Behavioral Implications of Causal Misperceptions[*]

Ran Spiegler[†]

December 5, 2019

Abstract

This review presents an approach to modeling decision making under misspecified subjective models. The approach is based on the idea that decision makers impose subjective causal interpretations on observed correlations, and borrows basic concepts and tools from the Statistics/AI literature on Bayesian Networks. While this background literature used Bayesian networks as a platform for normative and computational analysis of probabilistic and causal inference, here graphical models represent causal misperceptions and help analyzing their behavioral implications. I show how this approach sheds light on earlier equilibrium models with non-rational expectations, and demonstrate its scope of economic applications.

[*] Financial support by ERC Advanced Investigator grant no. 692995 is gratefully acknowledged. I thank Yotam Alexander, Daniel Barreto, Ignacio Esponda, Philippe Jehiel, Stephen Morris, Ariel Rubinstein and Heidi Thysen for helpful comments.

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

1 Introduction

Few social-science mottos are more familiar than "*correlation does not imply causation*". The motto even has its own Wikipedia entry. Yet its very popularity betrays how common it is for people to draw rash causal conclusions from observational data. E.g., observing a correlation between parenting style and children's personality, people instinctively jump into the conclusion that the former causes the latter - even though parental behavior could be a *reaction* to the child's temperament (see Harris (1998)). Likewise, prospective students are likely to interpret the correlation between academic degrees and future salaries as a causal effect, underplaying unobserved personal characteristics that partly explain this correlation. Finally, many of us exhibit the so-called "illusion of control" and attribute other people's behavior to our own actions rather than to less salient causes that lie outside our control.

Yet confusion between correlation and causation is not exclusively practiced by laymen. In fact, it occurs regularly on the pages of economic-theory papers. Consider a simple model of monopoly pricing. The seller determines the price x of his product, after receiving a partially informative signal s regarding an underlying demand parameter θ. The sold quantity is q. The seller chooses x to maximize his profit $\pi(x, q)$ with respect to the conditional probability

$$\Pr(q \mid s, x) = \sum_{\theta} \Pr(\theta \mid s) \cdot \Pr(q \mid x, \theta)$$

Each of the two terms of the R.H.S summand is a conditional probability - i.e. a description of some correlation between two variables. However, the two correlations have different causal meaning. The first term is *diagnostic*, answering the question "what is the underlying demand θ given the signal s?" In contrast, the second term answers a *causal* question: "What is the effect of the price x on sold quantity q, given the underlying demand θ?"

When economic theorists analyze such a model, they treat all of these terms as conditional probabilities, giving no thought to the causal/diagnostic distinction. In this sense, they practically "confuse correlation with causation". The reason they can afford to do so is threefold. First, standard analysis assumes that agents reason in terms of a correctly specified model. In particular, the above conditioning procedure mirrors the interaction's underlying causal structure, which is intuitively represented by the graph

$$\begin{tikzcd} \theta \arrow[r] \arrow[dr] & s \arrow[r] & x \arrow[dl] \\ & q & \end{tikzcd} \tag{1}$$

Second, the conditioning procedure tacitly follows sound rules for causal interpretation of probabilistic information (codified in Pearl (2009) - see Section 3.2). Finally, standard analysis presumes that the seller knows the prior joint distribution over all relevant variables - i.e. there are no "hidden" variables. This enables him to obtain a precise estimate of any conditional probability that plays a role in his reasoning (specifically, the terms $\Pr(\theta \mid s)$ and $\Pr(q \mid x, \theta)$).

When these conditions fail - as they do in many real-life situations - we can no longer afford to be sanguine about agents' causal reasoning. In particular, we need to address four questions: Does the decision maker (**DM**) interpret empirical regularities in his environment through the prism of a subjective causal model? Is this model correctly specified? Does the observational data at the DM's disposal enable the quantification of causal effects (especially the effect of his own actions on payoff-relevant consequences)? Which methods for drawing such causal inferences does he employ?

In this review I present a framework for modeling imperfect causal reasoning in decision-making contexts. This framework borrows the notion of a causal model - represented by a directed acyclic graph (DAG), of which (1) is an example - from the literature on graphical models that lies at the intersection of Statistics and Artificial Intelligence. Historically, DAGs provided a platform for automating probabilistic inference (Pearl (2014), Cowell et al. (1999), Koller and Friedman (2009)) and later causal inference (Pearl 2009). Recently, econometricians have debated the relevance of graphical causal models for empirical work in economics (Heckman and Pinto (2015), Imbens (2019)). In contrast, I adapt the DAG formalism to study *flawed* causal reasoning: DAGs will represent errors of causal attribution and provide tools for studying their behavioral implications. Thus, the same apparatus that undergirds *normative* probabilistic and causal reasoning in the Statistics/AI literature is here put to use for *descriptive* modeling of causal reasoning by boundedly rational agents. In this sense, my approach is closer in spirit to psychologists' use of graphical models for expressing aspects of intuitive, everyday causal reasoning (Sloman (2005)). The formalism also sheds light on earlier notions of "equilibrium with non-rational expectations". From this perspective, this review offers a partial synthesis of models of non-rational expectations through the lens of causal misperceptions.

2 An Example: The Dieter's Dilemma

The following example, adapted from Spiegler (2016), illustrates the modeling framework. A consumer who wishes to improve his health considers buying a nutritional supplement at a cost $k > 0$. Denote $a = 1$ if he buys the supplement and $a = 0$ otherwise. In reality, the supplement has no effect on the consumer's health: He is healthy (an outcome denoted $h = 1$) or unhealthy (denoted $h = 0$) with equal probability, independently of whether he consumes the supplement. The consumer's payoff is $h - ka$. Therefore, a consumer with rational expectations would not buy the supplement.

Now suppose that the binary variables a and h are correlated with a third, payoff-irrelevant binary variable c, which indicates the level of some chemical in the consumer's blood: $c = 0$ means that the chemical level is normal. The realization of c is a deterministic function of a and h , $c = (1-a)(1-h)$. That is, the consumer's chemical level is abnormal if and only if he is unhealthy *and* fails to take the supplement.

Although this is a one-shot decision situation, it is helpful to think of the consumer as taking his decision after many generations of identical consumers faced the same situation; a fraction α of this population chose $a = 1$. We can thus describe the historical realizations of a, h, c by a joint probability distribution p over (a, h, c): $p(a = 1) = \alpha$; $p(h = 1) = \frac{1}{2}$ independently of a; and $p(c = (1 - a)(1 - h) \mid a, h) = 1$ for every a, h.

Our consumer believes that this distribution has an underlying causal structure, which is represented by the following *directed acyclic graph* (**DAG**) $G : a \to c \to h$. That is, the consumer believes that his consumption decision affects his blood chemical level, which in turn is the sole direct cause of health. This causal belief is purely qualitative; it is entirely silent about the sign or magnitude of the causal effects.

How does the consumer turn this qualitative model into a quantitative belief? He simply measures the historical correlations between the variables he deems causally related and combines them in accordance with his causal model. Specifically, he measures the conditional probabilities $(p(c \mid a))$ and $(p(h \mid c))$, such that his subjective belief regarding the health consequences of his consumption decision is

$$p_G(h \mid a) = \sum_c p(c \mid a)p(h \mid c) \tag{2}$$

where p_G denotes the consumer's subjective belief; the notation conveys the idea of an objective distribution p viewed through the prism of the subjective causal model G. The consumer's causal model focuses his attention on particular correlations and instructs him how to combine them.

If the consumer's causal model were correct, it would be legitimate to write down $p(h \mid a)$ exactly as (2) and to interpret it as an estimated causal effect. The consumer's only error is that the objective distribution p is inconsistent with his subjective model causal. Indeed, it is consistent with a different causal model, according to which a and h are independent causes of c. This model can be represented by the DAG $a \to c \leftarrow h$. The consumer's subjective causal model deviates from this "true model" by *inverting* the link between c and h. We can thus regard the consumer's causal misperception as an instance of *reverse causality.* This is an illustration of the graphical language's ability to represent common errors of causal attribution.

Having defined the consumer's belief, let us turn to describing his behavior. As an expected-utility maximizer, he should choose the action a that maximizes $p_G(h = 1 \mid a) - ka$. However, note that $p_G(h \mid a)$ is ill-defined without knowledge of α - namely, the consumption frequency in the population. The reason is that the definition of $p_G(h \mid a)$ contains the term $p(h \mid c)$. This term involves no explicit conditioning on a because the consumer believes that c is the sole direct cause of h. However, by the specification of p, h and a are *not* independent conditional on c. (To see why, suppose we know that the consumer's chemical level is normal. If, on top of that, we learn that he has not taken the food supplement, we must infer he is in good health.) Indeed,

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

whereas

$$p(h = 1 \mid c = 1) = \frac{p(h = 1)p(c = 1 \mid h = 1)}{p(a = 0)p(h = 0)} = 0$$

Therefore, a change in α leads to a change in $(p(h \mid c))$, which implies a change in $p_G(h = 1 \mid a)$ and consequently a possible change in the subjective optimality of any given action.

This suggests that although we are dealing with single-agent decision making, a coherent description of the consumer's behavior requires an *equilibrium* definition. Accordingly, when $\alpha \in (0, 1)$, define it as a "personal equilibrium" if both actions $a = 0, 1$ maximize $p_G(h = 1 \mid a) - ka$. This implies the indifference condition

$$p_G(h = 1 \mid a = 1) - p_G(h = 1 \mid a = 0) = k \tag{3}$$

This notion of equilibrium captures a steady state in the population of ex-ante identical consumers who face the same Dieter's Dilemma.

To find the equilibrium, we need to calculate the relevant terms in (3):

$$\begin{aligned} p_G(h &= 1 \mid a = 1) \\ &= p(c = 1 \mid a = 1)p(h = 1 \mid c = 1) + p(c = 0 \mid a = 1)p(h = 1 \mid c = 0) \\ &= 1 \cdot \frac{1}{1 + \alpha} + 0 \cdot 0 = \frac{1}{1 + \alpha} \end{aligned}$$

and

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

It follows that when $k \in (\frac{1}{4}, \frac{1}{2})$, the unique personal equilibrium is $\alpha = (1 - 2k)/2k$.[1]

Thus, the consumer's reverse-causality error leads him to take the suboptimal action of consuming a useless product with positive probability. The prediction of a unique mixed strategy is an unmistakable *equilibrium* effect; it would never arise under conventional, single-agent expected-utility maximization. This effect is a hallmark of decision making under a misspecified model; it has precedents in the literature, which I will discuss later.

We can draw several lessons from this example. First, it suggests a language for systematic description of a DM with causal misperceptions: The DM forms a subjective belief by fitting a subjective causal model - represented by a DAG - to objective data. Second, the graphical representation of subjective causal models can capture errors of causal attribution such as reverse causality. Third, coherent description of decision making by an agent with a misspecified causal model requires an equilibrium approach. The next sections will develop these themes.

3 A Modeling Framework

An economic environment is described by a collection of variables $x_1, ..., x_n$. For every subset $B \subseteq \{1, ..., n\}$, denote $x_B = (x_i)_{i \in B}$. A steady state in this environment is described by an objective joint probability distribution p over these variables, with full support. A DM perceives this objective distribution through the prism of a *subjective causal model*. Formally, a causal model is

[1] When k it outside these bounds, personal equilibrium does not involve mixing over actions.

a DAG $G = (N, R)$, where $N \subseteq \{1, ..., n\}$ is the set of nodes and R is the set of directed links. For any node $i \in N$, let $R(i) \subset N$ denote the set of nodes $j \in N$ for which there is a link $j \rightarrow i$. A node corresponds to a variable, whereas a link represents a perceived direct causal relation. Directedness and acyclicity of the graph capture fundamental intuitions about causal relations.

Thus, the DAG (N, R) represents the DM's perception of qualitative causal relations in his environment. He quantifies them by estimating the causal relations - i.e., by $fitting$ his causal model to selective data extracted from the objective distribution. Formally, the DM *factorizes* p according to his DAG, as given by the following formula:

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

E.g., the DAG $G : 1 \rightarrow 3 \rightarrow 4 \leftarrow 2$ induces the belief

$$p_G(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3 \mid x_1)p(x_4 \mid x_2, x_3)$$

The full-support assumption ensures that the factorization formula is well-defined.

Formula (4) is familiar from the *Bayesian networks* literature, where a Bayesian network is defined by a DAG and a probability distribution that is consistent with the DAG in the sense that it is given by the R.H.S of (4). In that literature (e.g. Pearl (2014), Cowell et al. (1999), Koller and Friedman (2009)), this object provides a platform for computationally efficient probabilistic inferences. In addition, Bayesian networks are objects with rich mathematical structure that enables us to draw links between the structure of G and conditional-independence properties of p_G.

In the present context, the DAG G encodes a *belief distortion function*, which maps any objective distribution p to a subjective belief p_G. It captures the idea that the DM perceives empirical regularities through the prism of a subjective causal model. The subjective belief p_G is only defined over the variables that his causal model includes, and systematically distorts the objective distribution over these variables. The R.H.S of (4) describes this distortion. The view of (4) as a belief-distortion function is a novel feature of the present project, and (to my knowledge) does not appear in the Statistics/AI sources on which it draws. As we shall later see, this novel interpretation suggests new technical questions.

When the DAG is fully connected (any two nodes are directly linked), p_G is reduced to a standard chain rule, such that $p_G = p$ - i.e., the DM's subjective belief does not distort the objective distribution. At the other extreme, when the DAG is empty (it has no links), $p_G(x_N)$ is the product of the marginals $p(x_i)$ over all $i \in N$ - i.e., the DM perceives all variables to be independent.

The object p_G is a subjective belief to all intents and purposes. When the DM is interested in the marginal of p_G over some collection of variables x_M, he marginalizes it: $p_G(x_M) = \sum_{x_{N-M}} p_G(x_M, x_{N-M})$. Likewise, when the DM learns the value of some collection of variables x_M, he conditions his subjective belief on this information in accordance with standard Bayesian updating: $p_G(x_{N-M} \mid x_M) = p_G(x_M, x_{N-M})/p_G(x_M)$.

For a concrete image for this belief formation model, imagine the DM as a researcher who perceives the economic environment through the prism of a structural causal model, and asks a research assistant to estimate the structural equations of his model. The output of this process is an estimated model. More broadly, the mindset that (4) captures is that of someone who seeks statistical data to fit and quantify a prior model, not to test it.

3.1 Belief Errors as Causal Misperceptions

The simple model of belief formation presented above offers a language for describing systematic belief errors distilled from everyday observation and psychological research. In each case, the objective distribution p is consistent with a "true DAG" G^* and the DM's subjective DAG is $G \neq G^*$. The departure of G from G^* constitutes the DM's causal misperception, and the departure of p_G from $p_{G^*} = p$ is its resulting belief error. Section 2 already illustrated how to express a reverse-causality error in this language. This sub-section presents a few additional examples.

False belief in causality

Perhaps the most basic causal misperception is the very belief that empirical regularities have a causal basis. Indeed, many economic theories (e.g. competitive equilibrium) are fundamentally non-causal because they assume bidirectional influences among variables. When the steady state of an economic system is consistent with such a non-causal theory and yet the DM believes there is an underlying causal mechanism, he commits a belief error. How can we represent it with our formalism?

Recall that *every* objective distribution p is consistent with any fully connected DAG. Therefore, taken literally, the claim that a given distribution p is inconsistent with a causal model can never be correct. It only becomes meaningful if we require p to be consistent with a *non-trivial* causal model, in the sense that it has at least one missing link. Accordingly, say that p is inconsistent with a non-trivial causal model if the fully connected DAGs are the *only* DAGs with which it is consistent. In this case, if the DM's subjective DAG G is not fully connected, then $p_G \neq p$; and the DM's error is that he erroneously imposes *some* causal structure on his environment.

Fundamental attribution error

According to this idea from Social Psychology (Ross (1977)), people underestimate the situational nature of other people's behavior and tend to attribute it to fixed character rather than external circumstances. To capture this error in our framework, consider the true DAG $G^* : \theta_1 \rightarrow b \leftarrow \theta_2$, where b represents the behavior of some agent (other than our DM), while θ_1 and θ_2 are two independent variables that jointly describe external circumstances shaping the agent's behavior. Our DM's subjective DAG G departs from G^* by omitting the link $b \leftarrow \theta_2$. Then,

$$p_G(\theta_1, \theta_2, b) = p(\theta_1)p(\theta_2)p(b \mid \theta_1)$$

whereas the correct joint distribution is $p(\theta_1, \theta_2, b) = p(\theta_1)p(\theta_2)p(b \mid \theta_1, \theta_2)$.

If the DM learns (θ_1, θ_2), his subjective belief over b will be

$$p_G(b \mid \theta_1) = \sum_{\theta_2} p(\theta_2) p(b \mid \theta_1, \theta_2)$$

whereas the correct updated belief would be $p(b \mid \theta_1, \theta_2)$. That is, the DM neglects θ_2 and over-attributes the other agent's behavior to θ_1.[2]

Ignoring confounders
Suppose that the true DAG G^* is

$$\begin{tikzcd} & \theta \arrow[d] \arrow[dr] & \\ a \arrow[r] & y \arrow[r] & z \end{tikzcd} \tag{5}$$

where the node a represents the DM's action, y represents an intermediate outcome, z represents a final outcome and θ is an exogenous variable that is independent of a. The DM's subjective DAG is $G : a \to y \to z$. That is, he neglects the exogenous variable θ, such that his account of the relation between y and z exhibits an "omitted variable error". Then,

$$p_G(z \mid a) = \sum_{y} p(y \mid a) p(z \mid y)$$

whereas the true distribution of z conditional on a is

$$p(z \mid a) = \sum_{\theta} p(\theta) \sum_{y} p(y \mid a, \theta) p(z \mid y, \theta)$$

The DM's error is that he regards the correlation between y and z as a pure causal effect of y on z, whereas in reality part of the correlation is due to the confounding variable θ.

This error arguably lies behind most real-life manifestations of "confusing correlation with causation". Avoiding it is a primary concern of practitioners and teachers of causal inference (Angrist and Pischke (2008), Pearl and

[2] Ettinger and Jehiel (2010) formalize the Fundamental Attribution Error along similar lines, using the framework of analogy-based expectations (see Section 3.4).

Mackenzie (2018)). The Bayesian-network formalism enables us to describe this error and its resulting belief distortions.

Illusion of control

Langer (1975) coined the term "illusion of control" to describe a belief bias that exaggerates one's role in determining observed outcomes. To capture this error, let the true DAG be $G^* : a \leftarrow \theta \rightarrow y$, where a represents the DM's action, θ is an exogenous variable and y is an outcome. The DM's subjective DAG is $a \rightarrow y$. This is another example of confounder neglect. In reality, the DM's action and the final outcome are correlated only because they are conditionally independent consequences of the exogenous variables θ. However, the DM interprets this correlation as a causal effect of his action on the final outcome.

Neglecting indirect effects

The effect of actions on target variables often works through multiple causal channels. The following example captures a situation in which the DM fails to take all these channels into account. The true DAG G^* is

$$\begin{tikzcd}
a \arrow[r] \arrow[dr] & q_1 \arrow[r] & y \\
 & q_2 \arrow[ur] &
\end{tikzcd} \tag{6}$$

where q_1 and q_2 mediate the causal effect of a on y. The DM's subjective DAG is $G : a \rightarrow q_1 \rightarrow y$ - i.e., he neglects the channel that passes through q_2, possibly because he is unaware of this variable.

An example of this misperception that is common in discussions of economic policy is the neglect of equilibrium effects. For example, a is a policy maker's action; q_1 is a variable that the action directly impacts; q_2 represents a behavioral response to the action that results from changing incentives; and the policy maker neglects this response when evaluating the policy.

3.2 From Beliefs to Decisions

How does a DM who forms his beliefs according to (4) make decisions? The most conventional assumption would be that he maximizes expected utility with respect to p_G. Accordingly, endow the DM with the vNM utility function $u(x_1, ..., x_n)$. Suppose u is measurable with respect to x_N - otherwise, the DM's model omits variables that are essential for describing his preferences. Identify one of the n variables with the DM's action a. In addition, the DM may observe the realization of some collection of variables x_I, which enables him to condition his belief on this information. The DM chooses a to maximize

$$\sum_{x_N} p_G(x_N \mid a, x_I) u(x_N) \tag{7}$$

This innocent-looking formula conceals two subtleties, which I now discuss.

Confusing correlation with causation?

In the Introduction, I listed a number of ways in which the DM's causal reasoning can go astray. First, his perception of causal relations in his environment may be wrong. This type of error was our focus in Section 3.1, and it will continue to be at the heart of this review. Second, some of the variables in his causal model may be unobservable, which means that he is unable to estimate some of the causal relations he postulates *directly* from observational data. In this review I mostly abstract from this difficulty. Throughout this section and the next, every variable in the DM's causal model is observable.

Finally, there is the question of whether the DM draws correct causal inferences from observational data, taking his subjective causal model *as given*. Here the answer is quite subtle. In principle, the expected-utility maximization in (7) is performed with respect to a conditional probability distribution, $p_G(x_N \mid a, x_I)$, without any explicit attempt to disentangle its causal and diagnostic aspects - the same practice I described in the Introduction. Yet $p_G(x_N \mid a, x_I)$ need not represent the causal effect of a on other payoff-relevant variables (fixing x_I), even if we accept G as a correct causal model. In other words, the DM's procedure appears to "confuse correlation with causation".

In this review I sidestep this potential error of causal inference. In all the applications I will consider, the arrangement of the nodes that represent a and x_I in G is such that the DM's causal inference is sound. Specifically, Pearl's (2009) "do-calculus", which codifies correct causal inference from observational data given a DAG-represented causal model, would legitimize the interpretation of $p_G(x_N \mid a, x_I)$ as the causal effect of a (given x_I) on other variables. In particular, applications in which the DM is uninformed (such that there are no x_I variables) will invariably assume the node that represents a in G is ancestral - as in the Dieter's Dilemma. In this case, $p_G(x_N \mid a)$ would capture the true causal effect of a on other variables, if G were a correct model.

Presenting Pearl's machinery here would take a lot of space, and it would be irrelevant in the sense that the research described in this review has made no use of do-calculus. Therefore, I can only assure the reader that the DM's causal inferences in all the examples I present is sound *given* his subjective model; his sole error is that the model is wrong. In its study of imperfect causal reasoning, this review focuses on causal *misperceptions* - i.e. the error of having a misspecified causal model - while effectively ruling out errors of causal *inference* given the model. I will return to this issue in the concluding section.

Personal equilibrium

The other subtlety in (7) was pointed out in Section 2: The subjective conditional probability $p_G(x_N \mid a, x_I)$ need not be invariant to the objective conditional distribution $(p(a \mid x_I))$. The latter distribution formally describes the DM's strategy, interpreted as the long-run behavior of previous generations of DMs facing the same problem. This dependency would never arise under rational expectations: If the DM's causal model G were correctly specified, it would imply $p_G = p$, hence $p_G(x_N \mid a, x_I)$ would be invariant to $p(a \mid x_I)$ *by definition*. In contrast, the Dieter's Dilemma describes a situation in which the DM's subjective belief $p_G(h \mid a)$ varies with the long-run behavior given by $(p(a))$.

Thus, if the DM's long run behavior changed, so could his subjectively optimal decision. This suggests a need to define subjectively optimal behavior as an *equilibrium object*. A fully mixed strategy $(p(a \mid x_I))$ is a personal ε-equilibrium if, whenever $p(a \mid x_I) > \varepsilon$, a maximizes (7). A personal equilibrium is a strategy (not necessarily mixed) that is the limit of a sequence of personal ε-equilibria, where $\varepsilon \to 0$ along the sequence. I interpret trembles as blind experimentation. The need to introduce them arises because the terms in (4) are not given but *derived* from the joint distribution p, and therefore zero-probability events need to be ruled out.[3]

A personal equilibrium describes a steady state in which DMs take actions that are subjectively optimal, given the distorted belief that arises from fitting their subjective causal model to the steady-state distribution. The need to describe individual behavior in equilibrium terms is a more general feature of models of decision making under misspecified subjective models. Such equilibrium effects are *not* game-theoretic, because they arise even in *single-agent* decision problems.

Extending the model to interactive situations is straightforward. Consider a static game with incomplete information. From a player's point of view, the situation is the same as in the single-agent formulation, except that now some of the variables represent the characteristics and actions of other players, and possibly consequences of players' actions (e.g. final allocation in an auction). The definition of equilibrium is essentially the same: each player chooses subjectively optimal actions with respect to his belief, which results from fitting his subjective causal model to the equilibrium distribution.

3.3 Some Basic Bayesian-Network Tools

This review is obviously not the place for a proper introduction to the technical literature on Bayesian networks. For general references that contain relevant material, I refer the reader to Cowell et al. (1999), Pearl (2009) or Koller and Friedman (2009). Still, it will be helpful to describe briefly the *kind* of Bayesian-network tools that can be put to use by economic theorists studying behavioral implications of causal misperceptions. This is my task in this sub-section.

Spiegler (2016, 2017 and 2018) contain succinct, economic-theory-friendly adaptations of relevant material. These papers also contain technical results that are appear to be new (as stated) to the literature on graphical proba-

[3] The term "personal equilibrium" was coined by Kőszegi (2010) in the context of decision making under belief-dependent utility.

bilistic models - although a likely reason is that Bayesian-network specialists in Statistics or AI would probably find these results irrelevant or too obvious to highlight. This difference in perspective arises from the present framework's novel view of (4) as a belief-distortion function, which is essential for descriptive modeling of a DM with a misspecified causal model but seems orthogonal to what statisticians or computer scientists care about.

Equivalent DAGs

While I have emphasized the interpretation of DAGs as representations of subjective causal models, they can be viewed as mere representations of conditional-independence properties. If we only consider the belief distortion function p_G as such and ignore the causal interpretation of $p_G(x_N \mid a, x_I)$, this is an appropriate point of view. For instance, $1 \rightarrow 2 \rightarrow 3$ represents all the joint distributions that satisfy $x_1 \perp x_3 \mid x_2$. Different-looking DAGs represent different causal perceptions but they can be equivalent in terms of the conditional-independence properties they represent, and therefore in terms of the belief-distortion function they encode. The simplest example involves the DAGs $1 \rightarrow 2$ and $2 \rightarrow 1$. These DAGs clearly represent distinct causal models, yet they are equivalent in the sense that

$$p_{1\rightarrow 2}(x_1, x_2) \equiv p(x_1)p(x_2 \mid x_1) \equiv p(x_2)p(x_1 \mid x_2) \equiv p_{2\rightarrow 1}(x_1, x_2)$$

by the basic conditional-probability identity. Likewise, the DAGs $1 \rightarrow 2 \rightarrow 3$, $1 \leftarrow 2 \rightarrow 3$ and $1 \leftarrow 2 \leftarrow 3$ are all equivalent, in the sense that they generate the same belief-distortion function p_G.

Frydenberg (1990) and Verma and Pearl (1991) provided a complete characterization of this equivalence relation.[4] Two DAGs are equivalent if they have the same "skeleton" (i.e., they look identical if we ignore the direction of links) and the same "v-structure" (i.e., the same collection of triples of nodes i, j, k such that $i \rightarrow k \leftarrow j$ and yet i and j are not linked). E.g., $G : 1 \rightarrow 2 \rightarrow 3$ and $G' : 1 \rightarrow 2 \leftarrow 3$ are not equivalent: even though they

[4] Verma and Pearl defined DAG equivalence in terms of the classes of distributions that are consistent with each DAG. However, the result is the same for my notion of equivalence, which concerns the belief-distortion function associated with each DAG.

have the same skeleton, they have different v-structures. This means that $p_G \neq p_{G'}$ for some objective distribution p.

This characterization is valuable because it visualizes the equivalence relation. In particular, it facilitates checking whether a variable can be identified with an ancestral node in some DAG in the model's equivalence class. This is helpful, thanks to a property that is easy to glean from (4): If i is an ancestral node in G, then p_G never distorts the marginal distribution of x_i - i.e., $p_G(x_i) \equiv p(x_i)$.

The latter property is relevant in various contexts. For example, in many economic models, a DM's behavior is driven by an estimate of a particular variable x_i. The DM forms this estimate after observing the realization of another variable x_j. A key question is whether this conditional estimate is biased on average - i.e. whether

$$\sum_{x_j} p(x_j) p_G(x_i \mid x_j) \equiv p(x_i) \tag{8}$$

This is simply the "Bayes plausibility" property of conditional beliefs, which of course always holds under rational expectations. When p_G violates this property, the DM's beliefs of x_i are biased on average - e.g., product reviews may lead the consumer to systematically overestimate product quality. Spiegler (2018) provides a characterization of DAGs that satisfy the Bayes-plausibility property for any given pair of variables x_i, x_j. In particular, when both i and j are ancestral nodes in some DAG in the equivalence class of G, we can make a sequence of substitutions:

$$\sum_{x_j} p(x_j) p_G(x_i \mid x_j) \equiv \sum_{x_j} p_G(x_j) p_G(x_i \mid x_j) \equiv p_G(x_i) \equiv p(x_i)$$

such that the DM's conditional estimate of x_i is unbiased on average.

Perfect DAGs

A DAG with an empty v-structure is known as a perfect DAG. In a perfect DAG, "all parents are married": If the DAG contains the links $i \rightarrow k \leftarrow j$, then i and j must be directly linked. The causal interpretation is that if the

DM perceives two variables as direct causes of a third variable, he must also perceive a direct causal link between them.

Perfect DAGs have special properties. The equivalent-DAG relation implies that the direction of links in a perfect DAG is irrelevant, in the following sense: If a perfect DAG contains the link $i \rightarrow j$, there is an equivalent DAG that contains the inverse link $j \rightarrow i$. Relatedly, for any node in a perfect DAG, there is an equivalent DAG in which the node is ancestral. As a result, a perfect DAG does not distort the marginal distribution of *any* individual variable. Spiegler (2018) establishes that perfection is both necessary and sufficient for this property: If G is imperfect, there exist p and i such that $p_G(x_i) \neq p(x_i)$ for some x_i.

Junction trees

When a DAG is perfect, it is possible to construct an auxiliary *tree* whose nodes correspond to the original DAG's maximal cliques.[5] Furthermore, this tree has the following property: For any pair of tree nodes C and C', $C \cap C'$ is contained in every node along the unique tree path that connects C and C'. This construction is known as a "*junction tree*" (Cowell et al. (1999)), and it is instrumental in efficient probabilistic-inference algorithms. In economic applications, the junction tree offers a convenient representation of certain conditional distributions that are derived from p_G. Eliaz et al. (2019) show that if G is perfect, then for any given pair of variables x_i, x_j we can construct a simple causal chain $G' : x_i \rightarrow \cdots \rightarrow x_j$, such that G' contains weakly fewer nodes than G, and $p_{G'}(x_j \mid x_i) \equiv p_G(x_j \mid x_i)$ - provided that we can freely redefine the intermediate variables along the chain.

d-Separation

As mentioned above, a DAG can be viewed as a representation of conditional-independence assumptions. The notion of d-separation (Geiger et al. (1990)) is a complete graphical characterization of these assumptions. This notion is based on a more fundamental (and non-trivial) definition of "path blocking". It enables a simple algorithm (expressed in purely graphical terms) that

[5] A clique is a collection of nodes such that every pair of nodes is linked. A clique is maximal if it is not contained by another clique.

checks any conditional independence property $x_A \perp x_B \mid x_C$, for any triple of subsets of nodes A, B, C. When a given triple in G passes this graphical test, p_G satisfies the conditional-independence property for any p.

The concept of d-separation helps analyzing the decision model of Section 3.2. Recall that in the Dieter's Dilemma, individual optimization under a misspecified subjective DAG led to a genuine equilibrium effect. This effect would not arise if the subjective action-consequence mapping were invariant to the DM's strategy. Spiegler (2016) shows how to define this notion of "consequentialist rationality" as a conditional-independence property that can be easily tested using d-separation. When the DM's DAG passes the test, his behavior will not exhibit equilibrium effects, no matter how we parameterize his decision problem. Spiegler (2018) uses d-separation to characterize when the Bayes plausibility property (8) holds for *given* $i, j \in N$.

Summary
In this sub-section, we briefly encountered basic Bayesian-network concepts: Equivalent DAGs, perfect DAGs, junction trees and d-separation. Each concept arrives with mathematical tools that can be put to use in economic applications. In turn, questions that arise naturally in economic applications give rise to new Bayesian-network results. We will get a closer look into the use of these tools when we delve into economic applications in Section 4.

3.4 Relation to other Approaches

A DM with a misspecified causal model forms beliefs that deviate from rational expectations, by systematically distorting the steady-state distribution of his environment. It is therefore instrumental to place the Bayesian-network formalism in the context of earlier equilibrium concepts with non-rational expectations. In this sub-section I discuss a few prominent examples in the literature, from the point of view of their links to the Bayesian-network language. As a result, the exposition is very idiosyncratic; the interested reader should consult the original sources for their authors' original perspective.

Although these concepts were originally defined in the context of games, I adopt the perspective of a single DM, to facilitate comparison with the model of Section 3.2. Throughout this sub-section, our DM will play a simultaneous-move game with Nature. The DM's action is a, Nature's move is θ, and z is a consequence variable.

Analogy-based expectations

This concept was introduced by Jehiel (2005) for extensive-form games, and was later adapted to static Bayesian games by Jehiel and Koessler (2008). In both cases, the idea is that players partition contingencies into "analogy classes" and perceive endogenous variables of interest as a function of this analogy partition.

The following example illustrates how specifications of analogy-based expectations in static models can be reformulated in the Bayesian-networks language. The true DAG G^* is $a \rightarrow z \leftarrow \theta \rightarrow e$, where e represents the "analogy class" to which the state of Nature θ belongs - i.e., e is a coarse-grained description of θ. For a DM with rational expectations, e is irrelevant and can be omitted from his subjective model altogether. Our DM's subjective DAG is $G : a \rightarrow z \leftarrow e \leftarrow \theta$ - i.e., G departs from G^* by replacing the link $z \leftarrow \theta$ with $z \leftarrow e$. The interpretation is that the DM has a coarse perception of the mapping from θ to z. The definition of Analogy-Based Expectations Equilibrium (ABEE) effectively requires the DM's action to be subjectively optimal with respect to p_G.

In general, individual best-replying under ABEE in static Bayesian games can be rewritten as subjective optimization with a misspecified DAG G. This is attained by adding the analogy class as a distinct variable, and reorienting links from exogenous to endogenous variables such that the analogy variable becomes the link's origin.

Cursed equilibrium

Eyster and Rabin (2005) study an equilibrium concept for Bayesian games, which is closely related to ABEE. Inspired by the "Winner's Curse" phenomenon observed in auctions, Eyster and Rabin define the notion of "cursed" beliefs, which captures a player's failure to realize that his opponents' behavior depends on factors beyond those he is informed of.

To illustrate the concept, let the true DAG G^* be $a \leftarrow s \leftarrow \theta \rightarrow z$, where s represents the DM's noisy signal of the state of Nature θ. The DM's subjective DAG G departs from G^* by replacing the link $\theta \to z$ with $s \to z$, such that

$$p_G(\theta, z \mid s) = p(\theta \mid s)p(z \mid s) = p(\theta \mid s)\left(\sum_{\theta'} p(\theta' \mid s)p(z \mid \theta')\right)$$

whereas rational expectations would prescribe $p(\theta, z \mid s) = p(\theta \mid s)p(z \mid \theta)$. Thus, the "cursed" DM effectively believes that $z \perp \theta \mid s$, while the true conditional-independence property is $z \perp s \mid \theta$. (These "cursed" beliefs can be also redefined in the language of ABEE, with a suitable analogy partition.)

Eyster and Rabin (2005) define "partially cursed" beliefs as a convex combination of $p_G(\theta, z \mid s)$ and $p(\theta, z \mid s)$. The convexification parameter measures the strength of the DM's belief bias. In the Bayesian-networks framework, this can be viewed as a "model averaging" procedure: The DM entertains two possible causal models; he fits each model to objective data and forms his belief by taking a weighted average of the estimated models.

Eyster and Rabin (2010) propose "high-order" elaborations of their approach in the context of an observational-learning model. For instance, they assume that each player believes that his opponents have "cursed" beliefs. This can be described as a game with non-common priors, where a player's type is defined by his subjective causal model as well as his belief regarding the causal models held by his opponents.

$S(K)$ equilibrium

Osborne and Rubinstein (1998) present a solution concept for static games, in which each player postulates a direct mapping from his action to the game's payoff-relevant outcome, without forming an explicit belief regarding his opponents' behavior. The player estimates this mapping by sampling each action K (independent) times, and selecting the action that performs best in his sample. The player's misperception can be described as follows. The true DAG G^* is $a \to z \leftarrow \theta$, and u is purely a function of z. The DM's subjective DAG G is $a \to z$. In the current framework, the DM's belief $p_G(z \mid a)$ would be consistent with rational expectations. However, this is because he perfectly learns $p(z \mid a)$, whereas in Osborne and Rubinstein (1998), he naively extrapolates from a *finite* sample.

However, it should be emphasized that the decision errors that arise in $S(K)$ equilibrium are shaped not only by sampling but also by the DM's misspecified causal model, which omits θ as an explanatory variable of z. To see why, consider a variant in which the DM holds the correct subjective model G^* and he uses sampling to evaluate the terms $p(\theta)$ and $p(z \mid a, \theta)$. This variant is behaviorally distinct from the $S(K)$ model. More generally, $S(K)$ equilibrium suggests a natural extension of the belief-formation model given by (4), in which the terms in the factorization formula are replaced with finite-sample estimates.

The value of reformulating these concepts in the Bayesian-network language is not "unification" for its own sake. Rather, the point of this exercise is that it deepens our understanding of these concepts, by highlighting their implicit element of causal misperception. It also equips us with new tools for analyzing their behavioral implications. For instance, Spiegler (2016) shows that when the subjective DAGs G and G' are not fully linked, neither uniformly dominates the other in terms of the DM's welfare. This result explains why refining a DM's analogy partition in Jehiel's framework need not be welfare-improving. Likewise, Spiegler's (2016) characterization of consequentialist rationality enables us to see when concepts like ABEE or cursed equilibrium generate equilibrium effects in single-agent decision problems.

Other equilibrium concepts in the literature have a looser connection to the Bayesian-network approach. Esponda (2008) introduced a solution concept, called "naive behavioral equilibrium", which assumes that the DM has a prior belief that the state of Nature and the opponents' actions are independent conditional on his signal (as in Cursed equilibrium). In addition, Esponda specifies a number of feedback variables, and requires the DM's belief of each of these variables conditional on his signal to be correct. While the conditional-independence belief has a DAG representation, the empirical-consistency requirement lacks a tight Bayesian-network counterpart.

What is common to all these concepts is that they describes equilibrium beliefs that systematically distort an objective equilibrium distribution.

Contrast this with the familiar notion of *subjective priors*. Under the latter approach, the DM has a fixed belief that is *independent* of the objective distribution.

Berk-Nash equilibrium

Esponda and Pouzo (2016) present a general approach to modeling equilibrium beliefs under misspecified subjective models. They formulate the DM's prior model as a class Q of conditional probability distributions that map actions (as well as signals, in the case of a partially informed DM) to observable consequences. The model is misspecified if Q does not include the true stochastic action-consequence mapping.

In this environment, Esponda and Pouzo define a notion of equilibrium, according to which the DM optimizes against a subjective conditional belief that is the closest in Q to the true mapping, where the notion of closeness is a weighted version of Kullback-Leibler divergence (the weights are determined by the DM's equilibrium strategy). Esponda and Pouzo dub this solution concept *Berk-Nash equilibrium*, after Berk's (1966) asymptotic characterization of Bayesian learning under misspecified priors. The key difference is that while Berk examined passive learning, Esponda and Pouzo examine the case that is relevant for decision making, in which the DM's actions influence the sample he draws inferences from.

Heidhues et al. (2018) study a model that can be described as an application of Berk-Nash equilibrium. This example can illustrate the similarities and differences between Berk-Nash equilibrium and the Bayesian-network approach. The DM's action a is a real number, and the consequence is $y = Q(a, b)$, where Q is a deterministic function and b is an independently distributed, real-valued variable that represents exogenous conditions. The DM has a misspecified model, which is given by a subjective function $\tilde{Q}$ which is different from the true function Q. For instance, $\tilde{Q}$ can capture over-confidence when a measures the DM's level of involvement in a project, and $\tilde{Q}(a) > Q(a)$ and $\partial\tilde{Q}(a)/\partial a > \partial Q(a)/\partial a$ for every a. Heidhues et al. (2018) characterize Berk-Nash equilibrium for various specifications of Q and $\tilde{Q}$. For instance, in the case of over-confidence, the DM's equilibrium involvement is excessively high and justified by an overly pessimistic belief regarding exogenous conditions.

This model cannot be cast in terms of the Bayesian-network formalism, because the misspecification in the DM's subjective model is *parametric* - i.e., it assigns zero probability to the correct parameter that governs the mapping from a to y. In contrast, the Bayesian-network formalism is manifestly non-parametric. Belief errors due to incorrect functional forms of $p(y \mid a)$ are outside its scope.

Can we nest the Bayesian-network model in the Berk-Nash approach? When a and x_I in (7) are represented by an ancestral clique in the DAG G (or in some equivalent DAG), the answer is affirmative. That is, we can define Q to be the set of all conditional distributions $q(x_N \mid a, x_I)$ that are consistent with G, such that the DM's belief in Berk-Nash equilibrium will be $p_G(x_N \mid a, x_I)$. In this sense, the Berk-Nash approach subsumes the Bayesian-network approach as a special case.

From a conceptual perspective, the two approaches are different because of their treatment of learning feedback and how it relates to the DM's subjective model. First, Berk-Nash equilibrium is primarily interpreted as the limit of a dynamic process of active Bayesian learning. In contrast, the Bayesian-network approach captures an agent who fits his model to a large historical dataset once and for all. Second, in Berk-Nash equilibrium, the DM's feedback is defined separately from the prior model. In contrast, in the Bayesian-network approach, feedback is defined by the very collection of marginal and conditional probabilities that is required for quantifying the DM's causal model. Thus, the prior model and the feedback are intertwined. To put it differently, the Bayesian-network model reflects the idea that the data that is available to the DM depends on the questions he poses, which in turn depend on his prior model.

4 Economic Applications

In this section I survey a number of theoretical exercises that make use of the Bayesian-network formalism and illustrate the behavioral implications of causal misperceptions in economic contexts. In each case, I emphasize the type of causal misperception that the application involves.

4.1 Demand for Education

A number of economic applications can be written in terms of the "ignoring confounders" specification described in Section 3.1: The objective distribution p is consistent with the true DAG G^* given by (5), and the DM's subjective DAG is $G : a \to y \to z$. In these applications, the DM misinterprets the correlation between intermediate and final outcomes as a causal effect, neglecting the confounding role of an exogenous (and possibly hidden) variable.

Perhaps the most familiar idea in empirical labor economics is the need to address bias due to confounding when estimating the causal effects of variables such as education or immigration on labor-market outcomes. Econometricians employ sophisticated techniques in order to mitigate this problem. By contrast, a layperson is far less likely to appreciate the role of confounding, let alone employ sophisticated methods for drawing correct causal inferences from observational data.

The following example considers parental investment in education. It is extracted from Spiegler (2016), using a different parameterization. All four variables take values in $\{0, 1\}$, where $a = 1$ represents a parent's costly investment in his child's schooling; $y = 1$ represents high school performance by the child; $z = 1$ represents a good subsequent labor-market outcome; and $\theta = 1$ represents high innate "ability". The parent's payoff is $z - ka$, where $k > 0$ is the investment cost. The exogenous components of the objective distribution p are as follows: $p(\theta = 1) = \delta$, $p(y = 1 \mid \theta, a) = a\theta$, $p(z = 1 \mid \theta, y) = \theta + \beta y(1 - \theta)$, where $\beta \in (0, 1)$ is a constant. Thus, investment and ability are complements in the production of a good school outcome, whereas school performance and ability are substitutes in the production of a good labor-market outcome (the constant β measures the strength of substitutability). The endogenous component of p is the parent's mixed strategy, given by $p(a = 1) = \alpha$. Note that the parent is uninformed of θ and therefore his action is independent of θ.

If the parent had rational expectations, he would strictly prefer $a = 0$. The reason is that a good school outcome is attained only if the child has high ability, but in this scenario the school outcome is irrelevant for the labor-market outcome. However, the parent's subjective causal model G means that he misperceives the correlation between y and z as a purely causal effect, neglecting the confounding role of θ. To find personal equilibria, we need to compare the cost k of investing in the child's education with its perceived benefit

$$\begin{aligned} p_G(z &= 1 \mid a = 1) - p_G(z = 1 \mid a = 0) = \sum_y [p(y \mid a = 1) - p(y \mid a = 0)]p(z = 1 \mid y) \\ &= [p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0)][p(z = 1 \mid y = 1) - p(z = 1 \mid y = 0)] \end{aligned}$$

Let us calculate the terms in this expression:

$$\begin{aligned} p(y &= 1 \mid a = 1) = \delta \\ p(y &= 1 \mid a = 0) = 0 \\ p(z &= 1 \mid y = 1) = \frac{\alpha\delta \cdot 1}{\alpha\delta} = 1 \\ p(z &= 1 \mid y = 0) = \frac{\delta(1-\alpha) + (1-\delta) \cdot 0}{1 - \alpha\delta} = \frac{\delta - \alpha\delta}{1 - \alpha\delta} \end{aligned}$$

such that the perceived gross benefit from investment is $\delta(1-\delta)/(1-\alpha\delta)$. Note that the substitutability parameter β ends up being irrelevant. On the other hand, as in the Dieter's Dilemma, the long-run action frequency α affects the parent's perceived consequences of his action. Therefore, the parent's subjectively optimal behavior is an equilibrium object.

Unlike the Dieter's Dilemma, here an increase in α leads to a *higher* subjective evaluation of $a = 1$. As a result, there can be *multiple* personal equilibria. Specifically, if k is between $\delta(1-\delta)$ and δ, there are three personal equilibria: $\alpha = 0$, $\alpha = 1$ and $\alpha = (k - \delta(1-\delta))/k\delta$. When $k < \delta(1-\delta)$, the unique personal equilibrium is $\alpha = 1$. Thus, the parent sub-optimally invests in the child's schooling because neglecting the confounding role of θ leads to over-estimation of the value of schooling.

4.2 A Roy Model

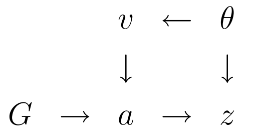
Another labor-economics classic is the Roy model of self-selection, which has been applied extensively to areas such as career choice or immigration (e.g. Borjas (1987)). To my knowledge, the empirical literature utilizing the Roy model invariably assumes rational expectations. However, it is likely that real-life agents draw naive inferences from observed correlations because they do not understand the selection mechanisms behind them.

The following example is a stylized Roy model of occupation choice. A DM is currently in occupation A and considers switching to occupation B. There are five variables: $a \in \{0, 1\}$ is the DM's action, where $a = 1$ indicates that he switches to occupation B; θ represents the DM's "ability"; v represents his earnings in the current occupation; z represents his observed earnings in occupation B, where $z = 0$ if the DM chooses not to switch to this occupation; and G is the DM's subjective DAG. The DM's utility is equal to his total earnings - i.e., $u(\theta, G, v, a, z) = az + (1 - a)v$.

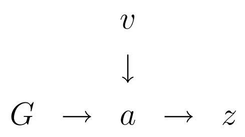
The objective distribution p is defined as follows. First, $\theta \sim U[-k, k]$, where $k > 1$ is a constant. Thus, $\theta > 0$ indicates above-average ability. Second, $v = \theta/k$ with certainty for every θ. Third, $z = \theta a$ with certainty for every θ, a. Since $k > 1$, occupation B has greater returns to ability than occupation A. The DM's DAG G is independent of θ and v. In the most basic Roy model, the DM learns his ability θ prior to making his decision. Instead, let us assume that the DM observes v rather than θ before taking his decision. This distinction would be immaterial under rational expectations, because θ and v are perfectly correlated. Thus, the DM's information consists of v and G.

Assume that G takes two values, G^* and G^0, with probabilities $1-\lambda$ and λ, respectively, and independently of θ and v. Thus, the objective distribution p is consistent with G^* in the following figure, whereas G^0 departs from G^*

by omitting θ (note that both DAGs include G as an ancestral variable):



$\mathbf{G}^*$



$\mathbf{G}^0$

The interpretation is that the population from which the DM is drawn consists of sophisticates and naifs. Sophisticates know the true process; in particular, they are aware that the latent variable θ acts as a confounding variable that affects the observed correlation between a and z. In contrast, naifs are unaware of this confounding effect, and interpret the correlation between a and z as a causal effect.

Let us characterize the equilibrium in this model. When $G = G^*$, the DM realizes that $z > v$ if and only if $v > 0$. Thus, in any equilibrium, a sophisticate will play $a = 1$ if and only if $v > 0$. Now consider a naive DM. According to his causal model, z is independent of (v, G) conditional on a. That is, he believes that he should play $a = 1$ if and only if $E(z \mid a = 1) > v$. This naive DM behaves as if he does not understand that the sample of switchers is selective. It follows that the strategy of a naive DM is to switch if and only if $v < v^*$, where v^* is the unique solution of

$$v^* = E(z \mid a = 1) = k \cdot \frac{\lambda \int_{-1}^{v^*} v dv + (1 - \lambda) \int_0^1 v dv}{\lambda \int_{-1}^{v^*} dv + (1 - \lambda) \int_0^1 dv}$$

This characterization has several noteworthy features. First, as in previous examples, the naive DM's subjectively optimal action is an equilibrium object: His evaluation of $a = 1$ depends on his perceived payoff from switching, which depends on the equilibrium switching patterns (including those of sophisticates). Second, while the sophisticate's switching strategy exhibits positive selection (DMs with above-average ability switch to the occupation with higher returns to ability), the naif's switching strategy exhibits *negative* selection. Yet, sophisticates exert an externality that weakens this negative selection - i.e., v^* decreases with λ. The reason is that by the positive selection of sophisticated switchers and the negative selection of naive switchers, a higher fraction of sophisticates in the population implies that on average, switchers have higher ability, which encourages naifs with large v to switch.[6]

4.3 Markets and Auctions

So far, we considered applications that involved a single DM. In the next few sub-sections, I turn to applications in which agents with causal misperceptions interact with rational agents or among themselves.

Esponda (2008) presents a monopsony example based on Samuelson and Bazerman (1985), where an uninformed buyer makes a take-it-or-leave-it offer to an informed seller. The original Samuelson-Bazerman example is a simple game-theoretic formulation of Akerlof's classic argument that adverse selection leads to market failure. Although Esponda (2008) used the example to illustrate naive behavioral equilibrium (mentioned in Section 3.4), it can be reformulated in terms of the "ignoring confounders" specification.[7]

Let a, θ, y and z represent the buyer's offer, the seller's valuation, the final allocation (i.e., whether trade takes place) and the buyer's gross payoff ($z = 0$ when there is no trade, and $z = \theta + b$ when trade occurs, where $b > 0$ is a constant). The buyer fails to take into account that θ confounds the correlation between y and z. Therefore, he does not realize that a higher bid increases not only the chances of trade but also the object's average quality conditional on trade. As a result, the market failure in the (unique) personal equilibrium is *more* pervasive than in the rational-expectations benchmark.

Eyster and Rabin (2005) apply "partially cursed equilibrium" to sealed-bid auctions, capturing the idea that bidders have limited understanding of the relation between opponents' bidding behavior and their private information. This enables Eyster and Rabin to represent bidders' failure to fully internalize the Winner's Curse. For example, an uninformed bidder with "fully

[6] The general form of true and subjective DAGs in this example can be applied in other contexts. Esponda and Pouzo (2014) study an electoral model in which voters' perceived consequences of their votes can be described in similar terms.

[7] Eyster and Rabin (2005) and Jehiel and Koessler (2008) investigated the same bilateral-trade example through the prism of the concepts of cursed equilibrium and ABEE. See Spiegler (2011, Ch. 8) for a pedagogical exposition.

cursed" beliefs correctly predicts the distribution over opponents' bids, but fails to perceive any correlation between bids and the object's value. As a result, he will erroneously consider truthful bidding to be weakly dominant in a second-price auction. When bidders are partially informed, cursed equilibrium generates rich patterns of over- and under-bidding relative to Nash equilibrium.

Antler and Bachi (2019) apply the concepts of ABEE and partially cursed equilibrium to an otherwise standard dynamic model of two-sided search with vertically differentiated, non-transferrable utility, and examine what happens when search frictions vanish. They show that even when the coarseness/cursedness friction is small, the equilibrium outcome is radically different from the perfectly assortative matching that emerges under rational expectations. In particular, a large group of agents with intermediate match values perpetually search and never find a match. Indeed, the share of these "eternal singles" in the population converges to one.

4.4 Monetary Theory

Monetary policy is another area in which systematic deviations from rational expectations can have important economic consequences. In a textbook model originated by Kydland and Prescott (1977) and Barro and Gordon (1983), a central bank controls a policy variable that affects inflation. Private-sector actors form an inflation forecast, possibly after observing the central bank's decision. In a simplified version of this model based on Sargent (1999), the objective distribution p is consistent with the following true DAG G^*:

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

where θ represents external variables that shape the central bank's preferences (especially how it trades off inflation and employment), a represents the central bank's action, π represents inflation, e stands for the private sector's inflation forecast and y represents real output.

Thus, private-sector expectations are relevant because real output (or employment) is determined by an "expectations-augmented" Phillips Curve, such that the real effect of inflation is at least partly offset when inflation is anticipated. To the extent that the central bank wishes to maximize expected output, it wants inflation to systematically exceed private-sector expectations. In other words, monetary policy involves "*expectations management*".[8]

Conventional models constrain expectations management by assuming that the central bank and the private sector share the same, correctly specified model of the macroeconomy. As a result, the private sector forms an unbiased inflation forecast conditional on its information, such that the central bank cannot generate systematic inflationary surprises.

Now relax the assumption that the private sector has rational expectations. In this context, causal models capture more than purely intuitive judgments. Professionals in financial markets and policy-making institutions often make use of explicit models, which sometimes involve deliberate causal assumptions based on theoretical preconceptions. The following are a few examples. First, $\theta \rightarrow a \rightarrow \pi \rightarrow y$ neglects the mediating role of expectations in the Phillips Curve. The DAG $\theta \rightarrow a \rightarrow y \rightarrow \pi$ combines this error with a reversal of inflation-output causation. Finally, the DAG

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

expresses a belief in monetary neutrality - i.e., a has no causal effect on y. This model maintains that any correlation between a and y is due to confounding by θ.

Spiegler (2018) examines the central bank's ex-ante optimal policy, under the assumption that the private sector observes the central bank's action $before$ forming its inflation forecast. The central bank's objective is to mini-

[8] As Woodford (2003, p. 15) argues, "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".

mize inflation and maximize output. The key question is whether the central bank can use monetary policy to enhance expected output in the long run. Under a linear Phillips Curve, where $E(y \mid \pi, e) = \pi - e$, the answer is negative when the private sector has rational expectations, because rational expectations imply $E(e) = E(\pi)$. Spiegler (2018) poses the following question: Which misspecified private-sector causal models retain the property that $E(e) = E(\pi)$?

This corresponds to the Bayes-plausibility property discussed in Section 3.3, and therefore it is amenable to the graphical characterization of this property given in Spiegler (2018). Let us revisit the three examples of misspecified private-sector causal models presented above. The DAGs $\theta \rightarrow a \rightarrow \pi \rightarrow y$ and $\theta \rightarrow a \rightarrow y \rightarrow \pi$ are $perfect$, hence the central bank cannot use monetary policy to generate systematic real effects.

In contrast, the DAG in (9) violates the graphical condition for unbiased inflationary forecasts - i.e., there are objective distributions for which $E(e) \neq E(\pi)$. Furthermore, it is possible to parameterize p (including a strategy for the central bank - i.e., how it chooses a as a function of θ) such that $E(e) < E(\pi)$. In this case, inflation systematically exceeds the private sector's conditional forecast, and therefore real output exceeds the rational-expectations benchmark.

However, certain conventional parameterizations rule out this possibility. Spiegler (2018) shows that when p is multivariate normal, any private-sector DAG implies $E(e) = E(\pi)$. This extends the classic impossibility result of Lucas (1972) and Sargent and Wallace (1975) far beyond the case of rational expectations, to the weaker assumption that the private sector generates its forecasts by fitting a recursive system of linear regression equations.

4.5 Contract Theory

A lively line of research examines principal-agent relationships when the agent has biased beliefs (see Spiegler (2011) and Kőszegi (2014) for reviews). Key questions are whether the agent's biases make him vulnerable to exploitation by the principal, whether the principal may have an incentive to "debias" the agent, and what form exploitative contracts take.

Schumacher and Thysen (2018) revisit the basic moral hazard model and relax the assumption that the agent correctly perceives the consequences of his actions. Their main example falls into the "neglecting indirect effects" mold of Section 3.1: The true DAG G^* is given by (6), where a is the agent's action; q_1 and q_2 are intermediate outcome variables; and y represents the final output, which is the only verifiable variable. The firm's payoff is $y - w$, whereas the agent only cares about w and a. One concrete story for G^* is that the agent is a mid-level manager who oversees a production team, and a represents his management style - e.g. whether he micromanages his staff; q_1 and q_2 represent the staff's workplace attendance and morale; micromanagement raises q_1 but lowers q_2; and output y increases additively in both q_1 and q_2.

Consider a parameterization in which the agent is risk-neutral and has limited liability. Micromanagement is more costly but yields higher output for the firm. If the agent had rational expectations, the principal would find it optimal to offer him a compensation scheme $w(y)$ that incentivizes the agent to micromanage his team despite the bad net effect this has on output. Now suppose the agent is unaware of the causal channel from a to y that passes through q_2 - i.e., he neglects the role of morale. His subjective DAG G is thus $a \rightarrow q_1 \rightarrow y$. The firm knows the correct model G^*. Its problem is to design a wage scheme $w(y)$ that maximizes $E(y-w)$ subject to the constraint that the agent plays a personal equilibrium given the wage scheme.

Under the above parameterization, the optimal contract induces a unique personal equilibrium, in which the agent always micromanages his team. This is not an obvious result because inducing the agent to play a mixed strategy could enhance his prediction error. The incentives are more weakly powered than in the rational-expectations benchmark. The firm's profit exceeds the rational-expectations benchmark, hence it has no incentive to debias the agent by teaching him the correct model.

Unlike other models in behavioral contract theory (e.g. Eliaz and Spiegler (2006)), exploitation of the boundedly rational agent in this model does not take place by inducing a wrong prediction of the consequences of his equilibrium action. Indeed, the agent correctly predicts the expected output on the equilibrium path. Instead, the agent's error is that he mispredicts the $counterfactual$ consequences of adopting an alternative managing style. In particular, because he neglects q_2, he ends up overestimating the effect of q_1 on output, which deters him from deviating to a more relaxed managing style. This enables the firm to adopt more weakly powered incentives, thus extracting part of the agent's informational rent.

Now suppose that q_1 joins y in the list of verifiable variables. Under rational expectations, the "informativeness principle" (Holmstrom (1979)) implies that the principal would prefer to contract on both q_1 and y. In contrast, the principal in the Schumacher-Thysen model faces two novel considerations. First, the agent believes that y is uninformative of a given q_1, hence conditioning on both q_1 and y introduces welfare-reducing noise from his (incorrect) point of view. Second, because the agent's subjective model does *not* distort $p(q_1 \mid a)$, conditioning on q_1 alone eliminates the principal's ability to exploit the agent's errors. The net result is that the principal prefers to condition on the wage on q_1 alone, a departure from the informativeness principle.

4.6 Simple Coarseness

Most of the misspecified causal models presented above involved an omitted variable. When it is a confounder or part of a causal chain from actions to consequences, non-trivial equilibrium effects can arise in single-agent decision problems. In these cases, the Bayesian-network language helps articulating the belief error and analyzing its behavioral implications.

The literature contains a number of models of omitted-variable errors, where a dependent variable of interest is objectively a function of several explanatory variables that lie outside the DM's control. The DM's subjective model omits some of these explanatory variables. The fundamental attribution error example in Section 3.2 is a case in point. I refer to such specifications of true and subjective DAGs as "simple coarseness". Because the variables involved are independent of the DM's action, the misperception is not as subtle as in other examples, and the Bayesian-network formulation is not as essential for understanding its behavioral implications. For instance, we can easily describe simple coarseness in terms of analogy-based expectations. In macroeconomics, the notion of "restricted perceptions equilibrium" (Evans and Honkapohja (2001)) describes beliefs that fall into the simple coarseness specification.

The following are a few examples from the literature. Mullainathan et al. (2006) study a persuasion problem in which the receiver of strategic communication reasons in terms of a coarse representation of the situation he is in. Kondor and Kőszegi (2015) model competitive supply of financial securities when investors exhibit coarseness. Piccione and Rubinstein (2003) study monopolistic price setting over time, where consumers vary in their ability to detect temporal patterns (captured by subjective models with different lags). Depending on how consumers' sophistication is correlated with their willingness to pay, the monopolist may be able to use temporally complex price patterns as a discrimination device. Eyster and Piccione (2013) examine a dynamic competitive asset market, in which traders use subjective models to account for price fluctuations, and these models differ in the explanatory variables they omit. Eyster and Piccione provide partial characterizations of equilibrium prices (their range and relation to the asset's fundamental value) as well as the returns that traders obtain. Jehiel and Samuelson (2012) consider long-run reputation, where short-run players correctly perceive the long-run player's average behavior but fail to perceive its history-dependence. Finally, Antler (2018) studies the design of pyramid schemes when potential distributors correctly perceive the average take-up rates but fail to understand how these change as the distribution network grows larger.

4.7 Endogenous Causal Models: Competing Narratives

So far, we have taken the DM's subjective causal model as given. However, in many contexts it is interesting to examine how it arises endogenously.

Consider the formation of beliefs about political questions such as: Does a dovish stance compromise national security? Do protectionist trade policies promote working-class welfare? Eliaz and Spiegler (2018) construct a model that explores the role of *narratives* in this belief-formation process. In the model, policies are promoted by narratives, formalized as DAGs that include action and consequence variables, as well as a selection of other variables. Thus, narratives differ in the variables they incorporate as well as in their location in the causal scheme.

A representative voter uses p_G to evaluate the policy that the narrative G is paired with, and selects the narrative-policy pair with the highest evaluation (i.e., the largest anticipatory utility). This captures the idea that voters are drawn to hopeful narratives. An equilibrium is a distribution over narrative-policy pairs that are selected in this manner, such that the distribution over policies is consistent with the equilibrium marginal distribution over actions. The need for an equilibrium notion of prevailing narratives mirrors the need for personal equilibrium in the decision model of Section 3.2.

Eliaz and Spiegler (2018) show that under mild conditions, the support of the equilibrium distribution contains multiple policies that are sustained by conflicting narratives. They also present stylized applications that demonstrate the formalism's ability to illuminate phenomena such as the tension between rational and "populist" narratives, the difference between the narratives that sustain hawkish and dovish foreign policies, and the role of narratives that deny or exaggerate the effect of policy.

5 A Dual Approach: Extrapolating Beliefs from Partial Data

In Section 3.4, I interpreted concepts like ABEE or $S(K)$ equilibrium in terms of DMs who perceive empirical regularities through the prism of a prior model. However, these concepts have a "dual" interpretation, according to which the DM has no prior model; he receives partial data and uses a "parsimonious" procedure for extrapolating a belief from the data. E.g., in ABEE, the DM receives coarse-grained data and extrapolates a belief that other variables of interest do not depend on finer details. Likewise, in $S(K)$ equilibrium, the DM extrapolates a belief that the sample is perfectly representative of the actual distribution.

A similar dual interpretation can be attempted for the Bayesian-network model. Recall the Dieter's Dilemma, and assume that the DM learns the joint distributions $(p(a,c))$ and $(p(c,h))$, but he lacks any data about $(p(a,h))$. This predicament is related to the "surrogate marker" problem in the context of drug approval.[9]

The DM has no prior causal model. Instead, he uses a simple rule for extrapolating a belief over a,c,h from his data: Choose the distribution over a,c,h that maximizes (Shannon) *entropy* subject to being consistent with $(p(a,c))$ and $(p(c,h))$. This distribution is precisely $p(a)p(c \mid a)p(h \mid c)$ - the same subjective distribution that our DM formed in Section 2. In other words, the belief $p_{a\rightarrow c\rightarrow h}(a,c,h)$ can be reinterpreted as the maximum-entropy extension of the partially specified distributions $(p(a,c))$ and $(p(c,h))$.

This idea is presented more generally in Spiegler (2017,2019). Recall the modeling framework of Section 3, where the objective distribution p is defined over the variables $x_1,...,x_n$. The DM's data access is defined by a collection $\mathcal{M}$ of subsets of $\{1,...,n\}$. For each $M \in \mathcal{M}$, the DM learns $(p(x_M))$. The DM's subjective belief is the distribution that maximizes Shannon entropy subject to being consistent with $(p(x_M))$ for all $M \in \mathcal{M}$. Maximum Entropy (MaxEnt) is a well-known criterion (going back to Jaynes (1957)), which generalizes the Lalplacian principle of insufficient reason. In the present context, it captures the idea that the DM refrains from assuming correlations unless he has data about them.

It turns out (see Hajek et al. (1992)) that when $\mathcal{M}$ satisfies the so-called "running intersection property", the MaxEnt extension of $(p(x_M))_{M\in\mathcal{M}}$ is p_G, where G is a perfect DAG whose set of maximal cliques is $\mathcal{M}$.[10] Spiegler

[9] See https://en.wikipedia.org/wiki/Surrogate_endpoint.

[10] The set $\mathcal{M}$ satisfies the running intersection property if its elements can be ordered

(2017) constructs a behaviorally motivated procedure of extrapolating from partial data, which mimics MaxEnt extrapolation in this case. Thus, when G is perfect, we can interpret p_G in two different ways: Fitting a prior causal model to data, or parsimonious model-free extrapolation from partial data. As the Dieter's Dilemma example illustrates, the two interpretations complement each other: Prior causal perceptions can lend greater confidence in MaxEnt-extrapolated beliefs.

Esponda's (2008) buyer-seller example, described in Section 4.3, can illustrate the dual interpretation. Suppose the buyer only learns the joint distributions $(p(a, z))$ and $(p(z, y))$ - i.e., he learns how bids map into trade as well as the payoff consequences of trade, but he lacks direct data about the correlation between bids and payoffs. Then, if the buyer employs MaxEnt to extrapolate a belief from his data, his subjective mapping from a to y will be as if he fits the causal model $a \rightarrow z \rightarrow y$ to p.

A related example is Jehiel's (2018) model of overconfident investors, which addresses the finding that entrepreneurs tend to be overly optimistic about their projects. A traditional explanation for this phenomenon is that entrepreneurs are temperamentally overoptimistic. Jehiel's alternative explanation is that entrepreneurs fail to recognize that the pool of undertaken projects (like the traded objects in Esponda's example) is selective. Spiegler (2017) reformulates this argument in terms of MaxEnt extrapolation from partial data.

The dual reinterpretation of p_G opens up other research directions. Spiegler (2019) presents a formalism of static games in which players' partial data access is part of the definition of their types. Jehiel (2011) considers an auction-design problem, in which the provision of partial data about bidders' equilibrium behavior is a novel instrument in the hands of the auction designer. Eliaz et al. (2018) study a communication game in which the sender submits a multi-dimensional message simultaneously with (correct but partial) data about the joint distribution of messages and the state of Nature.

$M_1, ..., M_K$, such that for every $i = 2, ..., K$, $M_i \cap (\cup_{j<i} M_j) \subseteq M_k$ for some $k < i$.

6 What Next?

The Bayesian-network framework provides a language for describing causal misperceptions and tools for exploring their behavioral consequences. The formalism derives its appeal from its proximity to natural language on one hand and the availability of rich mathematical techniques on the other hand. Although this review emphasized implications for equilibrium models, the formalism may be relevant for non-equilibrium models, too. Mailath and Samuelson (2018) study information aggregation when different individuals form conditional beliefs of some dependent variable based on private information and different subjective models. An individual agent's model can be represented by a DAG with two maximal cliques, only one of which includes the dependent variable. Agents (truthfully) communicate their beliefs repeatedly as in Geanakopolos and Polemarchakis (1982), such that an agent's belief in stage $k+1$ takes other agents' stage-k beliefs as an input. Mailath and Samuelson (2018) examine whether this process converges to a common belief and whether the correct distribution is in the convex hull of agents' limit beliefs.

Certain causal misperceptions are outside the expressive scope of DAGs. A DM's tendency to form a causal link between his action and some other variable may depend on the particular action taken. E.g., prescribing a medicine to a patient is more likely to be perceived as a cause of his subsequent recovery than inaction. DAGs cannot capture this distinction, which lies at the heart of Spiegler's (2013) model of strategic policy making when the public employs a naive heuristic for attributing observed outcomes to policy makers' moves. Likewise, the DM's belief in a causal influence may be "signed" - e.g. he may have a prior belief that investing in a child's education tends to *increase* his school performance. DAGs cannot encode this aspect of the causal belief. Finally, this review has dealt exclusively with *static* decision problems. The approach is applicable to dynamic models that exhibit some stationarity (e.g., the infinite-horizon processes with finite memory studied in Piccione and Rubinstein (2003) and Eyster and Piccione (2013), or Markov decision models in Esponda and Pouzo (2019)). However, non-stationary dynamic models are often outside DAGs' scope. Jehiel (2005) and Jehiel and Samet (2007) are examples of equilibrium concepts (capturing coarse reasoning) that are specialized for extensive-form games.

Within the DAG formalism, there is great potential for exploring how DMs perform *causal inferences*. In Section 3.2, I observed that in all the examples in this review, the DM draws correct causal inferences from observational data given his causal model; his only error is that the model is wrong. It is interesting to examine situations where this is not the case - especially when the DM's DAG includes unobservable variables. In these situations, if the DM continues to maximize expected utility with respect to the conditional probability $p_G(\cdot \mid a, x_I)$, he may indeed "confuse correlation with causation". Between this naive extreme and the opposite pole of normative, ultra-sophisticated causal inference, is there a middle ground of *partially sophisticated* causal inference that would describe real-life DMs? How can we model such a middle ground? Can Pearl's (2009) do-calculus formalism offer modeling tools?

This does not exhaust the interesting questions one can pose regarding the economic consequences of causal reasoning: How do DMs elicit causal models from observational data? How do they react to empirical refutations of their causal assumptions? What is the role of computational-complexity considerations in these processes? Such questions suggest that descriptive modeling of economic agents' causal reasoning will continue to benefit from ideas that originate in Psychology, Statistics and AI.

References

[1] Angrist, J. and J. Pischke (2008), Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press.

[2] Antler, Y. (2018), Multilevel Marketing: Pyramid-Shaped Schemes or Exploitative Scams?", mimeo.

[3] Antler, Y. and B. Bachi (2019), Searching Forever After, mimeo.

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

[5] Berk, R. (1966), Limiting Behavior of Posterior Distributions when the Model is Incorrect, Annals of Mathematical Statistics 37, 51-58.

[6] Borjas, G. (1987), Self-Selection and the Earnings of Immigrants, NBER Working Paper No. 2248.

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

[8] Eliaz, K. and R. Spiegler (2006), Contracting with Diversely Naive Agents, Review of Economic Studies 73, 689-714.

[9] Eliaz, K. and R. Spiegler (2018), A Model of Competing Narratives, mimeo.

[10] Eliaz, K., R. Spiegler and H. Thysen (2018), Strategic Interpretations, mimeo.

[11] Eliaz, K., R. Spiegler and Y. Weiss (2019), Cheating with (Causal) Models, mimeo.

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

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

[14] Esponda, I. and D. Pouzo (2014), Conditional Retrospective Voting in Large Elections, American Economic Journal: Microeconomics 9, 54-75.

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

[16] Esponda, I. and D. Pouzo (2019), Equilibrium in Misspecified Markov Decision Processes, mimeo.

[17] Esponda, I. and E. Vespa (2014), Hypothetical Thinking and Information Extraction in the Laboratory, American Economic Journal: Microeconomics, forthcoming.

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

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

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

[21] Eyster, E. and M. Rabin (2010), Naive Herding in Rich-Information Settings, American Economic Journal: Microeconomics 2, 221-43.

[22] Frydenberg, M. (1990), The Chain Graph Markov Property, Scandinavian Journal of Statistics 17, 333-353.

[23] Geanakoplos, J. and H. Polemarchakis (1982), We Can't Disagree Forever, Journal of Economic Theory 27, 192–200.

[24] Geiger, D., T. Verma and J. Pearl (1990), Identifying independence in Bayesian networks, Networks 20.5, 507-534.

[25] Hajek, P., T. Havranek and R. Jirousek (1992), Uncertain Information Processing in Expert Systems, CRC Press.

[26] Harris, J. (1998), The Nurture Assumption, London, Bloomsbury.

[27] Heckman, J. and R. Pinto (2015), Causal Analysis after Haavelmo, Econometric Theory 31, 115-151.

[28] Heidhues, P., B. Kőszegi and P. Strack (2018), Unrealistic Expectations and Misguided Learning, Econometrica 86, 1159-1214.

[29] Holmstrom, B. (1979), Moral Hazard and Observability, Bell Journal of Economics 10, 74-91.

[30] Imbens, G. (2019), Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics, Arxiv Working Paper 1907.07271.

[31] Jaynes, E. T. (1957), Information Theory and Statistical Mechanics, Physical Review 106, 620-630.

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

[33] Jehiel, P. (2011), Manipulative Auction Design, Theoretical Economics 6, 185-217.

[34] Jehiel, P. (2018), Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism, American Economic Review 108, 1582-97.

[35] Jehiel, P. and F. Koessler (2008), Revisiting Games of Incomplete Information with Analogy-Based Expectations, Games and Economic Behavior, 62, 533-557.

[36] Jehiel, P. and D. Samet (2007), Valuation Equilibrium, Theoretical Economics 2, 163-185.

[37] Jehiel, P. and L. Samuelson (2012), Reputation with Analogical Reasoning, Quarterly Journal of Economics 127, 1927-1969.

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

[39] Kondor, P. and B. Kőszegi (2015), Cursed Financial Innovation, WZB Discussion Paper.

[40] Kőszegi, B. (2010), Utility from Anticipation and Personal Equilibrium, Economic Theory, 44, 415-444.

[41] Kőszegi, B. (2014), Behavioral Contract Theory, Journal of Economic Literature 52, 1075-1118.

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

[43] Langer, E. (1975), The illusion of control, Journal of personality and social psychology 32, 311-328.

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

[45] Mailath, G. and L. Samuelson (2018), The Wisdom of a Confused Crowd: Model Based Inference, mimeo.

[46] Mullainathan, S. J. Schwartzstein and A. Shleifer (2008), Coarse Thinking and Persuasion, Quarterly Journal of Economics 123, 577-619.

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

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

[49] Pearl, J. (2014), Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Elsevier.

[50] Pearl, J. and D. Mackenzie (2018), The Book of Why: The New Science of Cause and Effect, Basic Books.

[51] 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.

[52] Ross, L. (1977), The Intuitive Psychologist and his Shortcomings: Distortions in the Attribution Process, Advances in Experimental Social Psychology, Vol. 10, Academic Press, 173-220.

[53] Samuelson, W. and M. Bazerman (1985), Negotiation under the Winner's Curse, Research in Experimental Economics 3, 105-38.

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

[55] 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.

[56] Schumacher, H. and H. Thysen (2018), Equilibrium Contracts and Boundedly Rational Expectations, mimeo.

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

[58] Spiegler, R. (2011), Bounded Rationality and Industrial Organization, Oxford University Press.

[59] Spiegler, R. (2013), Placebo Reforms, American Economic Review 103, 1490-1506.

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

[61] Spiegler, R. (2017), "Data Monkeys": A Procedural Model of Extrapolation From Partial Statistics, Review of Economic Studies 84, 1818-1841.

[62] Spiegler, R. (2018), Can Agents with Causal Misperceptions be Systematically Fooled? Journal of the European Economic Association, forthcoming.

[63] Spiegler, R. (2019), News and Archival Information in Games, mimeo.

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

[65] Woodford, M. (2003), Interest and Prices: Foundations of a Theory of Monetary Policy, Princeton University Press.