - cs] = \sum_t p(t) E_R(\theta | t) - cE(s)$$

The relation between the firm’s objective and the “systematic fooling” question is apparent from this expression.

If R is fully connected - such that the consumer has rational expectations - the firm’s objective function collapses into $E(\theta) - cE(s)$. In this case, the firm cannot use sponsored reviews to manipulate its average reputation, because it coincides with the product’s expected quality. The firm’s ex-ante optimal strategy is $p(s = 1 | \theta) = 0$ for every θ . (Of course, this policy will typically fail to be time-consistent; however, I focus entirely on the *ex-ante* perspective.)

Impose the following additional structure on p . Let $\theta \in \{0, 1\}$; the two values are equally likely. The firm’s strategy can thus be represented by two conditional probabilities:

$$\begin{aligned} \alpha &= p(s = 1 | \theta = 1) \\ \beta &= p(s = 1 | \theta = 0) \end{aligned}$$

Finally, $p(t | \theta, s)$ is degenerate: for every θ, s , $t = \theta + s$ with probability one. That is, sponsoring a review uniformly biases it upward.

In this example, the consumer’s DAG captures his understanding of the process that generates review content. The DAGs $\theta \rightarrow t$ and $s \leftarrow \theta \rightarrow t$ capture “naive” consumers, who are unaware of sponsorship, or fail to realize that it may directly affect review content. The DAG $\theta \rightarrow s \rightarrow t$ captures a “cynical” consumer, who not only recognizes that review content is potentially affected by sponsorship, but actually believes that it has nothing to do with the product’s quality once we condition on the sponsorship status. These three DAGs are all perfect, and therefore they would lead to quality estimates that are unbiased on average. As a result, the firm’s ex-ante optimal strategy under these DAGs coincides with the rational-expectations prediction.

In contrast, the DAG $R : \theta \rightarrow t \leftarrow s$ is imperfect. A consumer with this DAG realizes that sponsorship may affect reviews, but he believes that the prevalence of sponsorship is independent of the product's quality. This DAG treats s and θ as mutually independent primary causes of t , whereas in reality s may be caused by θ via the firm's strategy. We will now see that thanks to this correlation neglect, the firm can play a strategy that enhances its average reputation.

Proposition 4 *Let $R : \theta \rightarrow t \leftarrow s$. Then, the firm's ex-ante optimal strategy is $\alpha = 0$, $\beta = \frac{1}{2} - c$. The firm's average reputation under the ex-ante optimal strategy is*

$$\sum_t p(t) E_R(\theta | t) = \frac{1}{2} + \frac{1}{16}(1 - 4c^2) > \frac{1}{2} = E(\theta)$$

Note that even in this case, the firm's ability to manipulate its average reputation is limited - the firm is unable to increase it by more than $\frac{1}{16}$.

A natural question in this context is, why would a consumer who is aware of all three variables θ, s, t hold a causal model that does not fully link them? The answer is that my use of a simple three-variable example is due to my own bounded rationality (and possibly the reader's); this simplicity should not be mistaken for a simplicity of the real-life environment the example aims to capture. This environment would typically involve many variables: the quality of numerous types of products, as well as many reviewers and many outlets that publish their reviews. It would be hard for consumers to fully understand the intricate web of influences among these variables. A boundedly rational consumer is likely to make simplifying assumptions - e.g., that sponsorship patterns are independent of product quality. This particular assumption enables firms to use sponsored reviews to manipulate their average reputation.

4.2 Monetary Policy

In this sub-section I analyze the “monetary policy” example that was outlined in the Introduction. Recall the four economic variables: the exogenous “macro indicator” θ , the central bank’s action a , inflation π and real output y . The private sector’s inflation forecast is denoted e . The central bank’s payoff is $y - \theta\pi$.

The objective distribution p satisfies the following properties. First, $\theta \sim U[0, 1]$; the central bank privately observes θ prior to taking its action. Both π and a take values in $\{0, 1\}$, where $\pi = 0$ (1) represents low (high) inflation. The central bank’s strategy is defined by a collection of conditional probabilities: $\alpha(\theta) = p(a = 1 \mid \theta)$ for every θ . Inflation is a stochastic function of a , given by $p(\pi = 1 \mid a) = \beta a$, where $\beta \in (0, 1)$. That is, $a = 0$ is a safe action that induces low inflation with certainty, whereas $a = 1$ is a risky action that induces high inflation with probability β . The private sector forms its inflation forecast after observing the realization of a - i.e., $e = E_R(\pi \mid a)$. Output is given by the “Phillips Curve” $y = \pi - e + \eta$, where $\eta \sim N(0, \sigma_\eta^2)$ is independently distributed. Note that p is consistent with the following “true DAG” R^* defined over θ, a, π, y :⁴

$$\begin{array}{ccccc}
 \theta & \rightarrow & a & \rightarrow & \pi \\
 & & & \searrow & \downarrow \\
 & & & & y
 \end{array} \tag{6}$$

Plugging the Phillips Curve into the central bank’s payoff function, we obtain the following:

$$\begin{aligned}
 & \sum_{\theta} p(\theta) \sum_a p(a \mid \theta) [(1 - \theta) \cdot E(\pi \mid a) - E_R(\pi \mid a)] \\
 & = E(\pi) - \sum_a p(a) E_R(\pi \mid a) - \sum_{\theta} p(\theta) E(\pi \mid \theta) \theta
 \end{aligned}$$

If the private sector had rational expectations ($R = R^*$), this expression

⁴If we incorporated e as an explicit variable in the causal model, the direct link $a \rightarrow y$ would be replaced with the chain $a \rightarrow e \rightarrow y$. See Section 6.1 for a discussion of estimates as variables in DAGs.

would collapse into

$$-\sum_{\theta} p(\theta)E(\pi | \theta)\theta = -\beta \sum_{\theta} p(\theta)p(a = 1 | \theta)\theta$$

and the central bank's ex-ante optimal strategy would be $p(a = 1 | \theta) = 0$ for every $\theta > 0$.⁵

Consider two possibilities for the private sector's DAG:

$$R: \theta \rightarrow a \rightarrow \pi \leftarrow y \qquad R': \theta \rightarrow a \rightarrow \pi \rightarrow y$$

Each DAG represents a different narrative about how macro variables are interconnected. (Hoover (2001) describes historical controversies in macroeconomics in terms of conflicting causal mechanisms.) The DAG R' is misspecified; it deviates from the true DAG R^* by omitting the direct link $a \rightarrow y$. In other words, it neglects the causal channel from monetary policy to output via inflationary expectations. Because R' is perfect, Proposition 2 implies that inflation and output forecasts based on R' would be unbiased. In particular, the central bank cannot use monetary policy to create systematic output effects. As a result, the central bank's policy under R' will coincide with the rational-expectations prediction.

Now turn to the imperfect DAG R , which deviates from R' by *reversing* the causal link between inflation and output. This DAG may capture a prior belief in "classical" monetary neutrality, because it admits *no* causal channel from a to y . The private sector's conditional inflation forecast under R is

$$E_R(\pi | a) = \sum_{\pi} p_R(\pi | a)\pi = \sum_y p(y)p(\pi = 1 | a, y)$$

because according to R , y and θ are both independent of θ conditional on a . We will now see that under R , the central bank's ex-ante optimal strategy involves inflating (as the Phillips Curve noise vanishes).

⁵As in Section 4.1, I focus on ex-ante optimality, without addressing the time-consistency issue. Note that there would be no time-inconsistency problem if the private sector had rational expectations, because it observes a . This property is no longer assured when the private sector's DAG is misspecified.

Proposition 5 *In the $\sigma_\eta^2 \rightarrow 0$ limit, the central bank's ex-ante optimal policy is a cutoff strategy: $p(a = 1 | \theta) = 1$ if $\theta < \frac{1}{3}$ and $p(a = 1 | \theta) = 0$ if $\theta > \frac{1}{3}$. The expected output is $\frac{2}{9}\beta$.*

The intuition behind the result is as follows. When the central bank plays $a = 0$, it induces $\pi = 0$ with certainty, hence $p(\pi = 1 | a = 0; y) = 0$ for any y . As a result, the private sector's incorrect account for the variation in y does not lead to a biased inflation forecast. In contrast, when $a = 1$, inflation does fluctuate; the private sector's error is that it tries to account for these fluctuations by the variation in y , as if the latter is exogenous. Therefore, the private sector's inflation forecast conditional on $a = 1$ involves summing over all values of y , *without* conditioning y on $a = 1$. In the $\sigma_\eta^2 \rightarrow 0$ limit, this failure to condition on a translates to the identity $E_R(\pi | a = 1) = E_R(\pi)$. Thus, when the central bank plays $a = 0$, the private sector correctly updates its belief downward, whereas when the central bank plays $a = 1$, the private sector forms its inflation forecast as if it did not observe the central bank's action. This leads to systematic underestimation of expected inflation.

Testing and revising models

The notion of model-based predictions naturally raises the problem of testing and revising subjective causal models - especially under the “professional forecasting” interpretation, which takes R to be an explicit, formal model. The forecaster can test the model's assumptions or its predictions. A DAG embodies a collection of conditional-independence assumptions. E.g., the DAG R in this sub-section assumes (among other such properties) that $\pi \perp \theta | a$ and $y \perp a$. The forecaster can test these assumptions by directly measuring $p(\theta, \pi, a)$ and $p(a, y)$. He will then realize that p satisfies the former assumption but violates the latter. Likewise, the forecaster can test his model's predictions - e.g., measuring $E(\pi | a)$ and $E(y | a)$ and comparing them to the forecasts $E_R(\pi | a)$ and $E_R(y | a)$. He will then learn that his inflation forecast is systematically biased, whereas the output forecast is correct on average. The forecaster will face the dilemma of whether to revise his model in face of its mixed performance.

My perspective in this paper is that model revision is an infrequent activity. First, it is costly to measure, process and communicate correlations - indeed, this is one of the reasons I listed for using a sparse model that can generate many conditional predictions for a variety of purposes. The forecaster only measures the correlations that quantify his model; and he may avoid additional costly measurements required for testing it. Second, because the forecaster's model is a knowing simplification of a complex environment, he expects it to get some things wrong, and he may be unfazed by the model's mixed performance (defending his stance with familiar statements like "every model is wrong" or "it takes a model to beat a model"). Finally, if R reflects strong theoretical preconceptions, the forecaster will resist revisions that challenge them, and he will prefer minor complications of the model that leave core tenets untouched..

While modeling this process of testing and revising subjective causal models is outside this paper's scope, I now consider an example of an alternative DAG that could be adopted as a result of such a process. As mentioned earlier, R reflects a belief in *monetary neutrality* - i.e., denying any causal chain from a to y . But R makes the stronger assumption $a \perp y$. When the forecaster carries out an empirical test that refutes this assumption, it would be natural for him to tweak his model by adding the link $\theta \rightarrow y$ into his model, thus obtaining the DAG R'' given by (1). This allows a and y to be correlated, without challenging the core idea that a does not *cause* y .

Since the modified DAG R'' violates the sufficient condition of Proposition 3 for unbiased inflation forecasts, the central bank may still be able to use monetary policy for real effects. According to R'' , the private sector perceives the correlation between y and a only through their correlation with θ . Consequently, if the central bank plays a deterministic strategy (such as the cutoff strategy derived in Proposition 5), then $p_{R''}(y | a) = p(y | a)$, and therefore the private sector's inflation forecast will be consistent with rational expectations. It follows that the central bank must employ *randomization*. For example, it could play $p(a | \cdot) = \alpha$ for all θ , such that the private sector will perceive no correlation between a and y - i.e., $p_{R''}(y | a) = p(y)$ - and expected output will be $\beta\alpha(1 - \alpha)$, just as under R . However, this is not

the most cost-effective way of generating real effects. The following result derives the ex-ante optimal strategy.

Proposition 6 *In the $\sigma_\eta^2 \rightarrow 0$ limit, the central bank's ex-ante optimal strategy under R'' is $p(a = 1 \mid \theta) = \frac{1}{2}(1 - \theta)$ for every θ . The expected output in this limit is $\frac{1}{6}\beta$.*

Thus, the central bank plays a random strategy, where the probability that it tries to inflate is a simple decreasing function of the cost of inflation. The expected output that it generates is lower than under the original specification of R , because the private sector's partial ability to account for the correlation between a and y means that the central bank can no longer use the cost-effective cutoff strategy.

5 Extensions

In this section I extend the basic analysis in various directions.

5.1 Multivariate Normal Distributions

Proposition 2 means that an imperfect DAG exposes the agent to “systematic fooling” for *some* objective distribution. However, in applications we often restrict the domain of objective distributions, and this makes it harder to systematically fool our agent. In this section I examine the implications of a domain restriction that is common in economic models, namely that p is a multivariate normal distribution.

Proposition 7 *Let R be an arbitrary DAG, and let p be a multivariate normal distribution. Then, for every $i = 1, \dots, n$,*

$$\sum_{x_0} p(x_0) E_R(x_i \mid x_0) = E(x_i)$$

Thus, when p is multivariate normal, *any* DAG leads to unbiased forecasts; the mere assumption that the agent forms his beliefs by fitting a causal model to long-run data guarantees that he cannot be systematically fooled. The key to this finding is an existing result in the Bayesian-networks literature (see Koller and Friedman (2009, Ch. 7)), which establish an equivalence between multivariate normal distributions and “Gaussian network representations”. Consequently, factorizing p according to R preserves the nice properties of multivariate normal distributions, including the simple linear formulas for conditional expectations. In turn, this linearity ensures that conditional estimates are correct on average.

In each of the applications of Section 4, one of the variables was an *action* taken by some other agent (the firm in Section 4.1, the central bank in Section 4.2). Proposition 7 implies that in such cases, if that other agent plays a linear-normal strategy (and all other variables are linked by a system of linear-normal equations), our agent will never be systematically fooled. Thus, in linear-normal models, non-linear strategies are necessary for inducing systematically biased predictions.⁶

5.2 Observing Multiple Variables

So far, we have assumed that the agent conditions his estimates on a single observed variable x_0 . Now suppose that the agent’s signal is x_A , where $A \subset N$ is non-empty and may include more than one node. In a standard model with rational expectations, we can always redefine the agent’s signal as a single variable, without loss of generality. However, when the agent’s beliefs are based on a misspecified DAG, it is important to be explicit about the variables that constitute the agent’s signal. The agent’s estimate of x_i conditional on observing x_A is $E_R(x_i | x_A)$. We say that R induces unbiased estimates if for every objective distribution p and every $i \in N - A$, $\sum_{x_A} p(x_A) E_R(x_i | x_A) =$

⁶In the main part of the “monetary policy” example, the impossibility of systematically biased estimates extends to arbitrary non-linear central-bank strategies (as well as arbitrary distributions $p(\theta)$). It suffices to assume that inflation and output are determined according to linear-normal equations. The reason is that according to R , θ and a form an ancestral clique, and $\pi, y \perp \theta | a$.

$E(x_i)$.

Proposition 8 *A perfect DAG R induces unbiased estimates if and only if A is a union of mutually disconnected cliques.*

Thus, even when R is perfect, it may still give rise to biased estimates when the agent conditions his estimates on multiple variables that are connected by R but fail to form a clique. However, as long as A is a clique (or a collection of disconnected cliques), the agent's estimates are unbiased.

Example 5.1: A no-trade theorem

Another economic phenomenon in which the possibility of systematically biased estimates plays a key role is speculative trade in financial markets. In principle, when different traders have different subjective models, this can lead to belief heterogeneity and thus potentially gives rise to speculative trade.

Consider the following standard trading game. There is a collection of m risk-neutral traders. Each trader i has access to a set of trading actions S . Let θ be the state of Nature, and let t_i represent a signal that trader i receives prior to making his choice of trading strategy s_i . As usual, $s_i \perp (\theta, t_{-i}, s_{-i}) \mid t_i$ - i.e., the trader's action is independent of the state of Nature and the other traders' signals and actions, conditional on his signal. Let $z = (z_1, \dots, z_m)$ be a zero-sum vector of monetary transfers among traders, which is some stochastic function of θ and s_1, \dots, s_m , which satisfies the following property: there exists a default no-trade action $s^0 \in S$, such that if trader i plays $s_i = s^0$, he gets $z_i = 0$ with probability one, for all s_{-i} and θ . Trader i 's utility function is $u_i(z_i, s_i) = z_i - c \cdot \mathbf{1}(s_i \neq s^0)$, where $c > 0$ is the (arbitrarily small) cost of taking a non-default action.

The variables that are allowed to feature in the traders' causal models are θ and $(t_i, s_i, z_i)_{i=1, \dots, m}$. Assume that trader i 's DAG R_i admits the three variables t_i, s_i, z_i , and that it contains the link $t_i \rightarrow s_i$. A justification for this assumption is that because the trader considers conditioning his action on his signal, he acknowledges this as a causal effect. A strategy for trader i

is given by the conditional probabilities $(p(s_i | t_i))_{t_i, s_i}$. We say that a profile of strategies is an equilibrium if, whenever $p(s_i | t_i) > 0$,

$$s_i \in \arg \max_{s \in S} \sum_{z_i} p_{R_i}(z_i | t_i, s_i) u_i(z_i, s_i)$$

That is, if a trader plays an action with positive probability after observing some signal, this action must be a subjective expected-utility maximizer according to his updated subjective belief.

Proposition 9 *Suppose that R_i is perfect for every $i = 1, \dots, m$. Then, the unique equilibrium is $p(s_i = s^0 | t_i) = 1$ for every i and every t_i .*

The impossibility of biased estimates under perfect DAGs (in which a trader’s signal and his action constitute a clique) plays a crucial role in this result. Each trader’s prediction of his earnings conditional on his trading action and his information is unbiased on average, and this is what precludes speculative trade, despite the possible heterogeneity in the traders’ subjective models.

Ex-ante estimates

Suppose that the set of nodes A is empty. This corresponds to a situation in which the agent forms ex-ante, unconditional estimates of individual variables. The question then becomes whether the agent’s marginal subjective distributions over individual variables are correct. Formally, we will say that a DAG R induces unbiased ex-ante estimates if $E_R(x_i) \equiv E(x_i)$ for every objective distribution p and every $i \in N$. The following is a corollary of Proposition 2 in Spiegler (2016b): R induces unbiased ex-ante estimates if and only if it is *perfect*. Thus, perfection emerges as a critical property for unbiased predictions, whether they are conditional on unconditional.

5.3 Estimates as Variables

Throughout the paper, I assumed that the agent’s DAG does not admit his own estimates as variables. However, estimates or forecasts are themselves

variables that can play a role in the determination of economic outcomes - e.g., recall the Phillips Curve in the “monetary policy” example. In principle, they also could enter the agent’s subjective causal model. Denote $x_{i+n} = e_i$ for every $i = 1, \dots, n$, and $x = (x_0, x_1, \dots, x_{2n})$. Allow the set of nodes N in the agent’s DAG to be a subset of the enlarged set $\{0, 1, \dots, 2n\}$. Thus, when $i \in N$ for some $i > n$, this means that the agent’s causal model admits e_{i-n} as a variable. Recall our earlier restriction that $0 \in N$. The following is a sensible additional restriction.

Condition 2 *If $i \in N$ for some $i > n$, then $R(i) = \{0\}$ and $i - n \in N$.*

This condition requires two things. First, it says that the agent perceives x_0 to be the only immediate cause of his own estimates. The justification is that the agent conditions his estimates on x_0 and on no other variable; his causal model acknowledges this. Second, it requires that if the agent’s DAG includes an estimate of some variable, it must also admit the variable itself.

These two domain restrictions imply the following result.

Proposition 10 *Suppose that the domain of permissible objective distributions satisfies Condition 1 and that R satisfies Condition 2 (as well as the requirement that $0 \in N$). Then, there is a DAG R' that omits the nodes $n + 1, \dots, 2n$ altogether, such that $p_{R'}(x_{N-\{n+1, \dots, 2n\}}) \equiv p_R(x_{N-\{n+1, \dots, 2n\}})$ for every p in the restricted domain.*

This result means that our original assumption that the agent’s DAG omits his own forecasts is without loss of generality, as long as we accept the domain restrictions on p and R .

6 Related Literature

This paper contributes to a literature that studies decisions that are based on misspecified subjective models. Some of the existing approaches to modeling non-rational expectations can be embedded (at least for some domains) in

the DAG formalism. These approaches include the concepts of analogy-based expectations (Jehiel (2005)), “cursed” beliefs (Eyster and Rabin (2005)), or behavioral equilibrium (Esponda (2008)). See Spiegler (2016a) for a detailed explanation of these connections.

Within this literature, Piccione and Rubinstein (2003) share the “expectations management” aspect of the examples in Section 4. In their model, a seller commits to a deterministic temporal sequence of prices, taking into account that consumers (who play the role of the agent in this paper) can only perceive statistical patterns that allow the price at any period t to be a function of price realizations at periods $t - 1, \dots, t - k$, where k is a constant that characterizes the consumer. When the value of k is negatively correlated with consumers’ willingness to pay, the seller may want to generate a complex price sequence as a discrimination device. Relatedly, Ettinger and Jehiel (2010) study a bargaining model, in which a sophisticated seller employs deception tactics that lead a buyer who exhibits coarse reasoning to form a biased estimate of the traded object’s value.

Spiegler (2016b) interprets the Bayesian-network factorization formula as a representation of *objective* data limitations, such that the agent’s belief is a consequence of applying a certain extrapolation method to his limited data. In particular, Spiegler (2016b) shows that when R is perfect, p_R is the outcome of extrapolating a belief from incomplete datasets drawn from p , via an iterative variant on a method known as “conditional stochastic imputation”. From this point of view, perfect DAGs capture implicit data limitations rather than an explicit causal model.

The “monetary policy” example links the paper to a few works that examine monetary policy when the rational-expectations assumption is relaxed. Evans and Honkapohja (2001) and Woodford (2013) review dynamic macroeconomic models in which agents form non-rational expectations, and explore implications for monetary policy. Garcia-Schmidt and Woodford (2015) is a recent exercise in this tradition. The most closely related equilibrium concept that is employed in this literature is known as “restricted perceptions equilibrium”, which is based on a notion of coarse beliefs in the same spirit as Piccione and Rubinstein (2003) and Jehiel (2005). Sargent (2001), Cho

et al. (2002) and Esponda and Pouzo (2016) study models in which it is the *central bank* that forms non-rational expectations, whereas the private sector is modeled conventionally.

Finally, the general idea of modeling economic agents as econometricians or statisticians has many precedents. This is typically done in learning, non-equilibrium models (e.g. Bray (1982)). There are a few recent examples of *equilibrium* concepts that treat agents as (possibly flawed) statisticians - see Esponda and Pouzo (2016), Cherry and Salant (2016) and Liang (2016).

7 Conclusion

This paper explored the possibility of systematically fooling agents with causal misperceptions. Although I provided several examples that demonstrated this possibility, a somewhat surprising feature of the analysis was the ubiquity of *impossibility* results. Subjective causal models represented by perfect DAGs rule out systematically biased estimates; and if the objective distribution is multivariate-normal, this impossibility extends to *all* DAGs. Finally, even when biased estimates were possible, we saw that their magnitude in concrete examples was constrained. Thus, the mere process of forming beliefs by fitting a causal model to long-run data restricts a third party's ability to exploit the beliefs' departure from rational expectations.

In the “monetary policy” example, negative findings along these lines mean that classical results regarding the non-exploitability of the Phillips relation continue to hold even when the private sector forms beliefs according to a misspecified model. This lesson is intriguing, considering the heated historical debates over the exploitability of the inflation-output relation (see Klammer (1984)). The key assumption behind classical non-exploitability results (Lucas (1972), Sargent and Wallace (1975)) was allegedly the rationality of private-sector expectations. The impossibility results of this paper put these historical debates in a new perspective.

References

- [1] Athey, S., A. Atkeson and P. Kehoe (2005), “The Optimal Degree of Discretion in Monetary Policy,” *Econometrica* 73, 1431-1475.
- [2] Barro, R. and D. Gordon (1983), “Rules, Discretion and Reputation in a Model of Monetary Policy,” *Journal of Monetary Economics* 12, 101-121.
- [3] Cherry, J. and Y. Salant (2016), “Statistical Inference in Games,” Northwestern University, Mimeo.
- [4] Cho, I., N. Williams and T. Sargent (2002), “Escaping Nash Inflation,” *Review of Economic Studies*, 69, 1-40.
- [5] Cowell, R., P. Dawid, S. Lauritzen and D. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems*, Springer, London.
- [6] Esponda, I. (2008), “Behavioral Equilibrium in Economies with Adverse Selection,” *The American Economic Review*, 98, 1269-1291.
- [7] Esponda, I. and D. Pouzo (2016), “Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models,” *Econometrica* 84, 1093-1130.
- [8] Ettinger, D. and P. Jehiel (2010), “A Theory of Deception,” *American Economic Journal: Microeconomics* 2, 1-20.
- [9] Evans, G. and S. Honkapohja (2001), *Learning and Expectations in Macroeconomics*, Princeton University Press.
- [10] Eyster, E. and M. Rabin (2005), “Cursed Equilibrium,” *Econometrica*, 73, 1623-1672.
- [11] Garcia-Schmidt, M. and M. Woodford (2015), “Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis,” NBER Working Paper no. w21614.

- [12] Giacomini, R. (2015), “Economic Theory and Forecasting: Lessons from the Literature,” *Econometrics Journal* 18, C22-C41.
- [13] Hoover, K. (2001), *Causality in Macroeconomics*, Cambridge University Press.
- [14] Jehiel, P. (2005), “Analogy-Based Expectation Equilibrium,” *Journal of Economic Theory*, 123, 81-104.
- [15] Klamler, A. (1984), *The New Classical Macroeconomics: Conversations with the New Classical Economists and their Opponents*. Wheatsheaf Books.
- [16] Koller, D. and N. Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques*. MIT press.
- [17] Kydland, F. and E. Prescott (1977), “Rules rather than Discretion: The Inconsistency of Optimal Plans,” *Journal of Political Economy* 85, 473-491.
- [18] Liang, A. (2016), “Games of Incomplete Information Played by Statisticians,” Harvard University, Mimeo.
- [19] Lucas, R. (1972), “Expectations and the Neutrality of Money,” *Journal of Economic Theory* 4, 103-124.
- [20] Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge.
- [21] Piccione, M. and A. Rubinstein (2003), “Modeling the Economic Interaction of Agents with Diverse Abilities to Recognize Equilibrium Patterns,” *Journal of the European Economic Association* 1, 212-223.
- [22] Sargent, T. (2001), *The Conquest of American inflation*, Princeton University Press.
- [23] Sargent, T. (2003), “Rational Expectations,” *The Concise Encyclopedia of Economics*.

- [24] Sargent, T. and N. Wallace (1975), “‘Rational’ Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule,” *Journal of Political Economy* 83, 241-254.
- [25] Sloman, S. (2005), *Causal Models: How People Think about the World and its Alternatives*, Oxford University Press.
- [26] Spiegler, R. (2016a), “Bayesian Networks and Boundedly Rational Expectations,” *Quarterly Journal of Economics* 131, 1243-1290.
- [27] Spiegler, R. (2016b), “Data Monkeys: A Procedural Model of Extrapolation from Partial Statistics,” *Review of Economic Studies*, forthcoming.
- [28] Verma, T. and J. Pearl (1991), “Equivalence and Synthesis of Causal Models,” *Uncertainty in Artificial Intelligence*, 6, 255-268.
- [29] Woodford, M. (2013), “Macroeconomic Analysis without the Rational Expectations Hypothesis,” *Annual Review of Economics*, forthcoming.

Appendix: Proofs

Proposition 2

(Only if). Consider the special case in which $X_i = \{0, 1\}$ for every i , such that the expected value of any x_i w.r.t any distribution is equal to the probability that $x_i = 1$. I will explain at the end of the proof why this is w.l.o.g. When R is imperfect, it must contain a v -collider $i \rightarrow j \leftarrow k$. Let us consider objective distributions p with full support on X_N , for which all other variables are independent, such that

$$p_R(x_N) = p(x_i)p(x_k)p(x_j | x_i, x_k) \cdot \prod_{i' \in N - \{i, j, k\}} p(x_{i'})$$

This allows us to ignore all variables $i' \in N - \{i, j, k\}$ when calculating marginal or conditional distributions over x_j that are derived from p_R .

There are three cases to consider. First, suppose that $0 \notin \{i, j, k\}$ - i.e., 0 is not part of the v -collider. Then, $p_R(x_j | x_0) \equiv p_R(x_j)$. By Proposition

1, j is not an ancestral node in any DAG in the equivalence class of R . Therefore, by Lemma 1, we can find p for which $p_R \neq p$. (Our restrictions on p are w.l.o.g in this regard, because we can ignore all nodes $i' \neq i, j, k$ and set $R : i \rightarrow j \leftarrow k$.)

Second, suppose that $i = 0$. Then,

$$p_R(x_j = 1 \mid x_0) = \sum_{x_k} p(x_k) p(x_j = 1 \mid x_0, x_k)$$

Impose the following additional structure on p . First, $p(x_0 = 1) = \frac{1}{2}$. Second, $x_k = x_j = x_0$ with arbitrarily high probability. Third, $p(x_j = 1 \mid x_0 \neq x_k)$ is arbitrarily low. Then,

$$\sum_{x_0} p(x_0) p_R(x_j = 1 \mid x_0) = \frac{1}{2} \left\{ \sum_{x_k} p(x_k) [p(x_j = 1 \mid x_0 = 0; x_k) + p(x_j = 1 \mid x_0 = 1; x_k)] \right\}$$

is arbitrarily close to $\frac{1}{4}$, whereas $p(x_j = 1) = \frac{1}{2}$.

Finally, suppose that $j = 0$. Then,

$$p_R(x_i = 1 \mid x_0) = \frac{\sum_{x_k} p(x_k) p(x_i = 1) p(x_0 \mid x_i = 1; x_k)}{\sum_{x_k} p(x_k) \sum_{x_i} p(x_i) p(x_0 \mid x_i; x_k)}$$

Impose the following additional structure on p . First, $p(x_k = 1) = \frac{1}{2}$. Second, $p(x_i = x_k)$ with arbitrarily high probability. Third, $p(x_0 = 1 \mid x_i, x_k)$ is arbitrarily high when $x_i x_k = 1$ and arbitrarily low when $x_i x_k = 0$. Then, $p_R(x_i = 1 \mid x_0 = 1)$ is arbitrarily close to 1, and $p_R(x_i = 1 \mid x_0 = 0)$ is arbitrarily close to $\frac{1}{3}$, such that $\sum_{x_0} p(x_0) p_R(x_i = 1 \mid x_0)$ is arbitrarily close to $\frac{2}{3}$, whereas $p(x_i = 1) = \frac{1}{2}$.

Extending the proof to arbitrarily large X is straightforward - we only need to assume that the marginal of p over each of the variables x_i, x_j, x_k assigns arbitrarily high total probability to two arbitrary values, and that the small probability that is assigned to each of the other values is independently distributed.

Proposition 3

Denote $N^* = \{j \in N \mid jP0 \text{ or } jPi\}$. By definition, $0, i \in N^*$. It is immediate from (4) that when calculating $p_R(x_0, x_i)$, we can disregard any $k \notin N^*$. Therefore, we only need to consider the subgraph over N^* induced by R . By assumption, it is perfect. Therefore, the same proof of the “if” part of Proposition 2 applies here, and implies the result.

Proposition 4

The consumer’s quality assessment after observing $t = 2$ is

$$\begin{aligned} p_R(\theta = 1 \mid t = 2) &= \frac{p_R(\theta = 1, t = 2)}{p_R(t = 2)} = \frac{p(\theta = 1) \sum_s p(s) p(t = 2 \mid s, \theta = 1)}{\sum_\theta p(\theta) \sum_s p(s) p(t = 2 \mid s, \theta)} \\ &= \frac{p(\theta = 1) p(s = 1)}{p(\theta = 0) \sum_s p(s) \cdot 0 + p(\theta = 1) p(s = 1)} = 1 \end{aligned}$$

because the realization $t = 2$ is possible only when $\theta = 1$. Likewise, the realization $t = 0$ is possible only when $\theta = 0$, and a similar calculation yields $E_R(\theta \mid t = 0) = 0$. It follows that when $t \neq 1$, the consumer’s quality estimate is consistent with rational expectations.

Let us turn to the consumer’s quality assessment after observing $t = 1$:

$$\begin{aligned} p_R(\theta = 1 \mid t = 1) &= \frac{p(\theta = 1) \sum_s p(s) p(t = 1 \mid s, \theta = 1)}{\sum_\theta p(\theta) \sum_s p(s) p(t = 1 \mid s, \theta)} \\ &= \frac{p(\theta = 1) p(s = 0)}{p(\theta = 1) p(s = 0) + p(\theta = 0) p(s = 1)} \\ &= p(s = 0) = \frac{1}{2}(1 - \alpha) + \frac{1}{2}(1 - \beta) = 1 - \frac{1}{2}(\alpha + \beta) \end{aligned}$$

We can now calculate the firm’s expected payoff for any strategy (α, β) :

$$\frac{1}{2} \cdot \alpha \cdot 1 + \left[\frac{1}{2} \cdot (1 - \alpha) + \frac{1}{2} \cdot \beta \right] \cdot \left[1 - \frac{1}{2}(\alpha + \beta) \right] - c \cdot \left(\frac{1}{2} \cdot \alpha + \frac{1}{2} \cdot \beta \right)$$

The strategy (α, β) that maximizes this expression is $\alpha = 0, \beta = \frac{1}{2} - c$. That is, the firm sponsors reviews only when its quality is low, and even then only with some probability. Plugging the values of α, β into the expression for the firm’s average reputation yields the result.

Proposition 5

Denote $E_R(\pi | a) = e(a)$ and $\alpha = p(a = 1) = \sum_{\theta} p(\theta)p(a = 1 | \theta)$. Because $\pi \in \{0, 1\}$,

$$e(a) = \sum_y p(y)p(\pi = 1 | a, y)$$

Because η is normally distributed, $p(a, y)$ has full support, such that $e(a)$ never involves conditioning on zero-probability events. Because $y \perp \theta | a$ according to the true process, α is the only aspect of the central bank's strategy that is relevant for calculating $e(a)$.

Let us first calculate $e(0)$. Because $p(\pi = 1 | a = 0) = 0$, it follows that $p(\pi = 1 | a = 0, y) = 0$ for all y . Therefore, $e(0) = 0$. This in turn means that $E(y | a = 0) = 0$. It follows that if $\alpha = 0$, the central bank cannot induce strictly positive expected output. From now on, assume $\alpha > 0$.

Let us now calculate $e(1)$. First, note that $y \sim N(\mu, \sigma_{\eta}^2)$, where μ is random: $\mu = e(0) = 0$ with probability $1 - \alpha$, $\mu = 1 - e(1)$ with probability $\alpha\beta$, and $\mu = -e(1)$ with probability $\alpha(1 - \beta)$. A priori, two of these three values could coincide. However, we will now see that this is not the case. Because the normal distribution is symmetrically distributed around its mean, the ex-ante probability of $y < -e(1)$ is at least $\alpha(1 - \beta)/2$, whereas the ex-ante probability of $y > 1 - e(1)$ is at least $\alpha\beta/2$. Moreover, as σ_{η}^2 tends to 0, $p(\pi = 1 | a = 1, y < -e(1)) \rightarrow 0$ and $p(\pi = 1 | a = 1, y > 1 - e(1)) \rightarrow 1$. Therefore, in the $\sigma_{\eta}^2 \rightarrow 0$ limit,

$$0 < \frac{\alpha\beta}{2} \leq e(1) \leq 1 - \frac{\alpha(1 - \beta)}{2} < 1$$

It follows that as σ_{η}^2 approaches zero, μ gets *exactly* three values, $-e(1)$, 0 and $1 - e(1)$, and the gap between these values is bounded away from zero. In the $\sigma_{\eta}^2 \rightarrow 0$ limit, $p(\pi = 1 | a = 1, y) \rightarrow 1$ in the neighborhood of $y = 1 - e(1)$, whereas $p(\pi = 1 | a = 1, y) \rightarrow 0$ in the neighborhoods of $y = 0$ and $y = -e(1)$. Consequently, $e(1) \rightarrow p(\pi = 1) = \alpha\beta$ as $\sigma_{\eta}^2 \rightarrow 0$.

We have thus established that $E(\pi) = \alpha\beta$ and $\sum_a p(a)e(a) = \alpha \cdot \alpha\beta + (1 - \alpha) \cdot 0 = \alpha^2\beta$. Moreover, for any given α , it is optimal for the central bank to use a cutoff strategy - i.e., there will exist θ^* such that $F(\theta^*) = \alpha$,

and the central bank will play $a = 1$ if and only if $\theta < \theta^*$. The reason is that for any given α , the way the central bank allocates this probability to different values of θ does not change the expected output, and therefore the central bank would like to implement α in the least costly manner - which means playing $a = 1$ for low values of θ . Because θ is uniformly distributed over $[0, 1]$, it follows that the central bank will choose θ^* to maximize

$$\beta\theta^*(1 - \theta^*) - \beta \int_0^{\theta^*} \theta d\theta$$

The solution follows immediately.

Proposition 6

The argument that $p_{R''}(\pi = 1 | a = 0) = 0$ is the same as in Proposition 5. Denote $\alpha(\theta) = p(a = 1 | \theta)$ and $e(1) = p_{R''}(\pi = 1 | a = 1)$. Then, expected output is

$$\sum_{\theta} p(\theta)\alpha(\theta)[\beta - e(1)]$$

Now,

$$\sum_{\theta} p(\theta)\alpha(\theta)e(1) = \sum_{\theta} p(\theta)\alpha(\theta) \sum_y p_{R''}(y | a = 1, \theta)p(\pi = 1 | y, a = 1)$$

According to the true process, $\pi \perp \theta | (a, y)$, just as in the case of Proposition 5. Therefore, as in the previous result, in the $\sigma_{\eta}^2 \rightarrow 0$ limit, $p(\pi = 1 | y, a = 1) \rightarrow 1$ at $y = e(1)$ and $p(\pi = 1 | y, a = 1) \rightarrow 0$ at $y \neq e(1)$. It remains to calculate $p_{R''}(y | a = 1, \theta)$. According to R'' , $y \perp a | \theta$. Therefore, $p_{R''}(y | a = 1, \theta) = p_{R''}(y | \theta)$. Because the nodes y and θ form an ancestral clique of R'' , $p_{R''}(y, \theta) = p(y, \theta)$, and therefore

$$\begin{aligned} p_{R''}(y | \theta) &= p(y | \theta) = \sum_{a'} p(a' | \theta)p(y | a') \\ &= \alpha(\theta)p(y | a' = 1) + (1 - \alpha(\theta))p(y | a' = 0) \end{aligned}$$

In the $\sigma_{\eta}^2 \rightarrow 0$ limit, $p(y = e(1) | a' = 1) = p(\pi = 1 | a' = 1) = \beta$. It follows

that expected output is

$$\sum_{\theta} p(\theta) \alpha(\theta) [\beta - \alpha(\theta) \beta]$$

such that the central bank's objective function becomes

$$\sum_{\theta} p(\theta) \{ \alpha(\theta) [\beta - \alpha(\theta) \beta] - \alpha(\theta) \beta \theta \}$$

Note that this objective function is additivity separable in θ , such that for every θ , the optimal value of $\alpha(\theta)$ maximizes

$$\alpha(\theta) [1 - \alpha(\theta)] - \alpha(\theta) \theta$$

which immediately gives the solution.

Proposition 7

Let $p \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From now on, I will assume $\boldsymbol{\mu} = \mathbf{0}$. To see why this is w.l.o.g, note that we could define the auxiliary vector $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}$, such that for every i , $E_R(y_i | y_0) \equiv E_R(x_i | x_0) - \mu_i$ and $E(y_i) \equiv E(x_i) - \mu_i$. If we prove our result for the \mathbf{y} variables, it immediately implies the result for \mathbf{x} .

By a standard result (e.g., Theorem 7.4 in Koller and Friedman (2009)), $p(x_i | x_{R(i)})$ is multivariate normal. Specifically, we can write $p(x_i | x_{R(i)})$ as a linear regression equation with normally distributed noise:

$$x_i \sim N(\boldsymbol{\beta}^T x_{R(i)}, \Sigma_{i,i} - \Sigma_{i,R(i)} \Sigma_{R(i),R(i)}^{-1} \Sigma_{R(i),i})$$

where

$$\boldsymbol{\beta} = \Sigma_{R(i),R(i)}^{-1} \Sigma_{i,R(i)}$$

Thus, the collection $(p(x_i | x_{R(i)}))_{i=1,\dots,n}$ constitutes a Gaussian Bayesian network (see Definition 7.1 in Koller and Friedman (2009)). By Theorem 7.3 in Koller and Friedman (2009), $p_R \sim \mathbf{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$, where $\boldsymbol{\mu}' = \mathbf{0}$. Then, by the definition of conditional expectations under multivariate normal distributions, $E_R(x_i | x_0) = b x_0$, where b is some constant. Because $E(x_0) = 0$, it

then immediately follows that

$$\sum_{x_0} p(x_0) E_R(x_i | x_0) = 0 = E(x_i)$$

which completes the proof.

Proposition 8

(If). Suppose that A is a union of mutually disconnected cliques. This includes the possibility that A itself is a clique. Let $i \in N - A$. If i is disconnected from A , then $p_R(x_i | x_A) = p_R(x_i)$. Since R is perfect, 1 and Lemma 1 imply that $p_R(x_i) = p(x_i)$, hence $E_R(x_i | x_A) = E(x_i)$ for all x_A . Now suppose that i connected to A . By assumption, i is connected to at most one of the cliques that constitute A . Denote this clique by C . Then, $p_R(x_i | x_A) = p_R(x_i | x_C)$. Because R is perfect, Corollary 1 implies that we can take C or $\{i\}$ to be ancestral cliques. By Lemma 1, $p_R(x_C) \equiv p(x_C)$ and $p_R(x_i) \equiv p(x_i)$. Therefore, we can write

$$\sum_{x_C} p(x_C) p_R(x_i | x_C) \equiv \sum_{x_C} p_R(x_C) p_R(x_i | x_C) \equiv p_R(x_i) \equiv p(x_i)$$

which implies the claim.

(Only if). Suppose that A is not a union of mutually disconnected cliques (in particular, A is not a clique itself). Therefore, there exist nodes $j, k \in A$ such that there is a path in \tilde{R} that connects j and k , and yet j and k are not directly linked. Moreover, because R is perfect, there must be at least one such path that does not contain a collider. Without loss of generality, all the nodes along this path do not belong to A , except for j and k themselves. Finally, there must be a node i along this path, such that for some DAG R' in the equivalence class of R , there is a directed path in R' from i into j , as well as a directed path in R' from i into k .

Construct an objective distribution p for which all the variables that lie outside the above path are independent. Moreover, suppose that $x_j \perp x_k$ according to p , and $p(x_j = 1) = p(x_k = 1) = \alpha \in (0, 1)$. Therefore, we can ignore them when calculating $p_R(x_i | x_A)$. As before, we can consider

w.l.o.g the case in which every variable can only take the values 0 and 1. Let $\varepsilon > 0$ be arbitrarily small. Suppose that for every node j' (k') that lies along the path from i to j (k), $x_{j'} = x_j$ ($x_{k'} = x_k$) with independent probability $1 - \varepsilon$. Finally, suppose that $x_i = x_j x_k$ with independent probability $1 - \varepsilon$. By construction,

$$E_R(x_i | x_j, x_k) = p_R(x_i = 1 | x_j, x_k) = \frac{p_R(x_i = 1, x_j, x_k)}{\sum_{x'_i} p_R(x'_i) p_R(x'_i | x_j, x_k)}$$

Because we have assumed that all variables outside the above path are independent, we can ignore these variables and treat the node i as ancestral in R for the purpose of this calculation. Therefore, $p_R(x'_i) = p(x_i)$ for every x_i . Note that $R, x_j \perp x_k | x_i$. Therefore, and by the additional assumptions we imposed on p ,

$$p_R(x_i | x_j, x_k) \approx \frac{p(x_i)p(x_j | x_i)p(x_k | x_i)}{\sum_{x'_i} p(x'_i)p(x_j | x'_i)p(x_k | x'_i)}$$

To calculate this expression, note first that because $x_i = x_j x_k$ with probability close to one, $p(x_i = 1) \approx \alpha^2$ and $p(x_j = 1 | x_i = 1) = p(x_k = 1 | x_i = 1) \approx 1$, whereas

$$p(x_j = 1 | x_i = 0) = p(x_k = 1 | x_i = 0) \approx \frac{\alpha(1 - \alpha)}{1 - \alpha^2} = \frac{\alpha}{1 + \alpha}$$

Plugging these expressions into $p_R(x_i | x_j, x_k)$, we can verify that

$$\sum_{x_j, x_k} p(x_j, x_k) E_R(x_i | x_j, x_k) \neq p(x_i = 1) \approx \alpha^2$$

which completes the proof.

Proposition 9

First, we need to establish that when all traders play s^0 for every signal, this constitutes an equilibrium. Consider trader i 's choice. Because every trader

$j \neq i$ plays $s_j = s^0$, we have $p(z_j = 0) = 1$ for every $j \neq i$ - hence, because z is zero-sum, $p(z_i = 0) = 1$. This means that $p_{R_i}(z_i = 0 | s_i) = 1$ for every s_i , regardless of R_i . Therefore, $s_i = s^0$ is a subjective best-reply for trader i .

Let us now show that there are no other equilibria. By assumption, the action s^0 generates a sure payoff of zero. Therefore,

$$\sum_{z_i} p_{R_i}(z_i | t_i, s^0) u_i(z_i, s^0) = \sum_{z_i} p_{R_i}(z_i | t_i, s^0) z_i = 0$$

Now suppose that $p(s_i | t_i) > 0$ for some trader i , signal t_i and action $s_i \neq s^0$. For every such i, t_i, s_i , we must have

$$\sum_{z_i} p_{R_i}(z_i | t_i, s_i) z_i > 0$$

It follows that

$$\sum_{t_i} \sum_{s_i} p(t_i, s_i) \sum_{z_i} p_{R_i}(z_i | t_i, s_i) z_i \geq 0$$

for all i, t_i, s_i , with a strict inequality for some i, t_i, s_i . Therefore,

$$\sum_{i=1}^m \sum_{t_i} \sum_{s_i} p(t_i, s_i) \sum_{z_i} p_{R_i}(z_i | t_i, s_i) z_i > 0$$

By assumption, R_i contains the link $t_i \rightarrow s_i$. Therefore, the two variables constitute a clique in R_i . By Proposition 8,

$$\sum_{t_i} \sum_{s_i} p(t_i, s_i) \sum_{z_i} p_{R_i}(z_i | t_i, s_i) z_i = E(z_i)$$

hence

$$\sum_{i=1}^m \sum_{t_i} \sum_{s_i} p(t_i, s_i) \sum_{z_i} p_{R_i}(z_i | t_i, s_i) z_i = E\left(\sum_i z_i\right) > 0$$

a contradiction.

Proposition 10

Suppose that $i + n \in N$ for some $i = 1, \dots, n$. Then, by Condition 2, the factorization formula (4) contains the term $p(e_i | x_0)$. Also, $i \in N$. By

assumption, $p(E_R(x_i | x_0) | x_0) = 1$. Therefore, we can remove the term $p(e_i | x_0)$ from (4) altogether, and plug $e_i = E_R(x_i | x_0)$ in any term in (4) that conditions on e_i - which effectively means that such a term conditions on x_0 . We have thus obtained a DAG representation in which the node e is omitted, and any link from e to some node in R is replaced with a link from x_0 into the same node.