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

Ran Spiegler[†]

April 14, 2016

Abstract

Can a policy maker lead an agent to form systematically biased economic forecasts? This possibility is ruled out by the conventional rational-expectations postulate. I revisit this question and assume that the agent misperceives causal relations among economic variables, which may lead to non-rational expectations. The agent forms forecasts of economic variables after observing the policy maker's action. The agent's forecasts are based on fitting a subjective causal model - formalized as a direct acyclic graph, following the "Bayesian networks" literature - to objective long-run data. I show that the agent's forecasts can be systematically biased if and only if the agent's graph is not perfect - i.e., if the direction of at least some of the causal links he postulates is empirically meaningful. I characterize the policy maker's optimal strategy for several examples, mainly a stylized "monetary policy" application.

[*] Financial support by the Sapir Center is gratefully acknowledged. I thank Yair Antler, Kfir Eliaz, Rafaella Giacomini, Heidi Thysen, Todd Sarver, Vasiliki Skreta, Dimitri Vayanos and Michael Woodford 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 real-life interactions can be described as a principal-agent problem, in which the principal's ability to achieve his objectives depends on whether the agent correctly anticipates the principal's actions or their consequences. In some situations, the principal wishes to surprise the agent. For instance, success of a police crackdown on a drug-trafficking operation hinges on its unpredictability. Likewise, the immediate effect of a pay rise on worker morale is intuitively larger when it comes as a surprise. In other situations, the principal would like the agent to hold correct expectations. For example, when a company's management adapts its marketing strategy to changes in consumer demand, it would be better served if sales and service staff could anticipate the adaptation in advance, for the sake of swift implementation of the new strategy.

Monetary theory offers a prominent example of this general idea. In a well-known class of models, originated by Kydland and Prescott (1977) and Barro and Gordon (1983), the 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 (or unemployment) is determined by an "expectations-augmented" Phillips curve, such that the real effect of inflation is at least partly offset when inflation is anticipated. It follows that monetary policy involves "*expectations management*". To quote Woodford (2003, p. 15):

> "...successful monetary policy is not so much a matter of effective control of overnight interest rates as it is of shaping market expectations of the way in which interest rates, inflation and income are likely to evolve..."

Thus, to the extent that the central bank wishes to maximize expected output, it would like to set inflation systematically above private-sector expectations. And to the extent that the central bank wishes to minimize output fluctuations, it would like to avoid inflationary surprises.

In conventional models, the principal's ability to manage the agent's expectations is constrained by the assumption that the agent has "rational expectations" - i.e., he fully understands the statistical regularities in his environment, and thus forms unbiased forecasts of any economic variable conditional on his information. In this paper, I relax this assumption. Of course, one could model non-rational expectations in various ways. My approach is based on the following simple idea: *the agent derives his expectations by fitting a subjective causal model to objective long-run data.*

The idea that people reason about uncertainty via intuitive causal models has been studied extensively by experimental psychologists (e.g., see Sloman (2005)). In the more specific context of macroeconomic policy, private-sector agents hold intuitive, qualitative theories about the interconnection among macro variables; and such theories can sometimes be viewed as statements about causal relations. Indeed, Hoover (2001) describes historical controversies in macroeconomics in such terms. Furthermore, key private-sector actors (banks, financial-market speculators) regularly employ statistical models to form macroeconomic forecasts. While the exact specification of these models may be tweaked from time to time in order to get good empirical fit, their basic underlying causality assumptions are more likely to remain constant during times of relative stability. (For a study of how macroeconomic forecasters rely on models, see Giacomini et al. (2015).)

To formalize the notion that expectations are based on a subjective causal model, I employ a recent modeling framework (Spiegler (2015a)), which in turn builds on the Statistics and Artificial-Intelligence literature on *Bayesian networks* (Cowell et al. (1999), Pearl (2009)). The following example illustrates the modeling approach and its possible implications. It will serve as a running example throughout this paper.

An Example: Exploiting a Belief in Monetary Neutrality

Although this paper is a purely theoretical exercise, it will make use of a "monetary policy" scenario as a running example. Consider an environment in which the principal is a central bank and the agent is the private sector. I adapt a simple reformulation of the Barro-Gordon model due to Sargent (2001), Athey et al. (2005) and others. The central bank chooses an action a.

Inflation π is a stochastic function of a. The private sector forms its inflation forecast e after observing the central bank's move. Real output y is given by a "New Classical" Phillips curve, $y = \pi - e + \eta$ (where η is an independent, normally distributed variable with mean zero). Thus, only unanticipated inflation has real effects. The central bank has a single motive: maximizing expected output.

If the private sector had rational expectations, e would be equal to the true expected value of π conditional on a, and therefore ex-ante expected output would be zero, independently of the central bank's strategy. But now suppose that the private sector forms its expectations by reasoning in terms of a *causal model* that incorrectly links the relevant macro variables. Following Pearl (2009), a causal model is represented by a *directed acyclic graph* (DAG), where each node represents a variable and a direct link between two nodes signifies a perceived direct causal link between the variables they represent.

Specifically, suppose that the private sector's DAG, denoted R, is

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

This DAG represents a causal model according to which inflation is potentially a consequence of two independent causes: output and the central bank's action (the model omits the private sector's expectations). The causal model is entirely non-parametric - it does not assume anything regarding the sign or magnitude of causal relations - it merely postulates their existence and direction.

The causal model R is false because it perceives output to be exogenous and thus statistically independent of monetary policy, whereas according to the true process it is a consequence of the central bank's action via the Phillips Curve. Thus, the private sector subscribes to a "classical" worldview that postulates the absolute neutrality of monetary policy, whereas the true model allows for non-neutrality. Another way of expressing this disagreement is that R postulates that output causes inflation, whereas according to the true model, causation runs in the opposite direction. In other words, the private sector's subjective model exhibits *reverse causality.*

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

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

The formula $p_R(a, \pi, y)$ describes the private sector's subjective belief as a function of the true steady-state distribution p. It is an example of a "*Bayesian network factorization formula*" - it factorizes the steady-state distribution p into a product of conditional-probability terms, *as if* p were consistent with R. This is how I formalize the notion that the DM "fits a subjective causal model to the steady-state distribution". Because the causal model is entirely non-parametric, the private sector is always able to perfectly fit it to any objective distribution.

The subjective belief p_R systematically distorts the true correlation structure of the steady-state distribution p. The distortion arises because the private sector perceives statistical regularities through the prism of an incorrect causal model. Specifically, the private sector's inflation forecast after observing the central bank's action a is

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

This is in general different from the "rational" inflation forecast

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

The discrepancy arises because $p_R(\pi \mid a)$ involves an implicit expectation over y *without* conditioning on a.

The question is how the private sector's "non-rational" inflation forecast affects the central bank's considerations. It turns out that there are specifications of the exogenous processes (particularly the stochastic mapping from

a to π), for which the central bank can randomize over a in a way that leads the private sector to systematically underestimate inflation - i.e.,

$$\sum_a p(a)E_R(\pi \mid a) < \sum_\pi p(\pi)\pi$$

Consequently, the central bank can use monetary policy to enhance expected output. ■

In Section 2, I present a model that generalizes the above example. In the model, the agent forms forecasts of economic variables, after observing the realization of one particular variable (interpreted as the principal's action). The agent's forecasts are based on fitting a subjective causal model (which links nodes that represent variables and the agent's variable forecasts) to the joint probability distribution induced by the principal's strategy. The main question is: Is it possible for the agent to form a systematically biased forecast of any of the economic variables?

The main characterization result, given in Section 4, provides a simple answer to this question: systematically fooling the agent is possible (for a suitably chosen specification of the exogenous processes) if and only if the agent's DAG is *perfect*. A DAG is perfect if any two direct causes of a given variable must be directly linked themselves. The private sector's DAG in the above example violates perfection, because it perceives a and y as direct causes of π, and yet it does not postulate a direct causal link between the two. In contrast, the DAG $a \rightarrow y \rightarrow \pi$ is perfect, and therefore cannot give rise to systematically biased inflation forecasts.

Perfection is a key property in the literature on Bayesian networks, for various reasons. In the present context, it is important because it is equivalent to the property that the direction of any given causal link is not identified (because there exists a DAG that induces the same mapping from objective distributions to subjective beliefs, and reverses the link). Thus, the agent's misspecified causal model renders him vulnerable to biased forecasts if and only if it postulates empirically meaningful direction of causation.

Furthermore, Spiegler (2015b) showed that perfect DAGs can be interpreted as a representation of the agent's attempt to extrapolate a subjective belief from limited feedback. According to this interpretation, the agent does not have an explicit a-priori causal model. Instead, he gets access to partial observations regarding the objective long-run distribution (in the form of a large dataset with randomly generated missing values). The agent employs a procedurally rational rule for extending the data into a fully specified probability distribution over the economic variables, which distorts the objective distribution as if the agent tried to fit it to an explicit causal model, represented by a perfect DAG. The results in this paper imply that this procedure is sound, in the sense that it does not expose the agent to systematic forecast errors.

The perfection requirement can be weakened when we restrict the domain of permissible exogenous processes. For example, in our "monetary policy" example, if $\pi = a + \varepsilon$, where ε is an independent, normally distributed variable, the agent's inflation forecasts will be unbiased on average, regardless of the central bank's strategy.

This collection of results is intriguing, considering the heated historical debates over the exploitability of the inflation-output relation (see Klamer (1984)). The key assumption behind classical non-exploitability results (Lucas (1972), Sargent and Wallace (1975)) was allegedly the rationality of private-sector expectations. However, according to this paper, a considerably milder assumption - namely that the private sector forms its expectations by fitting a (potentially misspecified) causal model to long-run data - reproduces results in a similar vein.

If the agent's false causal perceptions do not make it easy for the principal to systematically fool him, this does *not* mean that they are irrelevant for his choice of strategy. For example, Section 5 studies an elaborate version of the above linear-normal specification of the "monetary policy" example, in which the central bank aims to target an exogenous inflation target. The central bank trades off the variance of real output and the mean square deviation of inflation from its target. The central bank's optimal policy is more rigid than in the rational-expectations benchmark. This rigidity is exacerbated when the variance of η becomes smaller relative to the variance of ε - i.e., when the Phillips relation becomes more reliable in relation to the transmission from central-bank actions to inflation.

This paper is related to a few works that examine monetary policy when the rational-expectations assumption is relaxed. Evans and Honkapohja (2001) and Woodford (2013) review dynamic models in which agents form non-rational expectations, and explore implications for monetary policy. See Garcia-Schmidt and Woodford (2015) for a recent exercise in this tradition. Sargent (2001), Cho et al. (2002) and Esponda and Pouzo (2015) study models in which it is the central bank who forms non-rational expectations, whereas the private sector is modeled conventionally.

More broadly, this paper contributes to the literature (reviewed in Spiegler (2015a)) that studies strategic interaction among agents who base their decisions on misspecified subjective models. Within this literature, Piccione and Rubinstein (2003) share the principal-agent "expectations management" aspect of the present paper. In their model, the principal is a seller who commits to a deterministic temporal sequence of prices, taking into account that consumers 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 price-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 have a biased estimate of the object's value.

2 The Model

Let $x_0, x_1, ..., x_n$ be a collection of real-valued economic variables. An agent observes the realization of x_0 and forms a subjective forecast e_i of each of the economic variables x_i, $i = 1, ..., n$. I use p to denote a joint distribution over all $2n+1$ variables. In all the applications in this paper, x_0 is interpreted as the action of a principal, possibly taken after having observed the realization of some of the other economic variables. Therefore, I will often refer to x_0 as an action and denote it by a.

If the agent's forecast is based on rational expectations, then p must satisfy the restriction that for every $i = 1, ..., n$, $p(e_i \mid x_0)$ assigns probability one to

$$E_p(x_i \mid a) = \sum_{x_i} p(x_i \mid a) x_i$$

Other models of belief formation would imply other restrictions on $p(e_i \mid a)$.[1]

Let us now introduce the idea that the agent forms his beliefs by fitting a misspecified causal model to long-run data. This will require basic concepts from the literature on Bayesian networks. The following exposition is standard (see Cowell et al. (1999) and Pearl (2009)), with a few minor adjustments (introduced in Spiegler (2015a) that serve the current purposes. To simplify exposition, denote $a = x_0$, and $e_i = x_{i+n}$ for every $i = 1, ..., n$, such that $x = (x_0, x_1, ..., x_{2n})$. Let $N^* = \{0, 1, ..., 2n\}$ be the set of variable indices. For every $M \subseteq N^*$, denote $x_M = (x_i)_{i \in M}$.

Define a *directed acyclic graph* (DAG) (N, R), where $N \subseteq N^*$ is the set of nodes and R is the set of directed links. (A directed graph is acyclic if it does not contain a directed path from a node to itself.) I use jRi or $j \rightarrow i$ interchangeably to denote a directed link from j into i. Observe that the binary relation R is asymmetric and acyclic. Abusing notation, let $R(i) = \{j \in N \mid jRi\}$ be the set of "direct parents" of node i. I will usually refer to R itself as the DAG.

Let $\tilde{R}$ be the *skeleton* (or undirected version) of R - i.e., $i\tilde{R}j$ if and only if iRj or jRi. A subset $M \subseteq N$ is a *clique* in R if $i\tilde{R}j$ for every $i, j \in M$. A clique M is *ancestral* if $R(i) \subset M$ for every $i \in M$. In particular, a node i is ancestral if $R(i)$ is empty. Given R, we say that a node j is an ancestor of another node i if R contains a directed path from j into i.

The agent is characterized by a DAG R. For any objective joint probability distribution p, the agent's subjective belief over x_N is

$$p_R(x_N) = \prod_{i \in N} p(x_i \mid x_{R(i)}) \tag{3}$$

[1] Throughout the paper, I use simple summations rather than integration, for notational clarity.

A probability distribution p is *consistent* with R if $p_R(x_N) \equiv p(x_N)$. Thus, R encodes a mapping that transforms every objective distribution p into a subjective belief p_R. When the DAG is fully connected, (3) is reduced to a textbook chain rule, such that $p_R = p$ for every p - i.e., the agent has "rational expectations". Note that in general, (3) may involve terms that condition on zero-probability events; when analyzing the model, I will need to rule out this possibility.

Following Pearl (2009), I interpret R as a *causal model.* The link $j \rightarrow i$ means that the agent regards the variable x_j to be an immediate cause of the variable x_i. While the agent presupposes the existence of this causal effect, he has no preconception regarding its sign or magnitude. In particular, this effect could be measured to be null. In other words, R is a "non-parametric model": even if p is governed by a parametric "true model", the agent does not impose any parametric restriction and fits its non-parametric causal model R to the data generated by the true model. For a concrete image to match this description, think of an analyst who tries to fit data with a system of structural equations. The analyst holds the collection of R.H.S variables in each equation fixed, but tweaks the exact functional form, until he gets good fit.

The agent's subjective distribution over any variable x_i, $i = 1, ..., n$, conditional on his observation of a is

$$p_R(x_i \mid a) = \frac{p_R(a, x_i)}{p_R(x_i)}$$

where $p_R(x_i) = \sum_{x_{-i}} p_R(x_i, x_{-i})$, as usual.

From now on, I impose one restriction on p and one restriction on R.

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

$$E_R(x_i \mid a) = \sum_{x_i} p_R(x_i \mid a) x_i \tag{4}$$

Condition 2 *The domain of permissible DAGs is restricted as follows. First,* $0 \in N$*. Second, if* $i \in N$ *for some* $i \in \{n+1, ..., 2n\}$*, then* $i - n \in N$ *and* $R(i) = \{0\}$*.*

Condition 1 is consistent with the interpretation of e_i as the agent's subjective forecast of x_i conditional on a. Condition 2 means that the agent's DAG perceives a to be the only immediate cause of his own forecasts (note that I use the notational convention $a = x_0$ and $e_i = x_{i+n}$). The justification for the latter restriction is that the agent actively conditions his forecasts on the principal's action and on no other variable, and it therefore makes sense to assume that his subjective causal model reflects this state of affairs. The condition also makes the self-evident requirement that the agent's DAG includes the observed signal as a variable, and that if the agent's DAG includes his forecast of some variable, it must also include the variable itself.

These two domain restrictions imply the following useful result.

Lemma 1 *Suppose that the domain of possible objective distributions satisfies Condition 1 and that* R *satisfies Condition 2. Then, there is a DAG* R' *that omits the nodes* $n+1, ..., 2n$ *altogether, such that* $p_{R'}(a, x_1, ..., x_n) \equiv p_R(a, x_1, ..., x_n)$ *for every* p *in the restricted domain. In particular, if* jRi *for some* $i \in \{1, ..., n\}$ *and* $j \in \{n+1, ..., 2n\}$*, then* $0R'i$*.*

Proof. Suppose that $i + n \in N$ for some $i = 1, ..., n$. Then, by Condition 2, the factorization formula (3) contains the term $p(e_i \mid a)$. Also, $i \in N$. By assumption, $p(E_R(x_i \mid a) \mid a) = 1$. Therefore, we can remove the term $p(e_i \mid a)$ from (3) altogether, and plug $e_i = E_R(x_i \mid a)$ in any other term in (3) that conditions on e_i - which effectively means that such a term conditions on a. 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 a into the same node. ■

This result means that we can assume w.l.o.g that the agent's DAG R is defined over $N \subseteq \{0, 1, ..., n\}$ - i.e., he omits his own forecasts from his causal model. I will follow this practice from now on. In addition, I will make minor assumptions (e.g., that p has full support on $(x_0, x_1, ..., x_n)$) that ensure that the factorization formula (3) does not involve terms that condition on zero-probability events.

The conditional expected value $E_R(x_i \mid a)$ is the agent's forecast of x_i after observing a. If the agent could - or felt the need to - test his causal model against historical data, he would discover the discrepancy between $E_R(\pi \mid a)$ and $E(\pi \mid a)$, thus refuting the model. I assume that no such "test for model misspecification" occurs. See Spiegler (2015a) for a detailed justification for this assumption.

3 "Monetary Policy" Example Revisited

The general problem in this paper will be: When is it possible for an agent with a misspecified DAG to form systematically biased economic forecasts? In applications, this question will be relevant because it is implied by the principal's objective function. To illustrate the problem, let us return to the "monetary policy" example of the Introduction, and show how a central bank may be able to exploit the private sector's misspecified causal model to generate systematic underestimation of inflation, thus enhancing expected real output.

In this example, there are only two economic variables: inflation π and real output y. Denote the private sector's inflation forecast by e. Both π and the central bank's action a take values in $\{0, 1\}$, where $\pi = 0$ (1) represents low (high) inflation. Assume that p satisfies $p(\pi = 1 \mid a) = \beta a$, where $\beta \in (0, 1)$. Thus, the action $a = 0$ induces low inflation with certainty, whereas the action $a = 1$ induces high inflation with probability β. For any given realization of π, e, $y = \pi - e + \eta$, where $\eta \sim N(0, \sigma_\eta^2)$ is independently distributed. That is, expected output is equal to the deviation of actual inflation from the private sector's forecast.

Note that p is consistent with the following “true DAG” R^*:

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

In contrast, the private sector’s DAG R is

$$a \rightarrow \pi \leftarrow y$$

Thus, as observed in the Introduction, the private sector’s causal model postulates that output fluctuations are exogenous, whereas inflation fluctuations are the consequence of output fluctuations as well as monetary policy. In relation to the true DAG R^*, R reverses the causal link between inflation and output, and it neglects the effect of inflationary expectations on output. The private sector’s conditional inflation forecast under R is

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

The central bank commits ex-ante to a probability distribution over a. Its strategy can be described by a single number, $\alpha = p(a = 1)$. Assume that the central bank has a sole objective: *maximizing expected output.* Plugging the “Phillips Curve” $y = \pi - e$, we obtain the following objective function:

$$\begin{aligned}
& \sum_{a} p(a) \left[E_p(\pi \mid a) - E_R(\pi \mid a) \right] \\
= \quad & E_p(\pi) - \sum_{a} p(a) E_R(\pi \mid a)
\end{aligned}$$

If the central bank could not systematically fool the private sector, the value of this objective function would be zero for any strategy that it might employ. However, we will now see that the central bank *is* able to cause the private sector to systematically underestimate inflation.

Proposition 1 *As $\sigma_\eta^2 \rightarrow 0$, the maximal expected output converges to $\frac{1}{4}\beta$.*

The level is attained by playing $\alpha = \frac{1}{2}$.

Proof. Denote $E_R(\pi \mid a) = e(a)$. Because $\pi \in \{0, 1\}$,

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

Because η is normally distributed, $p(a, y)$ has full support, such that $e(a)$ never involves conditioning on zero-probability events.

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 $E(y \mid a = 0) = 0$. It follows that if $\alpha = 1$, the central bank will not be able to induce strictly positive expected output. From now on, assume $\alpha > 0$.

Let us now calculate $e(1)$. First, note that conditional on $a = 1$, y is a normally distributed variable with a random mean: $y \sim N(1 - e(1), \sigma_\eta^2)$ with probability β, and $y \sim N(-e(1), \sigma_\eta^2)$ with probability $1 - \beta$. Therefore, the event in which $\pi = 1$ and $y > 1 - e(1)$ occurs with a probability greater than $\alpha\beta/2$. Likewise, the event in which $\pi = 0$ and $y < -e(1)$ occurs with a probability greater than $\alpha(1 - \beta)/2$. These two bounds imply that for any σ_η^2 and any fixed $\alpha > 0$, $e(1)$ is strictly between 0 and 1 and bounded away from both. In the $\sigma_\eta^2 \to 0$ limit, the ex-ante distribution of y has atoms at three points, $-e(1)$, 0 and $1 - e(1)$, such that $p(\pi = 1 \mid a = 1, y) \to 1$ in the neighborhood of $y = 1 - e(1)$, whereas $p(\pi = 1 \mid a = 1, y) \to 0$ in the neighborhoods of $y = 0$ and $y = -e(1)$. Thus, in the $\sigma_\eta^2 \to 0$ limit,

$$e(1) = \sum_y p(y)p(\pi = 1 \mid y) = p(\pi = 1) = \alpha\beta \tag{5}$$

We have thus established that $E(\pi) = \alpha\beta$ and $\sum_a p(a)e(a) = \alpha \cdot \alpha\beta + (1 - \alpha) \cdot 0 = \alpha^2\beta$. The central bank will choose α to maximize

$$\alpha\beta - \alpha^2\beta$$

which immediately gives the solution. ■

Thus, the central bank employs randomization to cause the private sector to systematically underestimate expected inflation, thus generating positive expected output. The intuition behind the result is as follows. When the realization of the central bank's strategy is $a = 0$, it induces $\pi = 0$ with certainty. In this case, the private sector's fallacious conditioning on y does not lead to a biased inflation forecast: $p(\pi = 0 \mid a = 0; y) = p(\pi = 0 \mid a = 0) = 1$, and therefore $p_R(\pi = 0 \mid a = 0) = 1$. In contrast, calculating the private sector's inflation forecast conditional on $a = 1$ involves summing over all values of y, without conditioning y on $a = 1$. In the $\sigma_\eta^2 \to 0$ limit, this failure to condition on $a = 1$ translates to the identity $E_R(\pi \mid a = 1) = E_R(\pi)$. Thus, when the central bank plays $a = 0$, the private sector correctly updates its belief, whereas when the central bank plays $a = 1$, the private sector forms its inflation forecast as if it has not observed the central bank's move! As a result, the private sector effectively "double counts" the episodes in which the central bank plays $a = 0$. This leads to systematic underestimation of expected inflation. Finally, note that β is completely irrelevant for the central bank's strategy, due to the linearity of $E_R(\pi \mid a = 1)$ in β.

4 General Analysis

In the previous section, we saw how a misspecified causal model may lead to a systematically biased forecast of some economic variable. On the other hand, other DAGs would always generate forecasts that are unbiased on average. A trivial example is when R is fully connected, such that the agent has rational expectations. For a somewhat less trivial example, consider an empty DAG R over $N = \{0, 1, 2..., n\}$ - i.e., $R(i) = \varnothing$ for every i. Then, it is easy to see from (3) that $p_R(x_i \mid a) \equiv p(x_i)$, such that $\sum_a p(a)E_R(x_i \mid a) = E_p(x_i)$.

Definition 1 *A DAG R induces unbiased forecasts if*

$$\sum_a p(a)E_R(x_i \mid a) = E_p(x_i)$$

for every $i \in N$ and every objective distribution p that satisfies Condition 1.

Our problem is to characterize the DAGs that induce unbiased forecasts. For this purpose, we need to introduce a few basic concepts and results from the Bayesian-networks literature.

Equivalent DAGs

A DAG encodes a mapping from objective distributions to subjective beliefs, which is given by (3). Two DAGs can be equivalent in the sense that they encode the same mapping.

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

For instance, the DAGs 1 → 2 and 2 → 1 are equivalent, by the basic identity $p(x_1)p(x_2 \mid x_1) \equiv p(x_2)p(x_1 \mid x_2)$. Moreover, a DAG that involves intuitive causal relations can be equivalent to a DAG that makes little sense as a causal model (e.g., it postulates a causal link between two variables in a direction that contradicts the temporal sequence of their realizations).

The following characterization of equivalent DAGs will be useful in the sequel. A *v-collider* in R is an ordered triple of nodes (i, j, k) such that iRk, jRk, $i\not Rj$ and $j\not Ri$ (that is, R contains links from i and j into k, yet i and j are not linked to each other). We will say in this case that there is a *v-collider into k*.

Proposition 2 (Verma and Pearl (1991)) *Two DAGs R and Q are equivalent if and only if they have the same skeleton and the same set of v-colliders.*

To illustrate this result, all fully connected DAGs have the same skeleton (every pair of nodes is linked) and an empty set of v-colliders, hence they are all equivalent (indeed, they all induce rational expectations because they reduce (3) to a textbook chain rule). In contrast, the DAGs 1 → 2 → 3 and 1 → 2 ← 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 DAG.

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

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

That is, a perfect DAG has the property that if iRk and jRk, then $i\tilde{R}j$ - i.e., 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. The DAG $a \rightarrow \pi \leftarrow y$ from our "monetary policy" example is imperfect, because it consists of a v-collider into π. In contrast, the DAG $a \rightarrow y \rightarrow \pi$ is perfect.

The following is an immediate implication of Proposition 2.

Corollary 3 *Two perfect DAGs are equivalent if and only if they have the same skeleton. In particular, if $M \subseteq N$ is a clique in a perfect DAG R, then M is an ancestral clique in some DAG in the equivalence class of R.*

This corollary means that the causal links postulated by a perfect DAG are not identified, in the sense that if iRj, there exists a DAG R' that is equivalent to R, such that $jR'i$. The direction of a causal link is empirically meaningful only when it is part of a v-collider.

The following lemma will be useful in the sequel. It establishes that if C is an ancestral clique in some DAG in the equivalence class of R, then the objective and subjective marginal distributions over x_C always coincide. Otherwise, we can find a distribution for which the two will diverge.

Lemma 2 (Spiegler (2015b)) *Let R be a DAG and let $C \subseteq N$. Then, $p_R(x_C) \equiv p(x_C)$ for every p if and only if C is an ancestral clique in some DAG in the equivalence class of R.*

We are now ready to state the main result of the paper. Recall that by the domain restrictions imposed in Section 2, we can assume w.l.o.g that the agent's DAG omits his own forecasts.

Proposition 3 *Suppose that the agent's DAG R is defined over $N \subseteq \{0, 1, ..., n\}$, where $0 \in N$. Then, R induces unbiased forecasts if and only if it is perfect.*

Proof. (**If**). Assume that R is perfect. Then, by Corollary 3, we can take 0 or i to be ancestral w.l.o.g. Then, by Lemma 2, $p_R(x_0) \equiv p(x_0)$ and $p_R(x_i) \equiv p(x_i)$. Therefore, we can write

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

which implies the claim.

(**Only if**). When R is imperfect, it must contain a v-collider $i \rightarrow j \leftarrow k$. Let us consider objective distributions p for which all other variables are independent, such that

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

This allows us to ignore all variables $i' \in N - \{i, j, k\}$ when calculating marginal or conditional distributions over x_j that are derived from p_R. In addition, suppose that the support of the marginal of p over any variable is of size two. W.l.o.g, we can assume for convenience that these two values are 0 and 1, such that the expected value of a variable w.r.t any probability distribution is equal to the probability that the variable takes the value 1.

There are three cases to consider. First, suppose that $0 \notin \{i, j, k\}$ - i.e., 0 is not part of the v-collider. Then, $p_R(x_j \mid x_0) \equiv p_R(x_j)$. By Proposition 2, j is not an ancestral node in any DAG in the equivalence class of R. Therefore, by Lemma 2, we can find p for which $p_R \neq p$. (Our restrictions on p are w.l.o.g in this regard, because we can ignore any variable $i' \neq i, j, k$ and set $R : i \rightarrow j \leftarrow k$. The only restriction that pertains to the variables x_i, x_j, x_k is that each takes two values, and that is innocuous for the application of Lemma 2.)

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

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

Let p have full support on (x_0, x_j, x_k), and impose the following additional structure. First, $p(x_0 = 1) = \frac{1}{2}$. Second, $x_k = x_j = x_0$ with arbitrarily high probability. Third, $p(x_j = 1 \mid x_0 \neq x_k)$ is arbitrarily small. Then,

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

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

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

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

Under the same p as in the previous case, $p_R(x_i = 1 \mid x_0 = 1)$ is arbitrarily close to 1, and $p_R(x_i = 1 \mid x_0 = 0)$ is arbitrarily close to $\frac{1}{3}$, such that

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

is arbitrarily close to $\frac{2}{3}$, whereas $p(x_i = 1) = \frac{1}{2}$. ■

Thus, as long as the agent's causal model is given by a perfect DAG (over some collection of variables, excluding the agent's own forecasts), he cannot be systematically fooled. Even if his conditional forecasts are incorrect, on average they are unbiased. For instance, in our running "monetary policy" example, if the private sector's DAG were $a \rightarrow y \rightarrow \pi$ or $\pi \leftarrow a \rightarrow y$, its output and inflation forecasts would be unbiased on average, even though the causal models these DAGs represent are misspecified. Conversely, if the agent's DAG is imperfect, there are objective distributions for which the agent's forecast of at least one of the economic variables is systematically biased.

As mentioned earlier in this section, perfect DAGs have the property that the causal links they postulate are unidentified, and in this sense completely spurious. Thus, the significance of Proposition 3 is that it demonstrates that the agent's misspecified causal model exposes him to systematic fooling if and only if the causal assumptions he makes are non-trivial.

Selective forecasts

The definition of unbiased forecasts that I utilized in this section is very demanding, because it requires the forecast of *any* variable to be unbiased. However, not all forecasts are necessarily economically relevant. For example, in the "monetary policy" example of Section 3, I assumed that the true process follows Sargent (2001). In particular, this meant that while the private sector's inflation forecast has implications for the realization of economic variables, its output forecast was irrelevant. In other conventional models of monetary policy - specifically, the so-called New Keynesian model - both inflation and output forecasts matter for the realization of macroeconomic variables (see Woodford (2003)). Thus, the forecasts that matter economically depend on the true model that underlies the objective distribution.

The following result is a sufficient condition for the agent's forecast of a *given* x_i to be unbiased. Fix a DAG (N, R) and consider a node $i \in N$. Let $\bar{R}$ be the transitive closure of R - i.e., $i\bar{R}j$ if there is a directed path in R from i to j. Define the following binary relation P. For every distinct $i, j \in N$, iPj if the following conditions hold: (i) it is not the case that $j\bar{R}i$; and (ii) $k\bar{R}j$ for some node k, such that $k = i$ or $k\bar{R}i$. Thus, iPj if x_i is a possibly indirect cause of x_j, or if the two variables have a common indirect cause without being indirectly caused by one another. I refer to $N - \{j \in N \mid iPj\}$ as the weak-upper-contour set of i induced by R, and denote it by $U_R(i)$. Note that $i \in U_R(i)$.

Proposition 4 *Let $i \in N - \{0\}$. Suppose further that the subgraph induced by R over $U_R(i)$ is perfect and contains 0. Then,*

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

Proof. It is immediate from the factorization formula (3) that all the variables x_j for which iPj are irrelevant for the calculation of $p_R(x_{U_R(i)})$. Therefore, we can ignore them. But by assumption, the subgraph over $U_R(i)$

induced by R is perfect. Because $0, i \in U_R(i)$, Proposition 3 implies the result. ■

Thus, as long as the violations of perfection occur "below" 0 and i in the causal hierarchy, they do not cause biased forecasts of x_i.

4.1 Unbiased Forecasts under Restricted Domains

Proposition 3 means that an imperfect DAG exposes the agent to being systematically fooled for *some* specification of the exogenous processes. However, in applications we typically impose additional structure on the exogenous processes, which restricts the domain of permissible objective distributions. Such domain restrictions extend the impossibility of systematically fooling an agent with causal misperceptions.

Throughout this sub-section, when I say that the agent's forecasts are unbiased, I mean that

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

for every $i = 1, ..., n$ and every p in the restricted domain (which, as usual, satisfies Condition 1).

The following pair of examples extend the "monetary policy" example of Section 3, by adding an economic variable θ that represents a state of Nature. The central bank privately observes θ before taking its action. For instance, θ may capture an inflation target that the central bank would like to implement. The other two economic variables, π and y, are assumed to be independent of θ conditional on a. That is, $p(\pi, y \mid \theta, a) \equiv p(\pi, y \mid a)$ for every objective distribution p in the restricted domain. Thus, p is consistent with the true DAG R^* given by

$$\begin{tikzcd} \theta \arrow[r] & a \arrow[r] \arrow[d] & \pi \arrow[d] \\ & e \arrow[r] & y \end{tikzcd} \tag{6}$$

Example 4.1 shows that when we restrict attention to distributions that are consistent with this DAG, an agent whose subjective DAG is imperfect will still have unbiased forecasts of inflation and output. Example 4.2 shows that when we further impose a conventional parametric specification of the domain of objective distributions, the imperfect DAG that led to unbiased forecasts in Section 3 will no longer do.

Example 4.1: An extended "monetary policy" example
Suppose that the private sector's DAG R is

$$\begin{tikzcd} \theta \arrow[r] \arrow[d] & a \arrow[d] \\ y \arrow[r] & \pi \end{tikzcd} \tag{7}$$

The economic interpretation of this causal model is exactly as in Section 3, except that now the private sector postulates that both the central bank's action and output are potentially influenced by the common exogenous variable θ.

Proposition 5 *Suppose that the private sector's DAG is (7). Then, the private sector's forecasts are unbiased.*

Proof. First, observe that π is irrelevant for calculating $p_R(\theta \mid a)$ or $p_R(y \mid a)$, and therefore for these purposes the node π can be removed, and the remaining DAG is perfect, such that the private sector's forecasts of θ and y are unbiased. Turning to the private sector's inflation forecast, observe that R is equivalent to a DAG that inverts the causal link between a and θ. Therefore, we can write

$$\sum_a p(a) p_R(\pi \mid a) = \sum_a p(a) \sum_\theta \sum_y p(\theta \mid a) p(y \mid a) p(\pi \mid \theta, y)$$

According to the true process, $y \perp \theta \mid a$. Therefore, $p(\theta \mid a)p(y \mid a) = p(\theta, y \mid$

a). It follows that

$$\begin{aligned}
\sum_a p(a)p_R(\pi \mid a) &= \sum_a p(a) \sum_\theta \sum_y p(\theta, y \mid a) p(\pi \mid \theta, y) \\
&= \sum_a p(a) \sum_\theta \sum_y \frac{p(\theta, y) p(a \mid \theta, y)}{p(a)} p(\pi \mid \theta, y) \\
&= \sum_\theta \sum_y p(\theta, y) p(\pi \mid \theta, y) \sum_a p(a \mid \theta, y) \\
&= p(\pi)
\end{aligned}$$

This completes the proof. ■

The argument behind the unbiased inflation forecast in this case relies on the particular conditional-independence property ($y \perp \theta \mid a$) that the objective distribution is assumed to satisfy.

Example 4.2: A linear-normal "monetary policy" example
In this example, I impose a conventional linear-normal structure on the mapping from the central bank's policy to inflation and output. Let e denote the private sector's inflation forecast. Then, π and y are given by

$$\begin{aligned}
\pi &= a + \varepsilon \\
y &= \gamma\pi - e + \eta
\end{aligned} \tag{8}$$

where $\gamma \geq 1$ is a constant, and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\eta \sim N(0, \sigma_\varepsilon^2)$ are independent. Note that this example goes beyond the example of Section 3 by introducing the parameter γ. When $\gamma > 1$, fully anticipated inflation has real effects. I do not impose any structure on the distribution of θ.

Assume that the private sector's DAG R is

$$\theta \rightarrow a \rightarrow \pi \leftarrow y \tag{9}$$

When $\gamma = 1$ and θ is constant, we are back with the example of Section 3.

Throughout this example, I use μ_z to denote the true expected value of any variable z.

Proposition 6 *Assume that the objective distribution satisfies (8). Then, the private sector's forecasts are unbiased.*

Proof. By the definition of R,

$$\begin{aligned} e_R(a) &= \sum_\pi p_R(\pi \mid a)\pi = \sum_\pi \sum_y p(y)p(\pi \mid a, y)\pi \\ &= \sum_y p(y)E(\pi \mid a, y) \end{aligned}$$

Because e is pinned down by a (i.e., it is independent of any other variable once we fix a), from now on we can replace any appearance of e in a term that holds a fixed with the notation $e(a)$. Since $\pi = a + \varepsilon$,

$$E(\pi \mid a, y) = a + E(\varepsilon \mid a, y)$$

Because $y = \gamma\pi - e + \eta$, we have

$$\gamma\varepsilon + \eta = y - \gamma a + e(a)$$

For given a and y, the R.H.S is a constant, whereas the L.H.S is a sum of two independent variables that are normally distributed with mean zero (and recall that the variance of $\gamma\varepsilon$ is $\gamma^2\sigma_\varepsilon^2$). Therefore, to calculate $E(\varepsilon \mid a, y)$, we can apply the standard formula for $E(X \mid X + Y)$ when X and Y are independent normal variables, and obtain

$$E(\varepsilon \mid a, y) = \frac{\beta}{\gamma}(y - \gamma a + e(a))$$

where

$$\beta = \frac{\gamma^2\sigma_\varepsilon^2}{\gamma^2\sigma_\varepsilon^2 + \sigma_\eta^2} \tag{10}$$

We can now write

$$\begin{aligned} e(a) &= \sum_y p(y)\left[a + \frac{\beta}{\gamma}y - \beta a + \frac{\beta}{\gamma}e(a)\right] \\ &= a(1-\beta) + \frac{\beta}{\gamma}e(a) + \frac{\beta}{\gamma}\mu_y \end{aligned}$$

Plugging the Phillips curve, we obtain

$$e(a) = (1 - \beta)a + \frac{\beta}{\gamma}e(a) + \frac{\beta}{\gamma}[\gamma\mu_a - E(e(a))]$$

This functional equation defines $e(a)$. Taking expectations, we obtain

$$E(e(a)) = (1 - \beta)\mu_a + \frac{\beta}{\gamma}E(e(a)) + \beta\mu_a - \frac{\beta}{\gamma}E(e(a))$$

such that

$$E(e(a)) = \sum_a p(a)e(a) = \mu_a = \mu_\pi$$

The latter identity follows from the assumption that $\pi = a + \varepsilon$ and ε is independent with mean zero.

This completes the proof. Nevertheless, it also enables us to get the following explicit solution for $e(a)$:

$$e(a) = \frac{\gamma - \gamma\beta}{\gamma - \beta}a + \frac{\gamma\beta - \beta}{\gamma - \beta}\mu_a$$

Plugging the expression for β, we obtain

$$e(a) = \frac{\sigma_\eta^2}{\gamma(\gamma - 1)\sigma_\varepsilon^2 + \sigma_\eta^2}a + \frac{\gamma(\gamma - 1)\sigma_\varepsilon^2}{\gamma(\gamma - 1)\sigma_\varepsilon^2 + \sigma_\eta^2}\mu_a \tag{11}$$

This expression will be useful in the sequel. ■

Equation (11) implies that when $\gamma = 1$, $e(a) \equiv a \equiv E_p(\pi \mid a)$. Thus, under the linear-normal specification with $\gamma = 1$, the private sector always makes optimal conditional inflation forecasts. When $\gamma > 1$, its conditional forecasts are incorrect because they assign positive weight to the ex-ante expected action. Nevertheless, the forecasts are correct on average.

4.2 More on Linear-Normal Models

Example 4.2 suggested that linear-normal specifications may give rise to unbiased forecasts, even when the agent's subjective DAG is imperfect and thus violates the condition for unbiased forecasts in unrestricted domains. In this sub-section I elaborate on this observation.

Let us return to a general environment with $n+1$ variables $x_0, x_1, ..., x_n$. Suppose that p is consistent with some true DAG R^*. Moreover, for every $i = 0, 1, ..., n$,

$$x_i = \sum_{j \in R^*(i)} \alpha_{ij} x_j + \varepsilon_i$$

where $\alpha_{ij} \neq 0$ for every $j \in R^*(i)$, and $\varepsilon_i \sim N(\mu_i, \sigma_i^2)$ is independently distributed. Thus, p is given by a recursive system of linear equations with independent normal error terms. We will say in this case that p is consistent with a linear-normal model. This is a conventional specification in economic applications. The following result shows that in this restricted domain, a very weak restriction on R is sufficient for unbiased forecasts.

Proposition 7 *Suppose that* 0 *is an ancestral node in some DAG in the equivalence class of* R*. Then, for every* $i = 1, ..., n$*,*

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

for every p *that is consistent with a linear-normal model.*

Proof. When p is consistent with a linear-normal model, we can rewrite the system of equations such that for every i,

$$x_i = \sum_{j \in R^{**}(i)} \gamma_{ij} \varepsilon_j$$

where R^{**} is an extension of R^* into a linear ordering (i.e., $jR^{**}i$ whenever R^* contains a directed path from j into i), and γ_{ij} is some constant (potentially zero). Thus, every x_i can be expressed as a sum of independent normal variables.

From now on, I will assume that $\mu_i = 0$ for every i. To see why this is w.l.o.g, note that this assumption means that

$$y_i = x_i + c_i$$

for every i, where c_i is a constant that involves μ and γ coefficients. It is therefore clear that $E_R(y_i \mid y_0) \equiv E_R(x_i \mid x_0) + c_i$ and $E_p(y_i) \equiv E_p(x_i) + c_i$, such that we can restate our result for y_i instead of x_i. This simplification means that $E_p(x_i) = 0$ for every $i = 0, ..., n$.

By assumption, we can regard 0 as an ancestral node in R. Also, it will simplify exposition if we align R with the natural order over $0, ..., n$, such that jRi implies $j < i/$. That is, for every $i = 1, ..., n$, $R(i) \subseteq \{0, ..., i-1\}$. Therefore, we can write

$$p_R(x_i \mid x_0) = \prod_{j=1,...,i} p(x_j \mid x_{R(j)})$$

such that

$$E_R(x_i \mid x_0) = \sum_{x_1} \cdots \sum_{x_{i-1}} \left(\prod_{k=1}^{i-1} p(x_k \mid x_{R(k)}) \right) \sum_{x_i} p(x_i \mid x_{R(i)}) x_i$$

The vector of random variables $x_{R(i)}$ can be expressed as a matrix times the vector $(\varepsilon_0, ..., \varepsilon_n)$. Because all the ε_i's are independent normal variables, $x_{R(i)}$ is jointly normal. Therefore, the expression

$$\sum_{x_i} p(x_i \mid x_{R(i)}) x_i = E(x_i \mid x_{R(i)})$$

is the expectation of a zero-mean normal variable conditional on the realization of a zero-mean multi-variate normal distribution. Hence,

$$E(x_i \mid x_{R(i)}) = \sum_{j \in R(i)} \gamma_{ij} x_j$$

where γ_{ij} is some constant. We have thus reduced $E_R(x_i \mid x_0)$ to

$$\sum_{j \in R(i)} \gamma_{ij} \sum_{x_1} \cdots \sum_{x_{i-1}} \left(\prod_{k=1}^{i-1} p(x_k \mid x_{R(k)}) \right) x_j$$

Consider the term that corresponds to some $j \in R(i)$. We can ignore the summation over all variables $k > j$, such that the term is reduced to

$$\sum_{x_1} \cdots \sum_{x_{j-1}} \left(\prod_{k=1}^{j-1} p(x_k \mid x_{R(k)}) \right) \sum_{x_j} p(x_j \mid x_{R(j)}) x_j$$

We can now repeatedly carry out this simplification in the same manner for each of these terms, until we eventually obtain

$$E(x_i \mid x_0) = b x_0$$

where b is some constant (potentially zero). Because $E(x_0) = 0$, it then immediately follows that

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

which completes the proof. ■

The condition that 0 is an ancestral node in some DAG in the equivalence class of R is a significant weakening of perfection (recall that in a perfect DAG, *every* node can be regarded as ancestral). The subjective DAGs in Section 3 and Example 4.2 satisfy this weaker property, and nevertheless gives rise to biased forecasts for a suitably specified p. However, as long as we restrict attention to objective distributions that are generated by linear-normal models, the agent's forecasts are always unbiased.

Note that the result of this sub-section does *not* imply Proposition 6, because the latter did not require the central bank's strategy to be linear-normal. Suppose that we adopt the interpretation of x_0 as an action taken by a principal, possibly after observing the realizations of some economic variables. If we force the principal to linear-normal strategies, Proposition 7 implies that the principal cannot generate biased forecasts (as long as R satisfies the sufficient condition). In principle, this may no longer hold when we remove this straitjacket and allow the principal to play any arbitrary strategy.

5 Conditional Forecast Errors

So far, the question we addressed was whether the agent's conditional forecasts of economic variables are unbiased *on average*. And indeed, in our running "monetary policy" example, this is all that mattered because we assumed that the central bank's sole objective was to maximize expected output. However, for many purposes, it also matters whether the agent's conditional forecasts depart from rational expectations for *given* realizations of a.

The requirement that the agent's forecasts are unbiased for every realization of a is of course more stringent than the requirement that his forecasts are unbiased *on average*. Correspondingly, we must seek more demanding conditions on R in order for this requirement to be satisfied. The following is an example of such a strengthening.

Suppose that R satisfies the sufficient condition of Proposition 4. If, in addition, $0Ri$, then $E_R(x_i \mid x_0) \equiv E_p(x_i \mid x_0)$. The reason is as follows. By assumption, the subgraph over $U_R(i)$ is perfect and contains both 0 and i. Since $0Ri$, $\{0, i\}$ is a clique in the subgraph. Perfection implies that we can regard it as an ancestral clique. Therefore, $p_R(x_0)$, $p_R(x_i)$ and $p_R(x_0, x_i)$ are all unbiased, which immediately implies the result.

In the remainder of this section, I present two examples in which the agent's misspecified causal model generates conditional forecasts errors, and I analyze the implications of these errors for the principal's choice of strategy.

5.1 The Exploitative Nutritionist

This sub-section is a variation on the "Dieter's Dilemma" example of Spiegler (2015a), showing how the principal can manipulate and take advantage of the agent's conditional forecast errors. The principal is a nutritionist who chooses whether to prescribe a food supplement to the agent, at a marginal cost k. The nutritionist's action a takes values in $\{0, 1\}$, where $a = 1$ means that he prescribes the supplement. There are two other relevant variables: the agent's state of health (denoted h), and the level of some chemical in the agent's blood (denoted c). Both c and h take values in $\{0, 1\}$, where $h = 1$

means that the agent is in good health, and $c = 1$ means that the chemical's level is abnormal. According to the true process, h is independent of a, and $p(h = 1) = \frac{1}{2}$; whereas c is a deterministic consequence of a and h given by $c = (1 - a)(1 - h)$.

Suppose that the agent's DAG is

$$R : a \rightarrow c \rightarrow h$$

That is, the agent's causal model reverses the direction of causation between h and c relative to the true process, which is consistent with the DAG $a \rightarrow c \leftarrow h$. Because the agent's DAG is perfect, it leads to health forecasts that are unbiased on average. However, as we shall see, the conditional health forecasts are typically incorrect.

Suppose that the nutritionist is able to exert monopoly power, such that the agent pays to the nutritionist an amount that is equal to his perceived value of the food supplement. The nutritionist's objective is to maximize its expected profit, which is thus given by

$$p(a = 1) \cdot [p_R(h = 1 \mid a = 1) - p_R(h = 1 \mid a = 0) - k]$$

If the agent had rational expectations, he would realize that $p(h = 1 \mid a = 1) = p(h = 1 \mid a = 0) = \frac{1}{2}$, because in reality h is independent of a.

Denote $p(a = 1) = \alpha$. Spiegler (2015a) shows that

$$\begin{aligned} p_R(h &= 1 \mid a = 0) = \frac{1}{2(1 + \alpha)} \\ p_R(h &= 1 \mid a = 1) = \frac{1}{1 + \alpha} \end{aligned}$$

such that the agent's willingness to pay for the supplement is

$$p_R(h = 1 \mid a = 1) - p_R(h = 1 \mid a = 0) = \frac{1}{2(1 + \alpha)}$$

The nutritionist's problem is thus reduced to choosing α to maximize

$$\alpha \cdot \left(\frac{1}{2(1+\alpha)} - k \right)$$

It follows that when $k \geq \frac{1}{2}$, the nutritionist is unable to profit from the agent's causal misperception. For $k < \frac{1}{2}$, the optimal solution is given by

$$\alpha^* = \min \left\{ 1, \sqrt{\frac{1}{2k}} - 1 \right\}$$

such that the agent's effective willingness to pay for the supplement is $\frac{1}{4}$ for $k \leq \frac{1}{8}$, and

$$\frac{\sqrt{2k}(1-\sqrt{2k})}{2}$$

for $k \in (\frac{1}{8}, \frac{1}{2})$.

5.2 Rigid Monetary Policy

For the last time in this paper, let us revisit the "monetary policy", adopting the linear-normal specification of Example 4.2. Unlike previous examples, here the central bank does not wish to exploit the private sector's conditional inflation-forecast errors. Rather, these errors are an impediment to achieving the central bank's objectives, and they constrain its ability to adapt monetary policy to changing circumstances.

Suppose that the exogenous variable θ represents an ideal inflation target. Let μ_z denote the true expected value of any variable z. The central bank's objective is to minimize

$$Var(y) + k \cdot E(\pi - \theta)^2$$

where $k > 0$ is a constant that captures the central bank's trade-off between its two motives (minimizing output variance and minimizing the mean square deviation of inflation from the target).

As a benchmark, suppose that the private sector has rational expectations. Then, its inflation forecast conditional on a is $E_p(\pi \mid a) = a$. Therefore,

$$y = (\gamma - 1)a + \gamma\varepsilon + \eta$$

and since ε and η are independent variables with mean zero, we can ignore them in the calculation of the objective function, which is reduced to

$$(\gamma - 1)^2 E[(a - \mu_a]^2 + k \cdot E[a - \mu_\theta]^2$$

Solving this problem is standard. The strategy that minimizes this objective function is

$$a^*(\theta) = \frac{k}{(\gamma - 1)^2 + k}\theta + \frac{(\gamma - 1)^2}{(\gamma - 1)^2 + k}\mu_\theta$$

for every θ. This solution does not rely on the normality assumption - it only requires ε and η to be independent zero-mean random variables.

The optimal policy under rational expectations exhibits some rigidity: it is a weighted average of the realized inflation target θ and the ex-ante average target μ_θ. A higher weight on the latter corresponds to a policy that is less responsive to fluctuations in the exogenous target. As γ approaches 1 - such that anticipated inflation matters less for output - the central bank's policy approaches perfect targeting.

The following result characterizes the central bank's optimal policy under the private sector's erroneous "classical" causal model.

Proposition 8 *Assume that the private sector's DAG is $R : \theta \to a \to \pi \leftarrow y$. Let $\pi = a + \varepsilon$, $y = \gamma\pi - e + \eta$, where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$; and $\eta \sim N(0, \sigma_\eta^2)$ are independent. Then, the central bank's optimal policy is*

$$a^{**}(\theta) = \frac{k}{\lambda(\gamma - 1)^2 + k}\theta + \frac{\lambda(\gamma - 1)^2}{\lambda(\gamma - 1)^2 + k}\mu_\theta$$

where

$$\lambda = \left(\frac{\gamma^2\sigma_\varepsilon^2 + \sigma_\eta^2}{\gamma(\gamma - 1)\sigma_\varepsilon^2 + \sigma_\eta^2}\right)^2$$

Proof. The central bank's problem is to choose a strategy (i.e., a potentially stochastic mapping from θ to a) that minimizes

$$Var(y) + kE(\pi - \theta)^2$$

subject to the constraints

$$\begin{aligned} \pi &= a + \varepsilon \\ y &= \gamma\pi - e(a) + \eta \end{aligned}$$

In Example 4.2, we saw that $e(a) = E_R(\pi \mid a)$ is given by (11). Therefore,

$$E(y \mid a) = (\gamma - \delta)a - (1 - \delta)\mu_a$$

such that

$$\mu_y = (\gamma - 1)\mu_a$$

where

$$\delta = \frac{\sigma_\eta^2}{\gamma(\gamma - 1)\sigma_\varepsilon^2 + \sigma_\eta^2}$$

Because ε and η are independent variables with mean zero, we can ignore them in the calculation of the objective function, which is reduced to

$$(\gamma - \delta)^2 E[(a - \mu_a]^2 + kE[a - \theta]^2 \tag{12}$$

This is exactly the same as in the rational-expectations case, except that the coefficient $(\gamma - \delta)^2$ replaces $(\gamma - 1)^2$. The policy that minimizes this expression is $a^{**}(\theta)$, as given in the statement of the proposition. Again, the derivation is standard and therefore omitted. ■

This result has a few noteworthy features. First, as observed in Section 4.1, the expression for $e(a)$ given by (11) implies that when $\gamma = 1$, the private sector's conditional inflation forecasts are always correct - despite having an incorrect causal model. In this case, the optimal policy under R coincides with the rational-expectations prediction for $\gamma = 1$ - namely, it fully tracks θ.

Deviations from the rational-expectations prediction occur only when $\gamma >$

1. In this case, the private sector's inflation forecast is a weighted average of a and its ex-ante expected value μ_a. That is, private-sector forecasts are not fully responsive to fluctuations in the central bank's actions. The intuition is the same as in Section 3: the private sector erroneously regards y as an exogenous variable that affects π, and therefore assigns some weight to the ex-ante expected value of y when forming its inflation forecast. Because y is in fact a consequence of a, the private sector ends up assigning weight to μ_a, thus failing to fully condition on the actual realization of a.

The *extent* of this failure depends on the relative magnitudes of σ_ε^2 and σ_η^2. As the Phillips relation becomes more reliable - or, equivalently, as the effect of monetary policy on inflation becomes less reliable - the erroneous weight on μ_a increases and the deviation from rational expectations is exacerbated.

The private sector's "expectational rigidity" impels the central bank toward a more rigid policy than in the rational-expectations benchmark. This can be immediately seen from the effective objective function (12). Since $\delta \leq 1$ by definition, the central bank places a larger weight on the consideration of minimizing the variance of a, compared with the rational-expectations benchmark. Excess rigidity of the optimal policy increases with $\sigma_\varepsilon^2/\sigma_\eta^2$.

6 Discussion

In this section I briefly discuss a few variations and extensions of the model.

6.1 Ex-ante Forecasts

Throughout this paper, I assume that the agent forms a forecast of each economic variable after observing the principal's action. A natural variant would assume that the agent forms his forecast without observing the principal's move. In this case, the question becomes whether the agent's marginal subjective distribution over any given economic variable (including the unobserved action) is unbiased on average.

Formally, we will say that a DAG R induces unbiased ex-ante forecasts if $E_R(x_i) \equiv E_p(x_i)$. The following result is a simple corollary of Proposition

2 in Spiegler (2015b). Suppose that the agent's DAG R is defined over $N \subseteq \{0, 1, ..., n\}$. Then, R induces unbiased ex-ante forecasts if and only if it is *perfect*. Thus, perfection turns out to characterize the property of unbiased forecasts, whether or not we assume that the agent observes the principal's move prior to forming his forecast.

6.2 The Principal's Commitment Problem

In all the versions of the "monetary policy" example that appeared in this paper, we looked for the central bank's ex-ante optimal strategy. This implicitly assumes that the central bank is able to commit ex-ante to its policy. Of course, the original Kydland-Prescott and Barro-Gordon models were developed to highlight the role of commitment when the private sector has rational expectations. However, note that in this paper, I assumed that the private sector *observes* the central bank's actions. If the private sector had rational expectations, there would be no role for ex-ante commitment, because the central bank would never be tempted to deviate from the ex-ante optimal action: the private sector would be able to monitor any deviation from the pre-committed action and adapt its rational forecasts accordingly.

In contrast, when the private sector has a misspecified causal model, a commitment problem does arise even when it perfectly monitors the central bank's actions. Suppose that $R : \theta \rightarrow a \rightarrow y \leftarrow \pi$. We saw in Section 4 that in this case, the private sector's inflation forecast is unbiased on average. Yet, at the same time it is entirely unresponsive to the realization of a. In other words, the private sector forms its inflation forecast as if it has rational expectations but cannot monitor the central bank's action - exactly as in the original Kydland-Prescott and Barro-Gordon models! To conclude, when the agent has causal misperceptions, the principal has a time-consistency problem even if the agent observes his move prior to making his forecasts.

6.3 Relevance to Dynamic Models

The basic model does not make any explicit assumptions regarding the temporal realization of economic variables. Yet all the applications we have seen were static. Nevertheless, the formalism can be applied to dynamic models. Consider a discrete-time environment with an infinite horizon. There is a collection of exogenous variables, $\theta = (\theta_1, ..., \theta_m)$, and a collection of endogenous variables $y = (y_1, ..., y_r)$. Let θ^t and y^t denote the realizations of θ and y at period t.

Imagine that the agent believes that the exogenous variables θ evolve according to some stochastic process with bounded memory, such that the realization of θ^t is a stochastic function of $\theta^{t-1}, ..., \theta^{t-K}$, where K is constant. In addition, the agent postulates that the endogenous variables evolve according to a "Markov equilibrium", such that y^t is a stochastic function of $(\theta^{t-K}, ..., \theta^t)$. These assumptions imply a belief that exogenous and endogenous variables jointly evolve according to a Markov process, whose invariant distribution plays the role of the objective distribution p in our model. The DAG R - defined over nodes that correspond to lagged variables - represents structural assumptions regarding this Markov process. I hope to pursue dynamic applications of the formalism along these lines in future works.

References

[1] Athey, S., A. Atkeson and P. Kehoe (2005), "The Optimal Degree of Discretion in Monetary Policy," *Econometrica* 73, 1431-1475.

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

[3] 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. and D. Pouzo (2015), "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models," mimeo.

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

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

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

[9] Giacomini, R., V. Skreta and J. Turen (2015), "Models, Inattention and Expectation Updates," CEPR Discussion Paper no. 11004.

[10] Hoover, K. (2001), *Causality in macroeconomics*, Cambridge University Press.

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

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

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

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

[15] 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.

[16] Sargent, T. (2001), *The conquest of American inflation*, Princeton University Press.

[17] 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.

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

[19] Spiegler, R. (2015a), "Bayesian Networks and Missing-Data Imputation," *Quarterly Journal of Economics*, forthcoming.

[20] Spiegler, R. (2015b), "On the 'Limited-Feedback' Foundation of Boundedly Rational Expectations," mimeo.

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

[22] Woodford, M. (2013), "Macroeconomic Analysis without the Rational Expectations Hypothesis," *Annual Review of Economics*, forthcoming.