Can Agents with Causal Misperceptions be Systematically Fooled?[*]

Ran Spiegler[†]

September 6, 2018

Abstract

An agent forms estimates (or forecasts) of individual variables conditional on some observed 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 average estimates coincide with the variables' true expected value (for any underlying objective distribution) if and only if the agent's graph is perfect - i.e., it directly links every pair of variables that it perceives as causes of some third variable. This result identifies neglect of direct correlation between perceived causes as the kind of causal misperception that can generate 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, Rafaella Giacomini, Todd Sarver, Vasiliki Skreta, Dimitri Vayanos, Michael Woodford, numerous seminar participants, the editor and two referees, 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

Many economic models assume that outcomes depend on some agents' estimates or predictions of particular variables. For instance, a manager's career depends on outside observers' estimate of his "quality". Similarly, in models of financial markets, speculative trade depends on whether traders predict positive expected monetary gains. Finally, in models of monetary policy, the central bank's policy can positively affect real variables to the extent that the private sector underestimates inflation.

In conventional models, an agent's estimates and 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 prediction errors cancel out on average, such that the average of the agent's predictions coincides with the average of the predicted variables. 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 depart from rational expectations in many ways; this paper focuses on the role of *causal perceptions* 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 preconception regarding the sign or magnitude of this effect). The agent fits his causal model to objective data obtained from a steady-state distribution, 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.

The agent's DAG has at least two interpretations. First, it can represent a lay person's intuitive causal perceptions, or a *narrative* that he employs to create an 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 multivariate probability distributions is cognitively demanding. One cannot perceive them directly as a whole, and instead must settle for learning several correlations among a relatively small number of variables. Thinking in terms

[1] Consider 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).

of a causal model - whether intuitive or formal, and especially if it can be represented by a *sparse* DAG - simplifies this task. The model alerts the agent to specific correlations and puts them together in narrative form, which makes it easier for him to grasp the system as a whole. Once the agent has quantified his causal model, he can use it to make any conditional prediction that is required by whatever the task he is facing.

The question that I study in this paper is whether an incorrect wrong causal model will nevertheless have properties that produce conditional predictions of individual variables that are correct *on average.* When these properties fail to hold, I analyze (in the context of specific applications) the extent to which an outside party will take advantage of the agent's belief biases. The following example illustrates the formalization of this question and its economic motivation.

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.

Consider the following simple version of this class of models, which I mostly borrow from Sargent (1999). A central bank chooses an action a that affects inflation π. The private sector forms an 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, such that only unanticipated inflation has real effects. The central bank's utility function is $y - \pi$ - i.e., it wants higher output and lower inflation.

Suppose that this system is in a steady state that is described by an *objective distribution* p over all variables a, π, e, y. If the private sector had rational expectations, e would be equal to the expected value of π conditional on the observed realization of a, according to p. As a result, expected output would be zero, independently of the central bank's strategy. In this case, the central bank's ex-ante optimal strategy would be to choose an action that minimizes expected inflation.

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

$$a \rightarrow \pi \leftarrow y \tag{1}$$

According to this causal model, inflation is a consequence of two causes: output and the central bank's action. The model is wrong because it perceives output to be independent of monetary policy whereas according to the true process, output is an indirect consequence of the central bank's action via the Phillips Curve. In particular, (1) reverses the direction of causality between output and inflation: it regards output as a cause of inflation, whereas according to the true model, inflation is among the causes of output. In addition, (1) excludes e - i.e., it does not recognize the role of private-sector expectations in the determination of macroeconomic variables (as did monetary theory prior to the seminal contributions of Phelps (1967) and Friedman (1968)). Thus, the private sector's causal model tells a "classical" story that postulates the *absolute neutrality of monetary policy*, as it contains *no causal path* from a to y. In contrast, the true model allows a to have an indirect causal effect on y, via inflationary surprises.

How does the private sector employ its causal model to forecast inflation? It simply *fits* the model to the true joint distribution p over a, π, y, according to the following formula:

$$p_R(a, \pi, y) = p(a)p(y)p(\pi \mid a, y) \tag{2}$$

The formula $p_R(a, \pi, y)$ describes the private sector's subjective belief as a function of the true distribution p. If p *were* consistent with R, it would be legitimate to write it in this form.

Expression (2) is an example of a "*Bayesian-network factorization for-*

mula", which factorizes p over a, π, y into a product of conditional-probability terms, *as if* p were indeed consistent with R. The terms on the R.H.S of (2) can be viewed as outcomes of specific correlation measurements that a forecaster makes in order to quantify 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 true distribution p. In particular, a correct way of factorizing p in this example is given by the textbook chain rule

$$p(a, \pi, y) = p(a)p(y \mid a)p(\pi \mid a, y)$$

The private sector's inflation forecast after observing the central bank's action a is[2]

$$E_R(\pi \mid a) = \sum_{\pi} p_R(\pi \mid a)\pi = \sum_{\pi} \sum_{y} p(y)p(\pi \mid a, y)\pi \tag{3}$$

Because the steady-state distribution p is affected by private-sector expectations, it is an "*equilibrium*" distribution; the equilibrium requirement is that $e = E_R(\pi \mid a)$ with probability one, for every a.

Note that $E_R(\pi \mid a)$ is in general different from the rational-expectations inflation forecast

$$E(\pi \mid a) = \sum_{\pi} p(\pi \mid a)\pi = \sum_{\pi} \sum_{y} p(y \mid a)p(\pi \mid a, y)\pi$$

The discrepancy arises because $p_R(\pi \mid a)$ involves an implicit summation over y *without* full conditioning on a. Moreover, the term $p(y)$ in (3) is not independent of a. As a result, a change in the central bank's strategy can lead to a change in $E_R(\pi \mid a)$. This dependency is what makes the central bank's problem interesting analytically.

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

Plugging (3) into the Phillips Curve, we obtain that expected output is

$$\sum_a p(a) \left[E_R(\pi \mid a) - E(\pi \mid a) \right] = \sum_a p(a) E_R(\pi \mid a) - E(\pi) = E(e) - E(\pi)$$

Thus, because the Phillips Curve is linear, expected output depends on whether the private sector's average inflation forecast deviates from average inflation - in other words, on whether the private sector's inflation forecasts are "*systematically biased*". The private sector's inflation forecast may depart from $E(\pi \mid a)$ for given realizations of a, and yet these errors will not affect ex-ante expected output if they cancel each other out on average. Thus, the central bank's ability to influence real activity depends on whether the private sector's causal model generates forecasts with a systematic bias.

In Section 4.1, I present a simple specification of $p(\pi \mid a)$ for which the central bank has a strategy that leads the private sector to systematically underestimate inflation - i.e., $\sum_a p(a) E_R(\pi \mid a) < \sum_\pi p(\pi)\pi$ - and therefore boosts expected output. My main task will be to find restrictions on R or the domain of p that would make such "systematic fooling" impossible in general. And when fooling *is* possible, I will use the formalism to examine its limits in specific settings.

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, and this produces a subjective distribution p_R over the variables that his model admits. The agent derives his conditional estimates of individual variables from p_R.

Can such an agent be systematically fooled, in the sense that *his conditional estimate of some individual variable deviates in expectation from the expected value of this variable?* Of course, this is only one aspect of how causal misperceptions distort decision makers' understanding of statistical regularities. However, it naturally comes up in economic applications where an agent's payoff is *linear* in his or some other agent's beliefs. The above "monetary policy" example is a case in point. Other examples that I consider in this paper include reputation (or career-concern) models and speculative trade among risk-neutral traders. Given the ubiquity of this linearity property in the economics literature, focusing on *average* estimates is of interest - especially if the characterization of the causal models that prevent agents from being systematically fooled in the above sense happens to be simple yet non-trivial.

The first main result, given in Section 3.1, provides such a characterization: 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 average forecast of *some* variable (in this case, inflation) is biased. In contrast, the DAG $a \rightarrow \pi \rightarrow y$ is perfect, and therefore cannot give rise to systematically biased predictions.

Perfection is a familiar property in the Bayesian-networks literature. Its significance in the present context is that it highlights the role of a particular form of correlation neglect in generating systematically biased estimates. Any DAG that omits a direct link between two variables captures some neglect of their correlation. However, not every type of correlation neglect can lead to average prediction errors; the main result identifies *neglect of direct correlation between perceived causes* as the potential source of systematically biased estimates. Indeed, in Section 3.2 I provide a graphical characterization of the causal models that can generate biased estimates of a *given* variable, which is based on this structural property, and explains why inflation forecasts in Example 1.1 can be biased, whereas output forecasts cannot.

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 wrong 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 of individual variables is economically relevant. I first provide a thorough analysis of the "monetary policy" example. Then, I present a simple example of a firm that considers the use of sponsored reviews to enhance its reputation among consumers. 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 average estimates are unbiased, *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 $x_0, x_1, ..., x_n$ 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 $M \subseteq \{0, 1, ..., n\}$, 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, $i \in N - \{0\}$, where $N \subseteq \{1, ..., n\}$ is some subset of the indices (or labels) of the economic variables. In some applications, I refer to e_i as the agent's forecast of x_i.

Let p be an objective joint distribution over all economic variables $x_0, ..., x_n$ as well as the estimate variables $(e_i)_{i \in N - \{0\}}$. This distribution represents steady-state statistical regularities in the agent's environment. I will later impose the condition that the e_i's are consistent with a specific model of belief formation. In particular, if they are based on rational expectations, then p must satisfy the restriction that for every $i \in N - \{0\}$, $p(e_i \mid x_0)$ assigns probability one to $E(x_i \mid x_0) = \sum_{x_i} p(x_i \mid x_0) x_i$.[3] (The reason I define p

[3] Throughout the paper, E without a subscript means expectation w.r.t the objective

over e_i's as well as x_i's is that in some applications (e.g. Example 1.1), the agent's beliefs affect the realization of economic variables. However, this is not necessary for the general analysis of Section 3, where we can afford to define p over the x_i's only.)

Our agent is characterized by a *directed acyclic graph* (DAG) (N, R), where $N \subseteq \{0, ..., n\}$ is the set of nodes and R is the set of directed links. I use jRi or $j \rightarrow i$ interchangeably to denote a directed link from j into i. Acyclicity means that the binary relation R is acyclic - i.e., the graph contains no directed path from a node to itself. Abusing notation, let $R(i) = \{j \in N \mid jRi\}$ be the set of "parents" of node i. In another abuse of notation, I will usually suppress N and refer to R itself as the agent's DAG, unless explicit reference to the set of nodes is important for intelligibility.

Following Pearl (2009), I interpret the DAG 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. 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 $e_1, ..., e_n$ - they are not represented by nodes in the DAG. I will provide a formal justification for the latter restriction in Section 5.3.

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

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

Thus, R encodes a mapping that transforms every objective distribution p into a subjective belief p_R. Marginalization and conditioning of p_R are defined as usual. For every $M \subset N$, the subjective marginal distribution over x_M is $p_R(x_M) = \sum_{x_{N-M}} p_R(x_M, x_{N-M})$. The agent's subjective distribution over x_i conditional on his observation of x_0 is $p_R(x_i \mid x_0) = p_R(x_0, x_i)/p_R(x_0)$.

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

distribution p.

structure of p. When R is fully connected, (4) is reduced to the textbook chain rule, such that $p_R(x_N) \equiv p(x_N)$. Thus, every objective distribution is consistent with a fully connected DAG. In other words, a fully connected DAG induces rational expectations.

In the general analysis, I impose the following domain restrictions on p.

Condition 1 *(i) The objective distribution p has full support over X_N, such that all the conditional probabilities in (4) are well-defined. (ii) For every x_0 and $i \in N - \{0\}$, $p(e_i \mid x_0)$ assigns probability one to*

$$E_R(x_i \mid x_0) = \sum_{x_i} p_R(x_i \mid x_0) x_i \tag{5}$$

In applications, I will sometimes be able to relax condition (i). Note that condition (ii) implies that the objective expectation of the agent's estimate of the variable x_i is

$$E(e_i) = \sum_{x_0} p(x_0) E_R(x_i \mid x_0) \tag{6}$$

The notations E_R and $E(e_i)$ do not explicitly invoke the objective distribution p. Whenever I use them (as well as the notation $E(x_i)$), the objective distribution to which they relate will be clear from the context.

The formula p_R describes how the agent employs his subjective causal model to form beliefs. I have in mind two more specific interpretations of this belief formation process. 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, and then interprets them causally in accordance with his subjective 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.

Alternatively, we can think of the agent as a *professional forecaster*, who has an *explicit formal model* of the economic environment; he fits the model to the steady-state distribution, and uses this "estimated model" to form forecasts of specific variables upon request. The forecaster's model 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. As a result, estimating the equation for x_i produces the true conditional distribution $p(x_i \mid x_{R(i)})$. It is as if the forecaster tweaks the equation's functional form until he gets perfect empirical fit, but he does not tamper with the equation's R.H.S variables - 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.

Although I have fixed x_0 as the variable that the agent gets to observe, this is merely an expositional device that is not needed for the general results. We should think of the agent as potentially facing many situations that involve the economic variables $x_1, ..., x_n$; every situation requires the agent to predict some variable x_i as a function of some other variable x_j, and these two variables vary across situations. (A subsequent extension of the basic model assumes that the agent conditions his prediction on *multiple* variables.) Grounding each of these numerous conditional predictions in a direct measurement of some conditional probability would be very costly. The "estimated model" p_R simplifies this task: it requires the agent to make a relatively small number of direct measurements once and for all, and enables him to draw on p_R whenever a situation calls for making a specific conditional prediction.[4]

We are now ready for the paper's central definitions.

[4] There 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.

Definition 1 (Unbiased estimate of a specific variable) *A DAG (N, R) induces an unbiased estimate e_i for some $i \in N - \{0\}$ if $E(e_i) = E(x_i)$ for every objective distribution p that satisfies condition 1.*

Definition 2 (Universally unbiased estimates) *A DAG (N, R) induces universally unbiased estimates if it induces an unbiased estimate e_i for every $i \in N - \{0\}$.*

Definitions 1 and 2 allow the agent to form estimates that depart from the rational-expectations benchmark - i.e., $E_R(x_i \mid x_0) \neq E(x_i \mid x_0)$ for some x_0. However, the errors even out when we integrate over x_0. The simplest example of a wrong DAG that induces universally unbiased estimates is an empty DAG ($R(i) = \varnothing$ for every $i \in N$). This DAG fails to capture correlations, because it satisfies $p_R(x_i, x_j) = p(x_i)p(x_j)$ for every i, j. However, this identity implies $p_R(x_i \mid x_0) \equiv p(x_i)$, hence $E(e_i) = E(x_i)$. My goal in the next section will be to characterize the class of DAGs that share the latter property.

3 General Results

I begin this section with a few graph-theoretic concepts and results, most of which are borrowed from the Bayesian-networks literature (though often with different notation and terminology).

Fix a DAG (N, R). The DAG's *skeleton* $(N, \tilde{R})$ is its undirected version - i.e., $i\tilde{R}j$ if and only if iRj or jRi. A subset $M \subseteq N$ is a *clique* in (N, R) if $i\tilde{R}j$ for every $i, j \in M$, $i \neq j$. 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 *collider* is an ordered triple of nodes (i, j, k) such that iRk and jRk. A collider (i, j, k) is referred to as a *v-collider* if $i\not{R}j$ 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). For instance, the DAG (1) contains a v-collider $y \rightarrow \pi \leftarrow a$. When discussing a v-collider, I will use the notations (i, j, k) and $i \rightarrow k \leftarrow j$ interchangeably.

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 3 *Two DAGs* (N, R) *and* (N, Q) *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 $p(x_1)p(x_2 \mid x_1) \equiv p(x_2)p(x_1 \mid x_2)$. A DAG that involves intuitive causal relations can be equivalent to a DAG that makes little sense as a causal model.

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

To illustrate this result, the DAGs $1 \rightarrow 2 \rightarrow 3$ and $1 \leftarrow 2 \leftarrow 3$ are equivalent because they have the same skeleton and an empty set of v-colliders. 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.

3.1 Universally Unbiased Estimates

I now turn to a characterization of the DAGs that induce universally unbiased estimates. We will see later that this characterization is a simple corollary of the characterization of DAGs that induce unbiased estimates of a *given* variable, which I provide in the next sub-section. Therefore, from a logical point of view, the order of the two sub-sections should have been reversed. However, presenting the universal case first is superior in terms of expositional simplicity, even if it leads to some redundancy in presentation.

The following lemma (borrowed from Spiegler (2016b)) begins the analysis of how graphical properties of the agent's causal model affect the structure of his belief distortions.

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.*

This lemma establishes that if C is an ancestral clique (in the DAG itself or in some equivalent DAG), then the subjective marginal distribution over x_C is always correct. Otherwise, we can find an objective distribution for which it will be distorted.

Definition 4 *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.

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 (N, R), then M is an ancestral clique in some DAG in the equivalence class of (N, R).*

Corollary 1 is an immediate implication of Proposition 1. It means that the causal links postulated by a perfect DAG are unidentified from observational data: if iRj, there exists a DAG R' that is equivalent to R, such that $jR'i$. A DAG contains causal links with observationally meaningful direction only when it contains a v-collider.

The following result is an immediate combination of Corollary 1 and Lemma 1.

Corollary 2 *In a perfect DAG R, $p_R(x_C) \equiv p(x_C)$ for any clique C.*

In particular, the agent's subjective marginal distribution over any variable coincides with its objective marginal distribution. In other words, an agent with a perfect DAG never distorts individual variables' marginal distributions. In contrast, an imperfect DAG induces incorrect subjective marginals for some variables and some objective distribution. If we viewed unbiased marginals as a desirable property that a plausible model of non-rational expectations "must" satisfy, then we would have to restrict attention to causal models that are represented by perfect DAGs. However, I do not share this view. As explained in the previous sections, I think of the agent's subjective model as a tool for making multiple conditional predictions in many possible choice contexts. The aspects of the objective distribution that the agent correctly perceives are the marginal and conditional distributions that quantify his causal model. Any other marginal or conditional subjective distribution is derived from his estimated model, and there is no a priori reason why it must be correct.

Proposition 2 *A DAG induces universally unbiased estimates if and only if it is perfect.*

Thus, as long as the agent's DAG is perfect, he cannot be systematically fooled about any variable. 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. As mentioned earlier, perfect DAGs have the property that the causal links they postulate are unidentified from observational data. Proposition 2 thus implies that the agent's causal misperceptions expose him to systematic fooling if and only if the direction of some of them is meaningful for observational data.

The sufficiency part of Proposition 2 has a three-line proof, thanks to the above preliminary results.

Proof of sufficiency part of Proposition 2. Suppose the DAG (N, R)

is perfect. By Corollary 2, $p_R(x_j) \equiv p(x_j)$ for every $j \in N$. Therefore,

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

which immediate implies $E(e_i) = E(x_i)$. ■

Thus, the key property of perfect DAGs that ensures universally unbiased estimates is that they induce correct marginals over all individual variables - including the conditioned variable x_0 and any estimated variable x_i.

In contrast, consider an imperfect DAG like $0 \rightarrow 2 \leftarrow 1$. Its neglect of the potential correlation between x_0 and x_1 can lead to a biased marginal subjective distribution over x_2. This in turn implies that the agent's average estimate of x_2 can be biased. The complete proof of the necessity part will be presented as a simple corollary of the characterization result in the next sub-section.

An immediate corollary of Proposition 2 is that universal unbiasedness is not monotone with respect to the thickness of the agent's DAG. When we add links to a DAG (e.g., from $0 \rightarrow 2 \quad 1$ to $0 \rightarrow 2 \leftarrow 1$), we may destroy perfection and therefore create a vulnerability to systematic fooling. Non-monotonicity results in this spirit appear in Eyster and Piccione (2013) and Spiegler (2016a).

Within the literature on equilibrium models with non-rational expectations, perfection emerges naturally in certain formulations of analogy-based expectations (Jehiel (2005)). For example, consider the DAG $s \leftarrow \theta \rightarrow \pi \rightarrow a$, where s represents the agent's signal, θ represents the state of Nature, π represents the analogy class to which θ belongs and a represents an opponent's action. Under this DAG, the agent perceives the opponent's behavior as a consequence of the analogy partition, whereas in fact it is a function of the state of Nature. As a result, the agent's conditional forecast of the opponent's action is overly coarse. Nevertheless, Proposition 2 implies that it is correct on average.

A recent explicit application of perfect DAGs is Schumacher and Thysen (2017), who study a principal-agent model in which the agent has a wrong causal model of the mapping from effort decisions to output. In particular, the agent's causal model is given by the perfect DAG $a \rightarrow x \rightarrow y$ (a, y and x represent effort, output and an intermediate outcome, respectively). The true process involves additional causal paths from a to y, which the agent's causal model neglects.

3.2 Unbiased Estimates of a Specific Variable

In this sub-section I hold the estimated variable fixed, and provide a necessary and sufficient condition on the agent's DAG under which it induces unbiased estimates of this variable. For this purpose, additional graphical definitions and notation are needed. Fix a DAG (N, R). A *path* is a sequence of directly linked nodes, ignoring the links' directions. For example, in the DAG $1 \rightarrow 2 \leftarrow 3$, there is a path between 1 and 3, even though there is no *directed* path between them. Define the binary relation P: for any $i, j \in N$, iPj if there is a directed path from i to j or if $i = j$. For example, in the DAG $1 \rightarrow 2 \rightarrow 3$, $1 \not{P} 3$ and $1 \not{P} 1$ yet $1P3$ and $1P1$. When iPj and $i \neq j$, we say that i is an *ancestor* of j and j is a *descendant* of i.

d-separation

One of the basic ideas in the Bayesian-network literature is that a DAG represents a collection of conditional-independence assumptions (common to all probability distributions that are consistent with a DAG, and violated by the distributions that are inconsistent with it). The concept of d-separation operationalizes this idea. It appears in any textbook on the subject (e.g. Pearl (2009)). I now present a version of d-separation that is specialized to our current needs. First, I introduce the notion of path blocking that is standard in the literature.

Definition 5 (Path blocking) *The node* 0 ***blocks*** *a path in the DAG if either of the following two conditions holds:* (1) *the path contains a segment of the form $k \rightarrow 0 \rightarrow j$ or $k \leftarrow 0 \rightarrow j$, for some nodes k, j;* (2) *the path contains a segment of the form $k \rightarrow m \leftarrow j$ for some nodes k, j, such that neither m nor its descendants are* 0*.*

To illustrate this definition, consider the DAG $1 \to 2 \leftarrow 0 \to 3$. The node 0 blocks the path between 1 and 3 - either because it contains the segment $2 \leftarrow 0 \to 3$ (thus satisfying condition (1)) in the definition), or because it contains the segment $1 \to 2 \leftarrow 0$ and 2 has no descendants (thus satisfying condition (2) in the definition). In contrast, in the DAG $1 \to 0 \leftarrow 2 \to 3$, 0 does not block the path between 1 and 3, because it does not contain a segment of the form $i \to 0 \to j$ or $i \leftarrow 0 \to j$, and the only segment of the form $i \to m \leftarrow j$ that it contains satisfies $m = 0$.

In what follows, let A and B be two disjoint subsets of $N - \{0\}$. In addition, write $x_A \perp_R x_B \mid x_0$ if every distribution that is consistent with R satisfies the conditional-independence property $x_A \perp x_B \mid x_0$.

Definition 6 (*d*-separation) *The node* 0 *d-separates* A *and* B *if* 0 *blocks every path between any* $j \in A$ *and any* $k \in B$.

Proposition 3 (Verma and Pearl (1990)) $x_A \perp_R x_B \mid x_0$ *if and only if* 0 *d-separates* A *and* B.

Thus, *d*-separation provides a convenient (and computationally simple) graphical rule for checking whether a conditional independence property is satisfied by all the distributions that are consistent with a DAG, hence by all subjective beliefs that can be induced by a given causal model.

The following simplification will be useful for the results of this subsection. It can be immediately seen from (4) that the only variables that are relevant for $p_R(x_0, x_i)$ - and therefore for $p_R(x_i \mid x_0)$ - are those represented by nodes j for which jPi or $jP0$. All other nodes can be ignored. Since this will simplify definitions and notation without affecting the analysis, in what follows, I assume that the DAG (N, R) satisfies the following:

$$\{j \in N \mid jPi \text{ or } jP0\} = N \tag{7}$$

In other words, any terminal node in the DAG must be 0 or i.

Proposition 4 *A DAG induces an unbiased estimate e_i if and only if for every v-collider $j \rightarrow k \leftarrow h$ in the DAG, 0 d-separates $\{i\}$ and $\{j, h\}$.*

This result traces the possibility of a biased estimate e_i to the existence of a v-collider in the agent's DAG - specifically, a v-collider whose upper nodes are ancestors of i or 0.[5] However, if 0 d-separates the node i from the v-collider's upper nodes, e_i will be unbiased. What is the meaning of the role of v-colliders in this characterization? Every DAG that is not fully connected distorts some objective distributions by neglecting certain correlations. However, not every type of correlation neglect leads to systematically biased estimates. Proposition 4 identifies *neglect of direct correlation between perceived causes* as the reason for biased estimates. However, if every perceived causal chain from these variables to the predicted variable x_i passes through the observed variable x_0, then by conditioning his estimate e_i on the observed x_0 protects him from the bias that can potentially result from the above correlation neglect.

The complete proof of Proposition 4 is in the Appendix. Here I will settle for an example that illustrates the result. Suppose that the agent's DAG R is

$$\begin{tikzcd} j \arrow[r] & 0 \arrow[d] & h \arrow[l] \arrow[dl] \\ & i & \end{tikzcd} \tag{8}$$

Here there is a direct link from one of the v-collider's upper nodes into i, hence 0 does not block it.

To see how this feature can generate a biased estimate e_i, construct the following objective distribution p. Suppose that all variables take values in $\{0, 1\}$. Let $p(x_j = 1) = \frac{1}{2}$, and assume that the other variables are given the by the following sequence of deterministic equations:[6]

$$x_h = x_j \qquad x_0 = 1 - (1 - x_j)(1 - x_h) \qquad x_i = x_0 x_h$$

[5] If we did not impose the simplifying assumption (7), the d-separation condition would only pertain to v-colliders $j \rightarrow k \leftarrow h$ in which $i'Pi$ or $i'P0$ for some $i' \in \{j, h\}$.

[6] This specification of p violates the full-support assumption. This is purely for expositional simplicity - small perturbations of p that restore this property would leave the argument intact.

Observe that the only feature of p that is inconsistent with R is the (perfect) correlation between x_j and x_h, whereas R assumes they are independent.

Under p, $E(x_i) = p(x_i = 1) = \frac{1}{2}$. Let us now calculate

$$E(e_i) = \sum\nolimits_{x_0} p(x_0) p_R(x_i = 1 \mid x_0)$$

where $p_R(x_i = 1 \mid x_0)$ is given by

$$\frac{p_R(x_0, x_i = 1)}{p_R(x_0)} = \frac{\sum_{x_j, x_h} p(x_j)p(x_h)p(x_0 \mid x_j, x_h)p(x_i = 1 \mid x_h, x_0)}{\sum_{x_j, x_h} p(x_j)p(x_h)p(x_0 \mid x_j, x_h)}$$

Note that $p(x_j = 1) = p(x_h = 1) = \frac{1}{2}$. All the other conditional-probability terms are 0 or 1, as given by our equations for x_0 and x_i. We obtain $p_R(x_i = 1 \mid x_0 = 1) = \frac{2}{3}$ and $p_R(x_i = 1 \mid x_0 = 0) = 0$, leading to

$$E(e_i) = p(x_0 = 1) \cdot \frac{2}{3} = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3} \neq \frac{1}{2} = E(x_i)$$

The calculation makes it clear that the reason for the bias is the neglect of the correlation between x_j and x_h.

Now suppose that the agent's DAG omitted the link $h \to i$ from (8). Then, $p_R(x_i \mid x_0)$ would be given by

$$\frac{p_R(x_0, x_i)}{p_R(x_0)} = \frac{\sum_{x_j, x_h} p(x_j)p(x_h)p(x_0 \mid x_j, x_h)p(x_i \mid x_0)}{\sum_{x_j, x_h} p(x_j)p(x_h)p(x_0 \mid x_j, x_h)} = p(x_i \mid x_0)$$

which implies that e_i is unbiased.[7]

Finally, we are able to derive the necessity part of Proposition 2 from Proposition 4.

Proof of necessity part of Proposition 2. Suppose the DAG (N, R) is imperfect. Then, it contains a v-collider $j \to k \leftarrow h$. There are two cases

[7] In this example, the agent's estimate coincides with rational expectations for *every* x_0, but this is only because of the direct link $0 \to i$. If we replaced this link with the segment $0 \to i' \to i$, e_i would depart from rational expectations but it would be unbiased on average.

to consider. First, suppose that $0 = k$. Then, 0 does not block the only path between h and j. As a result, the condition for unbiased e_h (or e_j) is violated. Second, suppose that $0 \neq k$. Obviously, $0 \neq j$ or $0 \neq h$. Assume the former case, w.l.o.g. Then, the only path between k and j is a direct link, which 0 obviously does not block. As a result, the condition for unbiased e_k is violated. ■

Recall that perfect DAGs induce an unbiased estimate of any x_i because they induce correct marginal subjective distributions over both x_0 and x_i. However, the latter property is not necessary. For instance, consider the following DAG:

$$\begin{tikzcd}
1 \arrow[r] & 0 \arrow[r] & 3 \arrow[d] \\
& 2 \arrow[u] & 4
\end{tikzcd}$$

This modified DAG can induce incorrect marginals over x_0, x_3 and x_4. Nevertheless, e_3 and e_4 are unbiased because the node 0 blocks the only causal path from the DAG's v-collider to the nodes 3 and 4.

4 Two Applications

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

4.1 Monetary Policy

In this sub-section I analyze a more elaborate version of Example 1.1. Recall the three economic variables: the central bank's action a, inflation π and real output y. Now introduce a fourth variable θ that the central bank *privately* observes before taking its action. Suppose that θ is real-valued and distributed over some finite subset of $(0, 1)$ - the exact distribution will be immaterial for our purposes. The central bank's utility function is $y - \theta\pi$. Thus, θ measures the central bank's trade-off between the two motives, but does not have any direct effect on macroeconomic variables.

Assume that both π and a take values in $\{0, 1\}$, where $\pi = 0$ (1) represents low (high) inflation. The central bank's strategy is thus 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.

The objective steady-state distribution p is consistent with the following "true DAG" R^* defined over θ, a, π, y:[8]

$$\begin{tikzcd}
\theta \arrow[r] & a \arrow[dl] \arrow[d] \\
\pi \arrow[r] & y
\end{tikzcd} \tag{9}$$

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

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

If the private sector had rational expectations ($R = R^*$ or fully connected), this expression would collapse into

$$-\sum_\theta p(\theta) E(\pi \mid \theta)\theta = -\beta \sum_\theta p(\theta) p(a = 1 \mid \theta)\theta$$

and the central bank's ex-ante optimal strategy would be $p(a = 1 \mid \theta) = 0$ for every $\theta > 0$ (I discuss the issue of dynamic consistency at the end of this sub-section).

[8] If we incorporated e as an explicit variable in the causal model, the direct link $a \to y$ would be replaced with the chain $a \to e \to y$. See Section 5.3 for a discussion of estimates as variables in DAGs.

Consider two possibilities for the private sector's DAG. Each DAG represents a different narrative about how macro variables are interconnected.[9] First, consider the DAG $\theta \rightarrow a \rightarrow y \rightarrow \pi$. Because it is perfect, Proposition 2 implies that inflation forecasts would be unbiased. As a result, the central bank's ex-ante optimal policy coincides with the rational-expectations prediction.

Now turn to the DAG R given by

$$\begin{tikzcd}
\theta \arrow[r] \arrow[d] & a \arrow[d] \\
y \arrow[r] & \pi
\end{tikzcd} \tag{10}$$

As in Example 1.1, this DAG reflects a "classical" belief in absolute monetary neutrality - namely, that a has no causal effect on y. The private sector's causal model allows a and y to be correlated, but only to the extent that both are caused by the exogenous variable θ.

Given this DAG R, the private sector's inflation forecast after observing a is

$$E_R(\pi \mid a) = \sum_{\pi} p_R(\pi \mid a)\pi = \sum_{\pi} \left(\sum_{y} \sum_{\theta} p(\theta \mid a) p(y \mid \theta) \right) p(\pi \mid a, y)\pi$$

whereas the "rational" conditional inflation forecast is

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

The discrepancy arises because $p_R(\pi \mid a)$ only acknowledges the correlation between a and y through their mutual correlation with θ.

The DAG given by (10) is imperfect. Moreover, it violates the condition in Proposition 4 because it contains a v-collider $y \rightarrow \pi \leftarrow a$, where there is a direct link between π and y (an upper node in the v-collider), which therefore cannot be blocked by a. (On the other hand, the DAG satisfies the condi-

[9] Hoover (2001) describes historical controversies in macroeconomics in terms of conflicting causal mechanisms.

tion with respect to *output* forecasts, because the only v-collider involves a *descendant* of y and a.) Therefore, we cannot rule out the possibility that the private sector's inflation forecasts will be systematically biased. Indeed, the following result establishes that under the current parameterization, the central bank's ex-ante optimal strategy involves inflating and the private sector systematically underestimates inflation (as the Phillips-Curve noise vanishes).

Proposition 5 *In the $\sigma_\eta^2 \to 0$ 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. Expected output in this limit is $(1-\theta^2)/4$ for every θ.*[10]

Proof. The following notation will be helpful in the proof:

$$\begin{aligned} e(a) &= E_R(\pi \mid a) = p_R(\pi = 1 \mid a) \\ \alpha_\theta &= p(a=1 \mid \theta) \\ \alpha &= \sum_\theta p(\theta)\alpha_\theta \end{aligned}$$

Because $\pi \in \{0,1\}$ and by the specification of R,

$$e(a) = \sum_\theta p(\theta \mid a) \sum_y p(y \mid \theta) p(\pi = 1 \mid y, a)$$

Let us first calculate $e(0)$. Because $p(\pi = 1 \mid a = 0) = 0$, it follows that $p(\pi = 1 \mid a = 0, y) = 0$ for all y. Therefore, $e(0) = 0$. This in turn means that $E(y \mid a = 0) = 0$, which means that the central bank's expected payoff can be reduced to

$$\sum_\theta p(\theta)\alpha_\theta[\beta - e(1) - \theta\beta] \tag{11}$$

[10] Because $p(\pi, y \mid a)$ and $p(e \mid a)$ are jointly determined, the central bank's problem is not a straightforward maximization problem in which the only object of choice is its own strategy. Rather, there could also be a need to select among multiple possible private-sector expectations (just as we apply equilibrium selection in mechanism design problems). However, the proof of the following result makes it clear that this issue does not arise in this example.

If $\alpha_\theta = 0$ for all θ, the central bank's payoff is zero. From now on, assume $\alpha_\theta > 0$ for some θ such that $\alpha > 0$.

Let us now calculate $e(1)$. Because $\alpha > 0$ and η is normally distributed (such that $p(y \mid \theta)$ has full support), the terms in the expression for $e(1)$ are all well-defined. For any fixed θ, $y \sim N(\mu, \sigma_\eta^2)$, where μ is random: $\mu = e(0) = 0$ with probability $1 - \alpha_\theta$, $\mu = 1 - e(1)$ with probability $\alpha_\theta \beta$, and $\mu = -e(1)$ with probability $\alpha_\theta(1 - \beta)$. By definition, $e(1) \in [0, 1]$. Let us verify that $e(1)$ is interior and does not converge to 0 or 1 in the $\sigma_\eta^2 \to 0$ limit, such that the above three values that μ can get are all distinct. Because the normal distribution is symmetrically distributed around its mean, the probability of $y < -e(1)$ given θ cannot be lower than $\alpha_\theta(1-\beta)/2$, whereas the probability of $y > 1 - e(1)$ given θ cannot be lower than $\alpha_\theta\beta/2$. Moreover, as $\sigma_\eta^2 \to 0$, $p(\pi = 1 \mid a = 1, y < -e(1)) \to 0$ and $p(\pi = 1 \mid a = 1, y > 1 - e(1)) \to 1$. Therefore, in the $\sigma_\eta^2 \to 0$ limit,

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

It follows that in the $\sigma_\eta^2 \to 0$ limit, the three possible values for μ - namely, $-e(1)$, 0 and $1 - e(1)$ - are all distinct. Moreover, in this limit, $p(\pi = 1 \mid a = 1, y = 1 - e(1)) = p(\pi = 0 \mid a = 1, y = -e(1)) = 1$. By comparison, $p(\pi = 1 \mid a = 1, y = 0)$ is not obvious because the joint realization $a = 1, y = 0$ gets zero probability in the $\sigma_\eta^2 \to 0$ limit. By (11) and the definition of $e(1)$, the central bank's payoff would be higher if we set $p(\pi = 1 \mid a = 1, y = 0) = 0$. Therefore, let us guess that this is the case, and verify this guess later on. It follows that

$$\sum_y p(y \mid \theta) p(\pi = 1 \mid y, a = 1) = \alpha_\theta \beta$$

and

$$e(1) = \beta \sum_{\theta'} p(\theta' \mid a = 1)\alpha_{\theta'} = \beta \sum_{\theta'} \frac{p(\theta')\alpha_{\theta'}}{\sum_{\theta''} p(\theta'')\alpha_{\theta''}} \alpha_{\theta'} \tag{12}$$

Expression (11) then becomes

$$\begin{aligned}
& \beta\left[\sum_{\theta} p(\theta)\alpha_\theta(1-\theta) - \sum_{\theta} p(\theta)\alpha_\theta \sum_{\theta'} \frac{p(\theta')(\alpha_{\theta'})^2}{\sum_{\theta''} p(\theta'')\alpha_{\theta''}}\right] \\
= \; & \beta\left[\sum_{\theta} p(\theta)\alpha_\theta(1-\theta) - \sum_{\theta'} p(\theta')(\alpha_{\theta'})^2 \left(\frac{\sum_{\theta} p(\theta)\alpha_\theta}{\sum_{\theta''} p(\theta'')\alpha_{\theta''}}\right)\right] \\
= \; & \beta \sum_{\theta} p(\theta)\left[\alpha_\theta(1-\theta) - (\alpha_\theta)^2\right] \qquad (13)
\end{aligned}$$

This objective function is additively separable in θ, such that for every θ, the optimal value of α_θ maximizes $\alpha_\theta(1-\theta) - (\alpha_\theta)^2$, which immediately gives the solution.

It remains to verify our guess that $p(\pi = 0 \mid a = 1, y = 0) = 1$ in the $\sigma_\eta^2 \to 0$ limit. Because $\alpha_\theta = \frac{1}{2}(1-\theta) \leq \frac{1}{2}$ for all θ, (12) implies that $e(1) \leq \frac{1}{2}\beta$. It follows that $|-e(1) - 0| \leq \frac{1}{2}\beta < \frac{1}{2}$ whereas $|1 - e(1) - 0| \geq 1 - \frac{1}{2}\beta > \frac{1}{2}$. The conditional distribution $p(y \mid e, a = 1)$ is distributed according to $N(\beta - e, \sigma_\eta^2)$. As we observed above, the only values of y that get positive probability in the $\sigma_\eta^2 \to 0$ limit conditional on $a = 1$ are $-e(1)$ and $1 - e(1)$. Furthermore, $p(\pi = 1 \mid a = 1, y = 1 - e(1)) = p(\pi = 0 \mid a = 1, y = -e(1)) = 1$. Since $-e(1)$ is closer to zero than $1 - e(1)$ in the $\sigma_\eta^2 \to 0$ limit (by a margin that is bounded away from zero in this limit), it follows that $p(\pi = 0 \mid a = 1, y = 0) \to 1$ as $\sigma_\eta^2 \to 0$, thus confirming our guess. ■

The intuition behind the result is as follows. When the central bank plays $a = 0$, it induces $\pi = 0$ with certainty. As a result, $E_R(\pi \mid a = 0) = 0$, as if the private sector had rational expectations. In contrast, when $a = 1$, inflation fluctuates, and the private sector's error is that it tries to account for these fluctuations by the variation in y - as if the latter were only caused by the exogenous variable θ. Therefore, the private sector's inflation forecast conditional on $a = 1$ involves summing over all values of y, weighting them according to the distribution $p_R(y \mid a = 1) = \sum_\theta p(\theta \mid a = 1)p(y \mid \theta)$.

If the central bank plays a deterministic strategy, $p_R(y \mid a = 1) = p(y \mid a = 1)$, such that the private sector's inflation forecast conditional on $a = 1$ is consistent with rational expectations. However, if the central bank employs randomization, it garbles the perceived correlation between a and y, such that the private sector's inflation forecast underreacts to the observed realization $a = 1$. This means that the private sector systematically underestimates inflation after observing $a = 1$. As long as $\theta < 1$, the output boost due to this systematic forecast error outweighs the cost of inflation.

Comment: Dynamic inconsistency

The central bank's optimal strategy is *dynamically inconsistent.* On one hand, playing $a = 0$ with positive probability is necessary for inducing the private sector's systematic prediction error. On the other hand, we saw that the realization $a = 0$ induces an unbiased private-sector inflation forecast, and therefore the central bank would want to switch to $a = 1$ if it could take $E_R(\pi \mid a = 1)$ as given. (In contrast, the central bank's strategy is dynamically consistent with respect to the exogenous variable θ. The proof of Proposition 5 establishes that the central bank's objective function ends up being additively separable in θ, and therefore it would not want to revise its strategy after θ is realized.)

This is a different sort of dynamic inconsistency than the one traditionally associated with this problem since Kydland and Prescott (1977). The latter arises when the private sector does not perfectly monitor the central bank's action before forming its inflation forecast. In contrast, in the present example, the private sector *perfectly monitors* the realization of a. Moreover, it also correctly grasps the central bank's strategy: $p_R(a \mid \theta) = p(a \mid \theta)$. This is a consequence of the fact that θ and a form an ancestral clique in R. The dynamic inconsistency here has an entirely different origin - namely, the need to create statistical patterns that the private sector will misperceive because of its wrong causal model.

Comment: Connection with Example 1.1

Suppose that the central bank's payoff is $y - \pi$, as in Example 1.1. Then, the variable θ is payoff-irrelevant. It can be shown that in this case, the central bank prefers to mix over a independently of θ. The intuition is that the central bank benefits from the private sector's misperception of the correlation between a and y; and when θ ceases to be payoff-relevant, it is optimal for the bank to maximize this misperception by making a and y appear (to the private sector) to be completely independent. As a result, θ becomes independent of all other variables, and therefore the model is effectively reduced to Example 1.1. Plugging $\alpha_\theta = \alpha$ for all θ in (13), we obtain that the central bank's optimal policy is to randomize uniformly over the two actions, independently of θ.

4.2 Manipulating Reputation

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). Although the consumer does not know whether a review is sponsored at the time he reads it, s is *not* an unobservable variable. Data about the historical frequency of sponsored reviews and their correlation with product quality or review content is available to the consumer, as he fits his causal model - given by a DAG R defined over three nodes that represent the variables θ, s, t - to the joint distribution p over these variables.

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. As usual in this paper, $e = E_R(\theta \mid t)$ with probability one for every t. 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 thus

$$\sum_\theta p(\theta) \sum_s p(s \mid \theta) \sum_t p(t \mid \theta, s) [e - cs] = E(e) - cE(s)$$

The relation between the firm's objective and the "systematic fooling" question is apparent from this expression. If the consumer had rational expectations, the firm's objective function would collapse 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 $p(s = 1 \mid \theta) = 0$ for every θ. (Of course, this policy will typically fail to be time-consistent; however, I focus entirely on the *ex-ante* perspective.)

In this example, the consumer's DAG tells a causal story about the process that generates review content. For instance, the DAG $\theta \to t \to s$ represents a "naive" story, according to which content is only influenced by the product's objective characteristics, and sponsorship is reactive (akin to tipping). By comparison, the DAG $\theta \to s \to t$ represents a "cynical" story, according to which content has nothing to do with the product's quality once we condition on the sponsorship status. Both DAGs are perfect, and therefore generate unbiased quality estimates. As a result, the firm's ex-ante optimal strategy under these DAGs coincides with the rational-expectations prediction.

In contrast, the DAG $R : \theta \to t \leftarrow s$ is imperfect. Specifically, it violates the condition of Proposition 4, because t does not block the only path between θ and s. 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. This type of correlation neglect falls into the category that Eyster and Rabin (2005) refer to as "*cursedness*". We will now see that the firm can play a strategy that exploits cursedness to enhance its average reputation.

For this purpose, impose the following additional structure on p. Let $\theta \in \{0, 1\}$; the two values are equally likely, such that $E(\theta) = \frac{1}{2}$. The firm's strategy can thus be represented by two conditional probabilities: $\alpha = p(s = 1 \mid \theta = 1)$ and $\beta = p(s = 1 \mid \theta = 0)$. Finally, $p(t \mid \theta, s)$ is degenerate: $t = \theta + s$ with probability one for every θ, s.

Proposition 6 *Let $R : \theta \to t \leftarrow s$. 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*

$$E(e) = \frac{1}{2} + \frac{1}{16}(1 - 4c^2)$$

Proof. 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):

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

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. ■

Note that the extent to which the firm can exploit the consumer's "cursedness" is limited: it can increase its perceived expected quality by at most $\frac{1}{16}$ on average.

Comment: Why is the consumer's DAG imperfectly connected?
This is a good opportunity to revisit an interpretational issue first mentioned in Section 2. 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 a pedagogical device; its simplicity should not be mistaken for a simplicity of the real-life environment it aims to capture. This environment would typically involve many variables: the quality of numerous types of products, numerous reviewers and various outlets that publish their reviews. It would be hard for consumers to fully understand the intricate web of influences among these variables. Furthermore, the consumer will encounter various situations that require him to make different conditional predictions: guessing whether a given review was sponsored, predicting the content of a review written by one author after seeing a review by another author (not knowing whether they are sponsored and by whom), predicting review content after learning that it was sponsored, etc. A boundedly rational consumer is likely to make simplifying assumptions that assist his attempt to understand statistical regularities in his environment. An example of such a simplifying assumption is that sponsorship is independent of product quality. This particular assumption enables firms to use sponsored reviews to manipulate their average reputation.

5 Extensions

In this section I extend the basic analysis in various directions. Proofs of all the results are in the Appendix.

5.1 Multivariate Normal Distributions

Proposition 2 means that an imperfect DAG exposes the agent to systematically biased estimates 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 the distribution over economic variables is multivariate normal.

Proposition 7 *Let R be an arbitrary DAG, and let the objective distribution over the variables $x_1, ..., x_n$ be multivariate normal. Then, $E(e_i) = E(x_i)$ for every $i = 1, ..., n$.*

Thus, the mere assumption that the agent forms his beliefs by fitting *some* causal model to the steady-state distribution guarantees that he cannot be systematically fooled - as long as the true distribution over economic variables is multivariate normal. The key to this finding is an existing result in the Bayesian-networks literature (see Koller and Friedman (2009, Ch. 7)): factorizing a multivariate normal distribution according to a DAG produces a multivariate normal distribution. Conditional expectations of variables in this class of distributions are simply weighted averages. While a wrong DAG can distort the weights, these distortions cancel out on average.

In each of the applications of Section 4, one of the variables was an *action* taken by some other agent (the central bank in Section 4.1, the firm 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, when the exogenous components of a model are linear-normal, non-linear strategies are necessary for inducing systematically biased predictions.

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, w.l.o.g. However, when the agent's beliefs are based on a wrong 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 $e_i = E_R(x_i \mid x_A)$. As before, we say that R induces universally unbiased estimates if $\sum_{x_A} p(x_A) E_R(x_i \mid x_A) = \sum_{x_i} p(x_i) x_i$ for every objective distribution p in the restricted domain and every $i \in N - A$.

The following result makes use of the following definition. Two sets of nodes $A, B \subset N$ are *mutually disconnected* in (N, R) if for every pair of nodes $i \in A$ and $j \in B$, there is no path in the skeleton $(N, \tilde{R})$ that connects i and j.

Proposition 8 *A perfect DAG induces universally 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 mutually disconnected cliques), the agent's estimates are universally unbiased. When A is empty, the result is reduced to the statement that the agent's ex-ante (unconditional) estimates of individual variables are correct.

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 risk-neutral traders have heterogeneous subjective models, this can lead to belief heterogeneity and thus allow for 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 action s_i. As usual, any objective distribution that is consistent with the game form satisfies $s_i \perp (\theta, t_{-i}, s_{-i}) \mid t_i$ for every trader i - i.e., the trader's action is independent of the state of Nature and other traders' signals and actions, conditional on his signal. Let $z = (z_1, ..., z_m)$ be a zero-sum vector of monetary transfers among traders, which is some stochastic function of θ and $s_1, ..., s_m$. This function 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. Assume that $p(\theta, (t_i)_{i=1,\ldots,m})$ has full support, but we do not need to assume that $p(z \mid \theta, (t_i)_i, (s_i)_i)$ has full support. Trader i's utility function is $u_i(z_i, s_i) = z_i - c \cdot \mathbf{1}(\mathbf{s}_i \neq s^0)$, where $c > 0$ is an arbitrarily small cost of taking a non-default action.

The variables that are allowed to feature in the traders' causal models are θ and $(t_i, s_i, z_i)_{i=1,\ldots,m}$. Assume that trader i's DAG includes at least three nodes that represent the 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. The following are examples of perfect DAGs for trader i that represent incorrect subjective causal models:

$$\theta \rightarrow t_i \rightarrow s_i \rightarrow z_i$$

$$\begin{tikzcd}
\theta \arrow[r] & t_i \arrow[r] \arrow[d] & t_j \arrow[d] \\
 & s_i \arrow[r] \arrow[ur] & z_i
\end{tikzcd}$$

A strategy for trader i is given by the conditional probabilities $(p(s_i \mid t_i))_{t_i,s_i}$. We say that a profile of strategies is an ε-equilibrium if $(p(s_i \mid t_i))_{t_i,s_i}$ has full support for every i and every t_i, and if whenever $p(s_i \mid t_i) > \varepsilon$,

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

That is, if a trader plays an action with probability greater than ε after observing some signal, this action must be a subjective expected-utility maximizer according to his updated subjective belief. The following result is a "no-trade theorem".

Proposition 9 *Suppose that R_i is perfect for every $i = 1, \ldots, m$. Then, for sufficiently small ε, every ε-equilibrium satisfies $p(s_i \mid t_i) \leq \varepsilon$ for every i, t_i and $s_i \neq s^0$.*

The impossibility of biased estimates under perfect DAGs (in which the nodes that represent a trader's signal and his action are linked) 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. The claim is not vacuous: if ε is sufficiently small, we can construct an ε-equilibrium in which every trader plays s^0 with probability $1 - \varepsilon \cdot (|S| - 1)$ and randomizes uniformly over all other actions.

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 could also enter the agent's subjective causal model. Denote $x_{i+n} = e_i$ for every $i = 1, ..., n$, and $x = (x_0, x_1, ..., x_{2n})$. Allow the set of nodes N in the agent's DAG to be a subset of the enlarged set $\{0, 1, ..., 2n\}$. 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 is aware that he conditions his estimates on x_0 alone. Second, it requires that if the agent's DAG includes an estimate of some variable, it must also admit the variable itself. This restriction on R implies the following result.

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

This result means that our original assumption that the agent's DAG omits his own estimates is w.l.o.g, as long as we accept the domain restrictions on p and R.

5.4 Conditional Estimates

Throughout the paper, the question I addressed was whether the agent's estimates of economic variables are unbiased *on average*. I provided three economic applications in which this is all that mattered. However, for many purposes, it also matters whether the agent's conditional estimates are consistent with rational expectations for *all* realizations of a (e.g., the agent may interact with a principal whose payoffs are non-linear in the agent's beliefs). Note that when this stronger requirement holds, the dynamic inconsistency problem discussed in Section 4 disappears. The following is a sufficient condition for the stronger requirement to hold for a given variable.

Claim 1 *Suppose that R is perfect and that $0Ri$ for some node $i \neq 0$. Then, $E_R(x_i \mid x_0) \equiv E_p(x_i \mid x_0)$ for every objective distribution p.*

Proof. By assumption, $\{0, i\}$ is a clique in a perfect DAG. Therefore, we can treat it as ancestral, such that $p_R(x_0, x_i)$ is unbiased. This immediately implies that $p_R(x_i \mid x_0) \equiv p(x_i \mid x_0)$ whenever $p(x_0) > 0$. ■

Imposing this condition on all possible $i \neq 0$ is a very strong requirement, because it means that the agent's causal model regards x_0 as a direct cause of all other variables.

6 Related Literature

This paper continues the research agenda set forth by Spiegler (2016a), where I initiated the use of the Bayesian-network formalism to model decision making under causal misperceptions. Specifically, in that paper I introduced the standard Bayesian–network factorization formula (4) as a representation of belief distortions that arise from fitting subjective causal models to objective data. I showed how the formalism can express familiar errors of causal reasoning (mistaking correlation for causation, reverse causality, omitting confounding variables, attributing outcomes to the wrong cause) and how it can be embedded in a model of individual decision making. I demonstrated that an equilibrium approach may be required to define subjectively optimal decisions under wrong causal models, and used basic tools from the literature on graphical models to characterize the causal misperceptions that call for such an approach.

This paper goes beyond Spiegler (2016a) in two main respects. First, it focuses on the question of whether wrong causal models generate systematically biased estimates of individual variables. Second, it extends the modeling approach from individual choice to interactive principal-agent settings such as those analyzed in Section 4, where the question of "systematic fooling" is a key aspect of the strategic interaction. In particular, the "monetary policy" example offers a prototype for future applications to macroeconomic theory.

More broadly, both Spiegler (2016a) and this paper contribute to the literature on equilibrium models under wrong subjective models. Prominent concepts in the literature include analogy-based expectations equilibrium (Jehiel (2005)), "cursed" equilibrium (Eyster and Rabin (2005)), behavioral equilibrium (Esponda (2008)) and Berk-Nash equilibrium (Esponda and Pouzo (2016)). In relation to the preceding literature, the factorization formula for p_R can be viewed as a class of models of how agents form subjective beliefs that systematically distort objective distributions' correlation structure. Spiegler (2016a) contains a detailed explanation of how the Bayesian-network representation relates to these previous approaches.

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 $t-1, ..., t-k$, 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. According to this interpretation, the agent measures particular correlations because these are the only ones that are available to him. As a result, 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 (1999), 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 examples of *equilibrium* concepts that treat agents as (possibly flawed) statisticians - see Osborne and Rubinstein (1998), 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, perhaps a 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 objective 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 wrong model. This lesson is intriguing, considering the heated historical debate over this question (see Klamer (1984)). The key assumption behind classical non-exploitability results (Lucas (1972), Sargent and Wallace (1975)) was allegedly the private sector's rational expectations, and this was perceived by many as the crux of the matter. The impossibility results of this paper put this historical debate in a new perspective.

Appendix: Proofs

Proposition 4

(**If**). Our objective is to show that if the agent's DAG satisfies the property that the node 0 d-separates i and the upper nodes of any v-collider, then the DAG induces an unbiased estimate e_i. Consider a DAG (N, R) in which any terminal node is either 0 or i. If the DAG has no v-colliders, then it is perfect, and the sufficiency part of Proposition 2 applies, such that the DAG induces unbiased estimates of any variable, including the variable represented by node i.

Now suppose that the DAG has v-colliders. Let $A \subset N$ be the set of upper nodes of these v-colliders as well as their ancestors - i.e., A is the union of all nodes i' such that $i'Pj$ and (j,h,k) is a v-collider for some $k,h \in N$ (recall that the triple (j,h,k) represents the form $j \to k \leftarrow h$ where j and h are not directly linked - that is, j and h are the upper nodes of the v-collider). By assumption, 0 blocks any path between i and the upper nodes of any v-collider. By the definition of path blocking and the assumption that any terminal node must be either i or 0, it must be the case that 0 blocks any path between i and A (in particular, $0, i \notin A$). By Proposition 3, $x_i \perp_R x_A \mid x_0$ - i.e., $p_R(x_i \mid x_A, x_0) \equiv p_R(x_i \mid x_0)$.

Consider the DAG (N,R') that modifies (N,R) as follows: A is a clique under R', and $R'(j) = R(j)$ for every $j \in N - A$. That is, the only difference between (N,R) and (N,R') is that (N,R') fully connects all the upper nodes of v-colliders and their ancestors in (N,R). I will now show that $p_{R'}(x_i \mid x_0) = p_R(x_i \mid x_0)$. First, by construction, A is an ancestral set of nodes in (N,R), in the sense that for every $j \in A$, $R(j) \subseteq A$. It immediately follows from (4) that

$$p_R(x_{N-A} \mid x_A) = \prod\nolimits_{j \in N-A} p(x_j \mid x_{R(j)}) \tag{14}$$

The same holds for (N,R'). And since $R'(j) = R(j)$ for every $j \in N - A$, we have $p_{R'}(x_{N-A} \mid x_A) \equiv p_R(x_{N-A} \mid x_A)$. In particular, this means that $p_{R'}(x_i \mid x_A, x_0) \equiv p_R(x_i \mid x_A, x_0)$. We can now rewrite $p_{R'}(x_i \mid x_0)$ as follows:

$$\begin{aligned} p_{R'}(x_i \quad \mid \quad x_0) &\equiv \sum\nolimits_{x_A} p_{R'}(x_A \mid x_0) p_{R'}(x_i \mid x_A, x_0) \\ &\equiv \sum\nolimits_{x_A} p_{R'}(x_A \mid x_0) p_R(x_i \mid x_A, x_0) \\ &\equiv \sum\nolimits_{x_A} p_{R'}(x_A \mid x_0) p_R(x_i \mid x_0) \equiv p_R(x_i \mid x_0) \end{aligned}$$

By construction, (N,R') is a perfect DAG. Therefore, by Corollary 2, $p_{R'}(x_i) \equiv p(x_i)$ and $p_{R'}(x_0) \equiv p(x_0)$. We can now write

$$\sum_{x_0} p(x_0) p_R(x_i \mid x_0) \equiv \sum_{x_0} p_{R'}(x_0) p_{R'}(x_i \mid x_0) \equiv p_{R'}(x_i) \equiv p(x_i)$$

which implies that (N,R) induces an unbiased estimate e_i.

(**Only if**). Let us first restrict attention to the case in which all variables get two possible values, 0 and 1. At the end of the proof I will explain why this is w.l.o.g. Our objective is to show that if the agent's DAG violates the property that the node 0 d-separates i and the upper nodes of any v-collider, then there is an objective distribution p for which $E(e_i) \neq E(x_i)$. Consider a DAG (N, R) in which any terminal node is either 0 or i. The DAG must contain a v-collider $j \rightarrow k \leftarrow h$. Moreover, for every such v-collider, $kP0$ or kPi. There are two cases to consider.

Case 1: The DAG contains a v-collider $j \rightarrow k \leftarrow h$, such that there is a directed path from k to i that excludes 0 (in particular, $0 \neq k$). Let M be the set of nodes consisting of j, k, h as well as all the nodes along the above directed path from k to i (in case $k \neq i$). Impose the following structure on p. First, for every $i' \notin M$, x_i is independently distributed. This enables us to ignore these variables entirely in our calculations. Second, if $k \neq i$, then for every i' along the directed path from k to i (including i itself), $x_{i'}$ is equal to x_k with arbitrarily high probability. This means that we can perform the calculations as if $i = k$, and this will be an arbitrarily precise approximation. There are now two subcases to consider.

Case 1.1: $0 \neq j, h$. Then, by assumption, $p(x_i \mid x_0) \equiv p(x_i)$ and $p(x_i \mid x_0) \equiv p(x_i)$. We can effectively assume that the DAG is $j \rightarrow i \leftarrow h$, and the objective is to construct p over the three variables x_j, x_h, x_i such that $p_R(x_i = 1) \neq p(x_i = 1)$. Lemma 1 establishes this is possible. (For an explicit example of such a distribution, see the proof of this lemma in Spiegler (2016b).)

Case 1.2: $0 = j$. We can effectively assume that the DAG is $0 \rightarrow i \leftarrow h$, and define p over the three variables x_0, x_h, x_i. Then,

$$p_R(x_i = 1 \mid x_0) = \sum_{x_h} p(x_h) p(x_i = 1 \mid x_0, x_h)$$

Impose the following additional structure on p. First, $p(x_0 = 1) = \frac{1}{2}$. Second, $x_h = x_i = x_0$ with arbitrarily high probability. Third, $p(x_i = 1 \mid x_0 \neq x_h)$ is arbitrarily low. Therefore $p(x_i = 1) \approx \frac{1}{2}$, whereas $\sum_{x_0} p(x_0) p_R(x_i = 1 \mid x_0)$

is equal to

$$\frac{1}{2}\left\{\sum_{x_h} p(x_h)\left[p(x_i = 1 \mid x_0 = 0; x_h) + p(x_i = 1 \mid x_0 = 1; x_h)\right]\right\} \approx \frac{1}{4}$$

Case 2: For every v-collider $j \to k \leftarrow h$ in the DAG, if there is a directed path from k to i, then it must include 0. Therefore, $kP0$ for every v-collider $j \to k \leftarrow h$. Moreover, by the assumption that 0 does not d-separate i and A, there must be a v-collider $j \to k \leftarrow h$ such that there is a path between j and i that does not include 0. As in case 1, we can impose structure on p that enables us to effectively define p over three variables x_i, x_h, x_0 and assume that the DAG is $i \to 0 \leftarrow h$, and all calculations will be arbitrarily precise approximations.

Impose the following additional structure on p. First, $p(x_h = 1) = \frac{1}{2}$. Second, $p(x_i = x_h)$ with arbitrarily high probability. Third, $p(x_0 = 1 \mid x_i, x_h)$ is arbitrarily high when $x_i x_h = 1$ and arbitrarily low when $x_i x_h = 0$. Then, $p(x_i = 1) \approx \frac{1}{2}$. Now,

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

Then, $p_R(x_i = 1 \mid x_0 = 1) \approx 1$ and $p_R(x_i = 1 \mid x_0 = 0) \approx \frac{1}{3}$, such that $\sum_{x_0} p(x_0)p_R(x_i = 1 \mid x_0) \approx \frac{2}{3}$.

Extending the proof to arbitrarily large X is straightforward - we only need to assume that the marginal of p over each of the variables 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 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, $E_R(y_i \mid y_0) \equiv E_R(x_i \mid x_0) - \mu_i$ 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)), $p(x_i \mid x_{R(i)})$ is multivariate normal. Specifically, we can write $p(x_i \mid x_{R(i)})$ 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 $(p(x_i \mid x_{R(i)}))_{i=1,\ldots,n}$ constitutes a Gaussian Bayesian network (see Definition 7.1 in Koller and Friedman (2009)). By Theorem 7.3 in Koller and Friedman (2009), $p_R \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}')$, where $\boldsymbol{\Sigma}'$ is some variance-covariance matrix. Then, by the definition of conditional expectations under multivariate normal distributions, $E_R(x_i \mid x_0) = bx_0$, where b is some constant. Because $E(x_0) = 0$, it then immediately follows that

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

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 $p_R(x_i \mid x_A) = p_R(x_i)$. Since R is perfect, Corollary 1 and Lemma 1 imply that $p_R(x_i) = p(x_i)$, hence $E_R(x_i \mid x_A) = E(x_i)$ for all x_A. Now suppose that i is connected to A. By assumption, i is connected to at most one of the cliques that constitute A. Denote this clique by C. Then, $p_R(x_i \mid x_A) = p_R(x_i \mid x_C)$. 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

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

which implies the claim.

(Only if). Suppose that A is not a union of mutually disconnected cliques (in particular, A itself is not a clique). 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 $p_R(x_i \mid x_A)$. As before, we can consider w.l.o.g the case in which every variable can only take the values 0 and 1. 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 and arbitrarily high probability. Finally, suppose that $x_i = x_j x_k$ with independent and arbitrarily high probability. By construction,

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

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, $x_j \perp x_k \mid x_i$. Therefore, and by the additional assumptions we imposed on p,

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

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 $p(x_j = 1 \mid x_i = 1) = p(x_k = 1 \mid x_i =$

1) ≈ 1, whereas

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

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

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

which completes the proof.

Proposition 9

By assumption, the action s^0 generates a sure payoff of zero. Therefore,

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

Now suppose that $p(s_i \mid t_i) > \varepsilon$ 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

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

in order for the action to be a subjective best-reply. It follows that if ε is sufficiently small,

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

for every trader i. Therefore,

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

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

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

hence

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

a contradiction.

Proposition 10

Suppose that $i + n \in N$ for some $i = 1, ..., n$. Then, by Condition 2, the factorization formula (4) contains the term $p(e_i \mid x_0)$. Also, $i \in N$. By assumption, $p(E_R(x_i \mid x_0) \mid x_0) = 1$. Therefore, we can remove the term $p(e_i \mid x_0)$ from (4) altogether, and plug $e_i = E_R(x_i \mid x_0)$ 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.

References

[1] Barro, R. and D. Gordon (1983), "Rules, Discretion and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics* 12, 101-121.

[2] Cherry, J. and Y. Salant (2016), "Statistical Inference in Games," Northwestern University, Mimeo.

[3] Cho, I., N. Williams and T. Sargent (2002), "Escaping Nash Inflation," *Review of Economic Studies,* 69, 1-40.

[4] Cowell, R., P. Dawid, S. Lauritzen and D. Spiegelhalter (1999), *Probabilistic Networks and Expert Systems,* Springer, London.

[5] Esponda, I. (2008), "Behavioral Equilibrium in Economies with Adverse Selection," *The American Economic Review,* 98, 1269-1291.

[6] Esponda. I. and D. Pouzo (2016), "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models," *Econometrica* 84, 1093-1130.

[7] Ettinger, D. and P. Jehiel (2010), "A Theory of Deception," *American Economic Journal: Microeconomics* 2, 1-20.

[8] Evans, G. and S. Honkapohja (2001), *Learning and Expectations in Macroeconomics*, Princeton University Press.

[9] Eyster, E. and M. Piccione (2013), "An Approach to Asset Pricing under Incomplete and Diverse Perceptions," *Econometrica* 81, 1483-1506.

[10] Eyster, E. and M. Rabin (2005), "Cursed Equilibrium," *Econometrica,* 73, 1623-1672.

[11] Friedman, M. (1968), "The Role of Monetary Policy," *American Economic Review* 58, 1–17.

[12] Garcia-Schmidt, M. and M. Woodford (2015), "Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis," NBER Working Paper no. w21614.

[13] Giacomini, R. (2015), "Economic Theory and Forecasting: Lessons from the Literature," *Econometrics Journal* 18, C22-C41.

[14] Hoover, K. (2001), *Causality in Macroeconomics*, Cambridge University Press.

[15] Jehiel, P. (2005), "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory,* 123, 81-104.

[16] Klamer, A. (1984), *The New Classical Macroeconomics: Conversations with the New Classical Economists and their Opponents.* Wheatsheaf Books.

[17] Koller, D. and N. Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques.* MIT press.

[18] Kydland, F. and E. Prescott (1977), "Rules rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy* 85, 473-491.

[19] Liang, A. (2016), "Games of Incomplete Information Played by Statisticians," Harvard University, Mimeo.

[20] Lucas, R. (1972), "Expectations and the Neutrality of Money," *Journal of Economic Theory* 4, 103-124.

[21] Osborne, M. and A. Rubinstein (1998), "Games with Procedurally Rational Players," *American Economic Review,* 88, 834-849.

[22] Pearl, J. (2009), *Causality: Models, Reasoning and Inference,* Cambridge University Press, Cambridge.

[23] Phelps, E. (1967), "Phillips Curves, Expectations of Inflation and Optimal Unemployment over Time," *Economica* 34, 254–81.

[24] "Money-Wage Dynamics and Labor Market Equilibrium," *Journal of Political Economy* 76, 678–711.

[25] 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.

[26] Sargent, T. (1999), *The Conquest of American inflation*, Princeton University Press.

[27] Sargent, T. (2003), "Rational Expectations," *The Concise Encyclopedia of Economics.*

[28] 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.

[29] Schumacher, H. and H. Thysen (2017), "Equilibrium Contracts and Boundedly Rational Expectations", mimeo.

[30] Sloman, S. (2005), *Causal Models: How People Think about the World and its Alternatives*, Oxford University Press.

[31] Spiegler, R. (2016a), "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics* 131, 1243-1290.

[32] Spiegler, R. (2016b), "Data Monkies: A Procedural Model of Extrapolation from Partial Statistics," *Review of Economic Studies*, forthcoming.

[33] Verma, T. and J. Pearl (1991), "Equivalence and Synthesis of Causal Models," *Uncertainty in Artificial Intelligence,* 6, 255-268.

[34] Woodford, M. (2013), "Macroeconomic Analysis without the Rational Expectations Hypothesis," *Annual Review of Economics* 5.1, 303-346.