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

Ran Spiegler[†]

June 29, 2016

Abstract

The conventional rational-expectations postulate rules out the possibility that agents will form systematically biased forecasts of economic variables. I revisit this question under the assumption that agents' expectations are based on a misperceived causal model. Specifically, I analyze a model in which an agent forms forecasts of economic variables after observing a signal. His 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 are never systematically biased if and only if his graph is perfect - equivalently, if the direction of the causal links he postulates has no empirical content. I demonstrate the relevance of this result for economic applications - mainly a stylized "monetary policy" example in which the inflation-output relation obeys an expectations-augmented Phillips curve.

[*] Financial support by the Sapir Center is gratefully acknowledged. I thank Yair Antler, Kfir Eliaz, Rafaella Giacomini, Todd Sarver, Vasiliki Skreta, Dimitri Vayanos, Michael Woodford and especially Heidi Thysen, for helpful comments and discussions.

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

1 Introduction

The outcome of many real-life interactions hinges on whether agents correctly forecast particular variables. For instance, success of a police crackdown on a drug-trafficking operation hinges on its unpredictability. Likewise, the immediate effect of a wage cut on worker morale may depend on whether it comes as a surprise. Finally, well-coordinated team production relies on one unit's ability to anticipate how another unit will adapt to an observed shock.

In conventional models, an agent's forecasts 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 economic variable conditional on his information. His predictions may miss the target, but the errors will cancel out on average. In other words, the agent cannot be "systematically fooled".

Indeed, economists sometimes identify this property with the rational-expectations principle itself. The following quote from an 2010 interview with John Cochrane is representative:

> "What is rational expectations? It is the statement that you [cannot] fool all the people all the time."[1]

However, rational expectations involve more than the requirement that agents' forecasts of individual variables are unbiased on average, because they demand a correct perception of the *entire* joint distribution over all variables. A priori, an agent's beliefs may satisfy the former while violating the latter.

In this paper, I relax rational expectations and revisit the question of whether agents can be systematically fooled. Of course, one can depart from rational expectations in many directions. I focus on the role of *causal misperceptions* in the formation of beliefs, and assume that the agent derives his expectations by fitting a misspecified causal model to objective long-run data.

[1] See http://www.newyorker.com/news/john-cassidy/interview-with-john-cochrane.

Example 1.1: Exploiting a belief in monetary neutrality

Perhaps the most well-known manifestation of the question of whether economic agents can be systematically fooled lies within monetary theory. 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 (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.

Although this paper is a purely theoretical exercise, it will make use of a running example that is based on a simple reformulation of the Barro-Gordon model studied by Sargent (2001), Athey et al. (2005) and others. 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 independent Gaussian noise. 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.

Now suppose that the private sector forms its expectations by reasoning in terms of a *causal model* that links the relevant macro variables. The idea that people reason about uncertainty via intuitive causal models has been studied extensively by experimental psychologists (see Sloman (2005)). In the specific context of macroeconomics, policy makers and private-sector actors often believe in basic narratives about how macro variables are interconnected. Such narratives are often causal - indeed, Hoover (2001) describes historical controversies in macroeconomics in such terms. Furthermore, key financial-sector actors employ statistical models to form macroeconomic forecasts. These models sometimes take the form of a recursive system of equations, which is consistent with a causal model. While the functional forms of these equations may be tweaked from time to time for the sake of empirical fit, their underlying causality assumptions are more likely to remain constant during times of relative stability.[2]

To formalize the notion that agents rely on causal models to form expectations, 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)). A causal model is represented by a *directed acyclic graph* (DAG); 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 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 postulates direct causal relations between variables without assuming anything regarding their sign or magnitude.

The causal model R is misspecified because it perceives output to be

[2] For a study of how macroeconomic forecasters rely on models, see Giacomini et al. (2015).

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 via inflationary surprises.

How does the private sector employ its causal model to forecast inflation? It simply *fits* the model to the true steady-state joint distribution p over a, π, y. If p *were* 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*", which factorizes the steady-state distribution p into a product of conditional-probability terms, *as if* p were consistent with R. Because the private sector perceives statistical regularities through the prism of an incorrect causal model, the subjective belief p_R may systematically distort the true correlation structure of the steady-state distribution p.

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

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

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. Note that because the steady-state distribution p is affected by the private sector's expectations, it is essentially an "equilibrium" distribution; the equilibrium requirement is that the private sector's inflation forecast e is $E_R(\pi \mid a)$ as given by (3).

How does the private sector's "non-rational" inflation forecast affect the central bank's considerations? In Section 3, I present a natural specification of the 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.

Plan of the paper

In Section 2, I present a general model, in which an agent forms forecasts of economic variables after observing the realization of one variable. The agent's subjective causal model is represented by a DAG over a set of nodes that correspond to some subset of the economic variables. He fits this model to a joint probability distribution over all variables, including possibly the agent's forecasts. The distribution satisfies an "equilibrium" condition that the agent's forecasts are consistent with his causal model.

I ask the following question: Can such an agent be systematically fooled? The main result, given in Section 4, provides a simple answer: The agent's forecasts are always correct on average if and only if his DAG is *perfect.* A DAG is perfect if any pair of direct causes of a given variable must be 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 the two. As a result, we could find *some* objective distribution for which the agent's forecast of *some* variable is biased on average. In contrast, the DAG $a \rightarrow y \rightarrow \pi$ is perfect, and therefore cannot give rise to systematically biased forecasts of inflation or output.

Perfection is a familiar property in the Bayesian-networks literature. In the present context, its significance is that in perfect DAGs (and only in such DAGs), the direction of any given causal link is unidentified (in the sense that there exists a DAG that induces the same mapping from objective distributions to subjective beliefs, for which this link is reversed). 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 when R is perfect, p_R can also be interpreted as the outcome of an attempt (by the agent himself or by an "analyst" he relies on) to extrapolate a subjective belief from partial statistical datasets drawn from p, via an intuitive procedure (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 main result thus implies that the extrapolation 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 are interested in forecasts of $specific$ variables, or when we restrict the domain of permissible exogenous processes. In particular, in Section 5 I show that in our "monetary policy" example, when $\pi = a + \varepsilon$ and ε is independent Gaussian noise, the agent's inflation forecasts are consistent with rational expectations for *any* realization of the central bank's action. In this case, the classical result regarding the non-exploitability of the Phillips relation continues to hold.

Impossibility results of this kind are 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.

So far, our discussion focused on whether the agent's forecasts are correct *on average*. In many contexts, however, conditional forecast errors matter even if they cancel out. In Section 6 I characterize the DAGs for which the agent's conditional forecast errors are always correct - a far more demanding characterization than perfection. I also present two examples that demonstrate the implications of DAGs that lead to conditional forecast errors, even though forecasts are systematically unbiased. First, I study a monopoly pricing example, in which demand for the monopolist's intrinsically useless product stems from consumers' "reverse causality" misperception. Second, I examine an elaborate version of the linear-normal specification of the "monetary policy" example, in which the central bank trades off the variance of real output and the mean square deviation of inflation from an exogenously distributed target. The central bank's optimal policy displays excess rigidity relative to the rational-expectations benchmark, which is exacerbated as the Phillips relation becomes less noisy.

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 other 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 $i = 1, ..., n$, $p(e_i \mid a)$ 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 $p(e_i \mid a)$.

Let us now introduce the idea that the agent forms his beliefs by fitting a subjective 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. I will sometimes denote $e_i = x_{i+n}$ (for every $i = 1, ..., n$) and $x = (x_0, x_1, ..., x_{2n})$. For every $M \subseteq \{0, 1, ..., 2n\}$, denote $x_M = (x_i)_{i \in M}$.

Define 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. (A directed graph is acyclic if it does not contain a directed path from a node to itself.) I use jRi

or $j \to i$ 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 "parents" of node i. I will usually refer to R itself as the DAG.

Let $\tilde{R}$ be the *skeleton* (i.e., undirected version) of R - i.e., $i\tilde{R}j$ if and only if iRj or jRi. A subset $M \subseteq N$ is a *clique* 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. A node j is an *ancestor* of another node i in R 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

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

Thus, R encodes a mapping that transforms every objective distribution p into a subjective belief p_R. A probability distribution p is *consistent* with R if $p_R(x_N) \equiv p(x_N)$. When the DAG is fully connected, (4) is reduced to a textbook chain rule, such that every p is consistent with R - i.e., the agent has "rational expectations". Note that in general, (4) may involve terms that condition on zero-probability events; when analyzing the model, I will need to rule out this possibility. The agent's subjective distribution over any variable x_i, $i = 1, ..., n$, conditional on his observation of a is $p_R(x_i \mid a) = p_R(a, x_i)/p_R(x_i)$, where $p_R(x_i) = \sum_{x_{-i}} p_R(x_i, x_{-i})$, as usual.

Following Pearl (2009), I interpret R as a *causal model.* The link $j \to i$ 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 link, 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". As a result, the agent is always able to perfectly fit it to any objective distribution. For a concrete image to match this description, think of an analyst who tries to fit data with a recursive system of 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.

From now on, I assume that $0 \in N$. This motivation for this restriction is that the agent's causal model should acknowledge a variable that he active conditions his forecast on. I also impose the following "equilibrium" restriction on the objective distribution p.

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*

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

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 long-run data, he would discover any discrepancy between $E_R(x_i \mid a)$ and $E(x_i \mid a)$, thus refuting the model. I assume that no such "test for model misspecification" occurs. See Spiegler (2015a) for a detailed justification for this assumption.

Should we admit forecasts as variables in the agent's causal model?
By assumption, the agent's DAG does not admit his own forecasts as variables. However, forecasts are themselves variables that can play a role in the determination of economic outcomes (e.g., see the expectational Phillips curve in Example 1.1). Therefore, in principle they could enter the agent's causal model. Recall the notation $e_i = x_{i+n}$ for every $i = 1, ..., n$. Allow the set of nodes N in the agent's DAG to be a subset of $\{0, 1, ..., 2n\}$. Thus, when $i \in N$ for some $i > n$, this means that the agent's causal model admits e_{i-n} as a variable. Recall our earlier restriction that $0 \in N$. When as admit forecasts as variables, 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 makes two requirements. First, it says that the agent perceives x_0 to be the only immediate cause of his own forecasts. The justification is that the agent actively conditions his forecasts on x_0 alone; his causal model should acknowledge this. Second, it requires that if the agent's

DAG includes a forecast of some variable, then it must also include the variable itself.

These two domain restrictions imply the following result.

Remark 1 *Suppose that the domain of permissible objective distributions satisfies Condition 1 and that R satisfies Condition 2 (as well as the requirement that $0 \in N$). Then, there is a DAG R' that omits the nodes $n+1,...,2n$ 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. 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 (4) 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 (4) altogether, and plug $e_i = E_R(x_i \mid a)$ in any term in (4) 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 our original assumption that the agent's DAG omits his own forecasts is w.l.o.g - as long as we accept the domain restrictions on p and R. Therefore, I will continue to follow this practice from now on.

3 The "Monetary Policy" Example

The general problem in this paper is: When will an agent with a misspecified causal model 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 Example 1.1. Recall that in this example, there are three economic variables: the central bank's action a, inflation π and real output y. The private sector's inflation forecast is denoted e. Both π and 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. Output is given by $y = \pi - e + \eta$, where $\eta \sim N(0, \sigma_\eta^2)$ is independently distributed.

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

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

In contrast, the private sector's DAG R is $a \rightarrow \pi \leftarrow y$. 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 (3).[3]

The central bank commits ex-ante to a probability distribution over a. Its strategy is defined by $p(a = 1) = \alpha$. Assume that the central bank has a sole objective: *maximizing expected output.* Plugging the Phillips curve, we obtain the following objective function:

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

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 can use a random strategy to cause the private sector to systematically underestimate inflation, thus enhancing expected output.

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 $\pi \in \{0,1\}$,

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

[3] Throughout the paper, I use simple summations rather than integration when writing down expressions for $E_R(x_i \mid x_0)$, for notational clarity.

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 $\alpha = 0$, the central bank cannot induce strictly positive expected output. From now on, assume $\alpha > 0$.

Let us now calculate $e(1)$. First, note that $y \sim N(\mu, \sigma_\eta^2)$, where μ is random: $\mu = e(0) = 0$ with probability $1 - \alpha$, $\mu = 1 - e(1)$ with probability $\alpha\beta$, and $\mu = -e(1)$ with probability $\alpha(1-\beta)$. A priori, two of these three values could coincide. However, we will now see that this is not the case. Because the normal distribution is symmetrically distributed around its mean, the ex-ante probability of $y < -e(1)$ is at least $\alpha(1-\beta)/2$, whereas the ex-ante probability of $y > 1 - e(1)$ is at least $\alpha\beta/2$. Moreover, as σ_η^2 tends to 0, $p(\pi = 1 \mid a = 1, y < -e(1)) \to 0$ 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 as σ_η^2 approaches zero, μ gets *exactly* three values, $-e(1)$, 0 and $1 - e(1)$, and the gap between these values is bounded away from zero. In the $\sigma_\eta^2 \to 0$ limit, $p(\pi = 1 \mid a = 1, y) \to 1$ 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 $y = -e(1)$. Consequently, $e(1) \to p(\pi = 1) = \alpha\beta$ as $\sigma_\eta^2 \to 0$.

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

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. Therefore, the private sector's failure to properly account for variations in y does not lead to a biased inflation estimate: because $p(\pi = 0 \mid a = 0; y) = 1$ for any y, we have $E_R(\pi \mid a = 0) = 0$. In contrast, when $a = 1$, inflation does fluctuate, and the private sector's error is that it tries to account for these fluctuations by fluctuations in y, as if the latter are exogenous. Therefore, the private sector's inflation forecast conditional on $a = 1$ involves summing over all values of y, *without* conditioning y on $a = 1$. In the $\sigma_\eta^2 \to 0$ 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 downward, whereas when the central bank plays $a = 1$, the private sector forms its inflation forecast as if it did not observe the central bank's action. This leads to systematic underestimation of expected inflation. Note that β is 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 DAG may lead to a systematically biased forecast of some economic variable. However, other DAGs *always* generate forecasts that are correct on average. A simple example is an empty DAG (i.e., $R(i) = \varnothing$ for every $i \in N$). It is easy to see from (4) that in this case, $p_R(x_i \mid a) \equiv p(x_i)$ and therefore $\sum_a p(a) E_R(x_i \mid a) = E_p(x_i)$.

Definition 1 *A DAG R induces unbiased forecasts if*

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

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

The role of the full-support restriction is to prevent p_R from including terms that condition on zero-probability events. Condition 1 plays no technical role, and I introduce it only for the sake of maintaining the equilibrium interpretation of p. 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 (4). Two DAGs can be equivalent in the sense that they encode the same mapping.

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

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

A *v-collider* in R is an ordered triple of nodes (i, j, k) such that iRk, jRk, $i\not{R}j$ and $j\not{R}i$ (that is, R contains links from i and j into k, yet i and j are not linked to each other). We 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. In contrast, the DAGs $1 \rightarrow 2 \rightarrow 3$ and $1 \rightarrow 2 \leftarrow 3$ are not equivalent: although their skeletons are identical, the former DAG has no v-colliders whereas $(1, 3, 2)$ is a v-collider in the latter.

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

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

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. If we think of a DAG as a recursive system of structural (non-parametric) equations, then perfection means that if x_i and x_j appear as explanatory variables in the equation for x_k, then there must be an equation in which one of these two variables is explanatory and the other is dependent.

The following is an immediate implication of Proposition 2.

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

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

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

Lemma 1 (Spiegler (2015b)) *Let R be a DAG and let $C \subseteq N$. Then, $p_R(x_C) \equiv p(x_C)$ for every p with full support on X_N if and only if C is an ancestral clique in some DAG in the equivalence class of R.*

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

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

Proposition 3 *A DAG R induces unbiased forecasts if and only if it is perfect.*

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

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

which implies the claim.

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

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

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.

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 1, we can find p for which $p_R \neq p$. (Our restrictions on p are w.l.o.g in this regard, because we can ignore all nodes $i' \neq i, j, k$ and set $R : i \rightarrow j \leftarrow k$.)

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

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

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

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

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

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

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

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

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

Thus, as long as the agent's DAG is perfect, he cannot be systematically fooled. Even if his conditional forecasts are incorrect, the errors cancel out on average. 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 average forecast of at least one of the economic variables is 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 *all* variables to be unbiased. However, not all forecasts need to be 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 $i \in N$. Define a binary relation P as follows. For every distinct $i, j \in N$, iPj if at least one of the following conditions hold in R: (i) i is an ancestor of j; (ii) i and j have a common ancestor and j is not an ancestor of i. Denote $L_R(i) = \{j \in N \mid iPj\}$. Observe that $i \notin L_R(i)$.

Proposition 4 *Let $i \in N - \{0\}$. Suppose further that the subgraph induced by R over $N - L_R(i)$ 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 (4) that if iPj, then x_j is irrelevant for the calculation of $p_R(x_{N-L_R(i)})$. Therefore, we can ignore all such variables. By assumption, the subgraph over $N - L_R(i)$ induced by R is perfect. Because $0, i \in N - L_R(i)$, 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.

5 Linear-Normal Models

Proposition 3 means that an imperfect DAG exposes the agent to being systematically fooled for *some* objective distribution. However, in applications we typically impose additional structure that restricts the domain of permissible objective distributions. Such domain restrictions extend the impossibility of systematically fooling an agent with causal misperceptions. In this section I focus on a common domain restriction, which assumes that variables are linked by a system of linear equations with Gaussian noise.

Example 5.1: A linear-normal "monetary policy" example
Modify the example of Section 3 by assuming that π and y are given by the following equations:

$$\begin{aligned} \pi &= a + \varepsilon \\ y &= \gamma\pi - e + \eta \end{aligned} \tag{6}$$

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. This example changes the mapping from a to π, and also introduces the new parameter γ. When $\gamma > 1$, fully anticipated inflation has real effects. Throughout this example, I use μ_z to denote the true expected value of any variable z.

Proposition 5 *Suppose that the private sector's DAG is $R : a \rightarrow \pi \leftarrow y$. Then, the private sector's forecasts are unbiased for every objective distribution that satisfies (6).*

Proof. Because y is an ancestral node in R, Proposition 4 implies that the private sector's output forecast is unbiased. Let us turn to the private sector's inflation forecast. As in Section 3, I use $e(a)$ to denote the forecast after observing the realization a. By the definition of R,

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

Since $\pi = a + \varepsilon$, $E(\pi \mid a, y) = a + E(\varepsilon \mid a, y)$. By the second equation in (6), 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

$$\beta = \frac{\gamma^2\sigma_\varepsilon^2}{\gamma^2\sigma_\varepsilon^2 + \sigma_\eta^2} \tag{7}$$

We can now write

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

Since $\pi = a + \varepsilon$ and $E(\varepsilon) = 0$, $\mu_\pi = \mu_a$. 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

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

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

$$e(a) = \frac{\sigma_\eta^2}{\gamma(\gamma-1)\sigma_\varepsilon^2 + \sigma_\eta^2}a + \frac{\gamma(\gamma-1)\sigma_\varepsilon^2}{\gamma(\gamma-1)\sigma_\varepsilon^2 + \sigma_\eta^2}\mu_a \tag{8}$$

This expression will be useful in Section 6.2. ■

Equation (8) 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 - as if it has rational expectations. 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.

Example 5.1 suggests that linear-normal specifications may give rise to unbiased forecasts, even when the agent's subjective DAG is imperfect. Let us now elaborate on this observation. Return to the 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 $\alpha_{ij} \neq 0$, 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.* Note that this is *not* a generalization of Example 5.1, because the latter did not require the central bank's random strategy to be normally distributed.

Proposition 6 *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. 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 $j < i$. Therefore, we can write

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

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 product of some matrix and 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,

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

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

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

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 significantly weaker than perfection (recall that in a perfect DAG, *every* node can be regarded as ancestral). The subjective DAG in Section 3 satisfies this weaker property. Under this restriction, the agent's forecasts are unbiased when the objective distribution is generated by a linear-normal model.

6 Conditional Forecast Errors

So far, the question we addressed was whether the agent's forecasts of economic variables are unbiased *on average*. Indeed, in our running "monetary policy" example, this is all that mattered because the central bank's sole objective was to maximize expected output. However, for many purposes, it also matters whether the agent's conditional forecasts are consistent with rational expectations for *all* realizations of a. The following is a sufficient condition for this stronger requirement to hold for a given variable.

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 $N - L_R(i)$ is perfect and contains both 0 and i. Because $\{0, i\}$ is a clique in the subgraph, perfection implies that we can regard it as ancestral. Therefore, $p_R(x_0, x_i)$ - and consequently $p_R(x_0)$ as well - are all unbiased, which immediately implies the result.

In the remainder of this section, I present two principal-agent 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.

6.1 Monopoly Pricing

This sub-section is a variation on the "Dieter's Dilemma" example of Spiegler (2015a). A monopolistic firm facing one consumer produces a food supplement at a constant marginal cost $k > 0$. The firm's action a takes values in $\{0, 1\}$, where $a = 1$ means that the firm sells the supplement. There are two other relevant variables: the consumer's health (denoted h), and the level of some chemical in his blood (denoted c). Both c and h take values in $\{0, 1\}$, where $h = 1$ means that the consumer is in good health, and $c = 1$ means that the chemical's level is abnormal. According to the true process, $p(h = 1) = \frac{1}{2}$, independently of a, and c is a deterministic consequence of a and h given by $c = (1 - a)(1 - h)$. The true process is thus consistent with a "true DAG" $a \rightarrow c \leftarrow h$.

The assumption that h is independent of a implies that if consumers had rational expectations, their willingness to pay for the supplement would be zero. Now suppose that the consumer's DAG is $R : a \rightarrow c \rightarrow h$. This DAG reverses the direction of causation between h and c relative to the true DAG.

Because the consumer'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.

Because the firm has monopoly power, it fully extracts the consumer's willingness to pay for the supplement, which is $p_R(h = 1 \mid a = 1) - p_R(h = 1 \mid a = 0)$. The firm commits to a mixture over actions, which is interpreted as the long-run frequency with which it sells the supplement. Its objective is to maximize the total profit

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

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

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

for every $a = 0, 1$, such that the consumer's willingness to pay for the supplement is $1/2(1 + \alpha)$ - note that it decreasing in the selling frequency. The firm'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 firm is unable to profit from the consumer'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 consumer's willingness to pay for the supplement is $\frac{1}{4}$ for $k \leq \frac{1}{8}$, and $\sqrt{2k}(1 - \sqrt{2k})/2$ for $k \in (\frac{1}{8}, \frac{1}{2})$.

At first glance, the comparative statics w.r.t k depicts a conventional response to changes in marginal cost: as k goes down, the firm sells a greater total quantity at a lower price. Normally, we would interpret this response as sliding down a downward sloping demand curve. However, the logic is different here. A decrease in k increases the firm's incentive to produce; a larger selling frequency leads to a lower endogenous willingness to pay for the supplement, and therefore the firm needs to lower the price.

6.2 Rigid Monetary Policy

For the last time in this paper, let us revisit the “monetary policy”, adopting the linear-normal specification of Example 5.1. 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.

Extend the basic example by adding an exogenous variable θ, which the central bank privately observes θ before taking its action. This variable represents the inflation target that the central bank would like to implement. The other two economic variables, π and y, are independent of θ conditional on a. In particular, they obey the linear-normal equations (6). No structure is imposed on the distribution of θ. The true process is consistent with the true DAG

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

Let μ_z denote the true expected value of any variable z.

The central bank’s objective is to minimize

$$Var(y) + k \cdot E(\pi - \theta)^2 \tag{10}$$

where $k > 0$ is a constant that captures the central bank’s trade-off between 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,

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

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 of ε and η.

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 former corresponds to a policy that is more responsive to fluctuations in the target. As γ approaches 1 - such that anticipated inflation matters less for output - the central bank's policy approaches perfect targeting.

The private sector's DAG is

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

Thus, the private sector's causal model agrees with the true model about the way θ and a are jointly distributed; the only disagreement is about the way output and inflation are determined, along the same lines as in Section 3.

Proposition 7 *Given the private sector's DAG R, 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 (10) subject to the constraints that $\pi = a+\varepsilon$ and $y = \gamma\pi - e(a) + \eta$. (Recall the notation $e(a) = E_R(\pi \mid a)$.)

In Section 5, we saw that $e(a)$ is given by (8). Because the true process in the current example has the feature that $\pi, y \perp \theta \mid a$, the same expression for $e(a)$ continues to hold. Therefore,

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

where

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

Thus, $\mu_y = (\gamma - 1)\mu_a$.

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

$$(\gamma - \delta)^2 E(a - \mu_a)^2 + kE(a - \theta)^2 \tag{11}$$

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 5, the expression for $e(a)$ given by (8) implies that when $\gamma = 1$, the private sector's inflation forecasts are consistent with rational expectations, hence the optimal policy fully tracks θ. Deviations from the rational-expectations prediction occur 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 the central bank's action. 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 (relative to the reliability of the effect of monetary policy on inflation), 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 (11). 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 $\sigma_\varepsilon^2/\sigma_\eta^2$.

7 Discussion

In this section I briefly discuss a few variations and extensions of the model, as well as the paper's relation to some works on non-rational expectations.

7.1 Ex-ante Forecasts

Throughout this paper, I assumed that the agent forms forecasts after observing a signal. A natural variant would assume that the agent forms his forecasts without observing anything. 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): 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 the agent conditions his forecast on a signal.

7.2 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 assumed that the central bank is able to commit ex-ante to a random 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 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 despite the perfect monitoring of the central bank's actions. Suppose that $R : a \rightarrow y \leftarrow \pi$. By our analysis in Section 4, the private sector's inflation forecast is correct 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, principal-agent situations are vulnerable to a time-consistency problem when the agent has causal misperceptions, even if he perfectly monitors the principal's move.

7.3 Relevance to Dynamic Models

The model of this paper 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 current and lagged variables - represents structural assumptions regarding this Markov process.

7.4 Related Literature

This paper contributes to a 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 "expectations management" aspect of the examples in 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.

The paper is also 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.

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), "Data Monkies: A Model of Naive Extrapolation from Partial Statistics," 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.