Bayesian Networks and Boundedly Rational Expectations*

Ran Spiegler†

December 8, 2015

Abstract

I present a framework for analyzing decision making under imperfect understanding of correlation structures and causal relations. A decision maker (DM) faces an objective long-run probability distribution p over several variables (including the action taken by previous DMs). He is characterized by a subjective causal model, represented by a directed acyclic graph over the set of variable labels. The DM attempts to fit this model to p, resulting in a subjective belief that distorts p by factorizing it according to the graph via the standard Bayesian-network formula. As a result of this belief distortion, the DM's evaluation of actions can vary with their long-run frequencies. Accordingly, I define a "personal equilibrium" notion of individual behavior. The framework enables simple graphical representations of causal-attribution errors (such as coarseness or reverse causation), and provides tools for checking rationality properties of the DM's behavior. I demonstrate the framework's scope of applications with examples covering diverse areas, from demand for education to public policy.

*This paper has benefitted from ESRC grant no. ES/L003031/1. I am grateful to Yair Antler, In-Koo Cho, Philip Dawid, Kfir Eliaz, Erik Eyster, Philippe Jehiel, Ehud Lehrer, Heidi Thysen and Michael Woodford, an editor and referees, as well as seminar and conference audiences, for many helpful conversations and comments.

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

1 Introduction

The rational-expectations postulate entails that agents in an economic model perfectly understand its equilibrium statistical regularities - in particular, the structure of correlations among variables. In recent years, economists have become increasingly interested in equilibrium models that relax this extreme assumption. This paper proposes an approach to modeling decision makers (DMs) with an *imperfect* understanding of equilibrium correlations, based on the idea that such flaws arise from an attempt to fit a misspecified causal model to the equilibrium distribution.

Consider a DM whose vNM utility function u is defined over a collection of variables $x = (x_1, ..., x_n)$, where x_1 is the DM's action. Imagine that before choosing how to act, the DM gets access to a "historical database" consisting of (infinitely) many joint observations of the relevant variables - including the actions taken by previous DMs facing the same decision problem. The empirical distribution p over x in the database obeys the textbook chain rule:

$$p(x) \equiv p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_n \mid x_1, ..., x_{n-1}) \tag{1}$$

To represent a DM who might misperceive the correlation structure of long-run distributions, I propose an *extension* of the chain rule. The DM is characterized by a *directed acyclic graph* (DAG) R over the set of nodes $\{1, ..., n\}$.[1] The DM's subjective belief distorts every objective long-run distribution p by "factorizing it according to R", via the formula

$$p_R(x) = \prod_{i=1}^{n} p(x_i \mid x_{R(i)}) \tag{2}$$

where $R(i)$ denotes the set of direct parents of the node i in the DAG, and $x_{R(i)}$ is the projection of x on $R(i)$. For instance, if $R : 1 \to 2 \to 3 \leftarrow 4$, then $p_R(x) = p(x_1)p(x_4)p(x_2 \mid x_1)p(x_3 \mid x_2, x_4)$.

A DAG R and the set of distributions representable by (2) define what is

[1] A directed graph is defined by a set of nodes and a set of directed links between nodes. The graph is acyclic if it does not contain any directed path from a node to itself.

known as a *Bayesian network*. This concept was introduced by statisticians as a representation of conditional-independence assumptions, and has become ubiquitous in Artificial Intelligence as a platform for efficient probabilistic-inference algorithms (see Cowell et al. (1999) and Koski and Noble (2009) for textbooks). In the present context, the DAG R is the DM's "type", and (2) describes how this type distorts every objective distribution p into the subjective belief p_R. When R is fully connected, it reduces (2) to a standard chain rule, thus representing a DM with rational expectations. At the other extreme, when R is empty, it represents a DM who cannot perceive *any* correlations that might actually exist: $p_R(x) = p(x_1) \cdots p(x_n)$.

Pearl (2009) advocated the view of DAGs as *causal structures* that underlie observed statistical regularities: the link $j \to i$ means that x_j is an immediate cause of x_i. Sloman (2009) presented psychological evidence that people use intuitive causal models to perceive uncertain environments, and employed DAGs to represent such models. The causal interpretation is consistent with the directedness and acyclicity properties of R: a causal chain from x_i to x_j should preclude a causal chain in the opposite direction. It also gives content to the factorization formula (2): to predict x_i conditional on its causes, we only need to know the realization of its *immediate* causes.

Following Pearl and Sloman, I interpret R primarily as a *subjective causal model*, such that p_R is the outcome of the DM's attempt to fit his (possibly misspecified) causal model to "historical data". The causal model is entirely non-parametric: it only posits the existence of certain causal links. The DM extracts the correlation between x_i and $x_{R(i)}$ (for each $i = 1, ..., n$) from the "historical database" represented by p. He focuses on these particular correlations because these are the ones that are required to quantify his causal model. The following example illustrates the behavioral implications of this idea.

An example: The Dieter's Dilemma

A DM who wishes to improve his health considers a diet that involves abstaining from a food he likes. In reality, the DM's choice and his health are statistically independent, yet they are potentially correlated with the level of some chemical in the DM's blood. The latter variable is payoff-irrelevant.

Therefore, a DM with rational expectations would choose *not* to diet.

Let a, h, c denote the DM's action (i.e., dieting decision), state of health and chemical level. Since a and h are independent, the objective long-run distribution p can be written as $p(a, h, c) = p(a)p(h)p(c \mid a, h)$. Thus, p is consistent with a "true DAG" R^* : $a \rightarrow c \leftarrow h$ - i.e., a causal model that posits a and h as independent causes of c.[2] This is "as if" consistency: although R^* may well describe an actual causal mechanism that underlies p, as modelers we are free to regard R^* as a mere representation of purely statistical independence between a and h.

Assume that our DM's subjective DAG is R : $a \rightarrow c \rightarrow h$ - i.e., he *inverts the causal link* between the chemical level and his health, relative to the true DAG R^*. The DM's attempt to fit his causal model to the long-run distribution p generates the subjective belief $p_R(a, h, c) = p(a)p(c \mid a)p(h \mid c)$. Guided by this belief, the DM will choose a to maximize

$$\sum_h p_R(h \mid a)u(a, h) = \sum_h \sum_c p(c \mid a)p(h \mid c)u(a, h) \tag{3}$$

Now impose additional structure on the true process: the chemical level is normal when the DM is healthy *or* when he diets; otherwise, it is abnormally high. Consequently, as long as not all historical DMs chose to diet, c and h will exhibit non-degenerate correlation in the database, and therefore $p_R(h \mid a)$ will be a non-constant function of a. Thus, although in reality a and h are independent, the DM's estimated causal model leads him to perceive an indirect causal effect of a on h. The reason is that although the DM correctly perceives the steady-state correlation between c and h, he mistakes it for a direct causal effect of c on h, which (according to R) implies an indirect causal effect of a on h.

Furthermore, the magnitude of this perceived effect is sensitive to the frequency of dieting in the database. To see why, note that the DM's subjective causal model postulates that a and h are independent *conditional* on c (as

[2] In diagrams, I will often name nodes by the variables' intuitive labels (rather than by the indices $1, ...n$). Although this practice involves some abuse of notation, it facilitates reading the graphs.

evident from (3)). However, this is *false*: if we knew that the chemical level is normal, learning whether the DM is dieting would affect our prediction of his state of health. It follows that the term $p(h \mid c)$ in (3) is sensitive to the marginal of p over a. In particular, a lower long-run frequency of dieting strengthens the estimated correlation between h and c, and hence the DM's subjective evaluation of dieting.

We are thus led to think of the DM's steady-state behavior as an *equilibrium* concept: the long-run distribution p is in equilibrium if it only assigns positive probability to actions a that maximize the expectation of $u(a, h)$ with respect to $p_R(h \mid a)$. In Section 3.1, I show that when the direct disutility from dieting is not too large, the DM's equilibrium probability of dieting is positive. For a range of parameter values, the unique equilibrium is *mixed* - a genuine equilibrium effect that is impossible under standard expected-utility maximization.[3]

I present the equilibrium model of individual choice in Section 2. For expositional simplicity, I assume that the DM cannot condition his action on any signal (Appendix C relaxes this assumption). An objective steady-state distribution p is a "*personal equilibrium*" if whenever an action x_1 is played with positive probability, it maximizes the expectation of u with respect to the conditional distribution $p_R(x_2, ..., x_n \mid x_1)$. A conventional "trembling hand" criterion handles zero-probability events.[4]

The integration of the Bayesian-network factorization formula (2) into an equilibrium model of choice constitutes the paper's first main contribution. It provides a framework for analyzing the behavior of a DM who forms his beliefs by fitting a subjective causal model to objective distributions. Because graphical causal models are entirely non-parametric, they are applica-

[3]The idea that the DM's long-run behavior may affect his evaluation of actions when he misperceives correlation structures has precedents in the literature (see Sargent (2001) and Esponda (2008)). To my knowledge, this paper offers the first general articulation of this idea.

[4]The term "personal equilibrium" was introduced by Kőszegi and Rabin (2006) and Kőszegi (2010) in the context of decision making with reference-dependent preferences, when the reference point is a function of the DM's expectation of his own choice. Geanakopolos et al. (1989) study a model in which the DM's payoff is a direct function of his prior belief, and this can lead to equilibrium effects in individual choice.

ble to *any* static decision problem. The framework thus provides a "general recipe" for transforming a standard rational-expectations model into an equilibrium model with non-rational expectations: substitute $p_R(x_2, ..., x_n \mid x_1)$ for $p(x_2, ..., x_n \mid x_1)$ in the definition of individual best-replying, where p is the true equilibrium distribution.

The paper's other major contributions can be summarized as follows.

Capturing errors of causal/statistical reasoning

Section 3 and Appendix C present applications of personal equilibrium to various domains. Each example is characterized by a "true DAG" R^*; the DM's DAG R is obtained by *performing a basic operation on* R^* - removing, inverting or reorienting links. Different operations capture different errors of causal attribution - link inversion captures "reverse causation", link removal captures "coarseness", etc. In specific contexts, these errors translate to well-known statistical fallacies. E.g., Section 3.2 studies an example of parental investment in education. The parent's DAG distorts R^* by removing the links flowing from a node that represents the child's "latent ability" into nodes that represent his school and labor-market outcomes. This removal of links captures the fallacy of ignoring a confounding variable. In (potentially multiple) personal equilibria, the parent over-invests in education. As the examples show, the Bayesian-network framework offers a language for describing errors of causal attribution and for analyzing their behavioral implications.

General characterizations of choice behavior

The framework is more than a language; it also provides tools for checking general rationality properties of personal equilibrium. Section 4.1 states a necessary and sufficient condition for the possibility of equilibrium effects, in terms of the structural relation between R and R^*. This condition is easy to operationalize, thanks to a basic concept from the Bayesian-networks literature called d-separation (explained in Appendix B). Section 4.2 presents a necessary and sufficient condition on R for the DM's behavior to be consistent with rational expectations in all environments that share the same restriction on the subset of payoff-relevant variables. Results of this kind are valuable

because they illuminate the robustness of an economic model's predictions to departures from full-fledged rational expectations. Finally, Section 4.3 shows that two subjective DAGs never dominate one another in terms of objective expected payoff, unless exactly one of them is fully connected.

Bayesian networks as a unifying framework

This paper offers a fresh look at the growing literature on equilibrium models with non-rational expectations. Osborne and Rubinstein (1998) studied games with players who misperceive the consequences of their actions, due to naive extrapolation from small samples. Eyster and Rabin (2005) assumed that Bayesian-game players underestimate the correlation between opponents' actions and signals. Madarasz (2012) modeled players who suffer the "curse of knowledge". In Esponda (2008) and Esponda and Pouzo (2014a), agents neglect the counterfactual effect of their actions on the distribution of payoff consequences. Piccione and Rubinstein (2003), Jehiel (2005), Jehiel and Koessler (2008), Mullainathan et al. (2008), Eyster and Piccione (2013) and Schwartzstein (2014) studied models in which agents' beliefs are measurable with respect to a *coarse* representation of the set of contingencies (by omitting variables from their subjective model or by clumping contingencies into "analogy classes").[5] Section 5 shows that some of these concepts can be reformulated as special cases of the present framework (defined by suitable R^* and R), or as refinements and extensions thereof. Bayesian networks thus offer a unifying framework, highlighting the thread of flawed causal reasoning that runs through equilibrium models with non-rational expectations.[6]

2 The Modeling Framework

Let $X = X_1 \times \cdots \times X_n$ be a finite set of *states*, where $n \geq 2$. I refer to each x_i as a *variable*, and $N = \{1, ..., n\}$ is the set of variable labels. For every $M \subseteq N$ and $x \in X$, denote $x_M = (x_k)_{k \in M}$. The set X_1 represents the set

[5] Similar elements of "coarse reasoning" appeared in macroeconomics under the title "restricted perceptions equilibrium" (Evans and Honkapohja (2001), Woodford (2013)).

[6] Graphical probabilistic models were introduced into economics in other contexts: to facilitate computation of Nash equilibria (Kearns et al. (2001), Koller and Milch (2003)), or to discuss causality in econometric models (White and Chalak (2009)).

of *actions* that are available to a decision maker (DM). Accordingly, I will often use the notation $X_1 = A$, $x_1 = a$, $x = (a, y)$, $y = (x_2, ..., x_n)$. Note that the DM's action is part of the description of a state. The DM is entirely uninformed of y when he acts (Appendix C relaxes this assumption).

2.1 Beliefs

Let $p \in \Delta(X)$ be an *objective* probability distribution over states. To capture limited understanding of the correlation structure of p, I introduce a new primitive. A *directed acyclic graph* (DAG) is a pair (N, R), where N is the set of nodes and R is the set of directed links. To describe a link from j to i, I use the notations jRi and $j \to i$ interchangeably. Let $R(i) = \{j \in N \mid jRi\}$ denote the set of "parents" of the node i. E.g., in the DAG $1 \to 3 \leftarrow 2$, $R(1) = R(2) = \varnothing$ and $R(3) = \{1, 2\}$. In what follows, I identify the DAG with R.

The DM is characterized by a "*subjective DAG*" R. For any objective distribution p, the DM's subjective belief over X is p_R, given by the factorization formula (2).[7] Thus, R is a short-hand for a *mapping* that assigns a subjective belief to every objective distribution. It is instructive to compare this to the traditional notion of subjective priors. Under the latter approach, the DM has a fixed belief that is *independent* of the objective distribution p. In contrast, according to (2), the DM's subjective belief *changes systematically* with p.

We will say that p is *consistent with a DAG* R if $p_R(x) \equiv p(x)$. If p is consistent with R, it is necessarily consistent with every DAG that adds links to R. For any three disjoint subsets $B, C, D \subset N$, the notation $x_B \perp_R x_C \mid x_D$ means that x_B and x_C are independent conditional on x_D, for every p that is consistent with R. Appendix B presents a basic tool from the Bayesian-networks literature, called d-separation, which characterizes the conditional-independence properties satisfied by all distributions that are consistent with a given DAG. Thus, a DAG can be viewed as a representation of a list of

[7] The formula contains potentially ill-defined terms, because it is possible that $p(x_{R(i)}) = 0$ for some i and x. This does not pose any difficulty for us, because we can exclude zero-probability realizations of x when performing expected-utility calculations.

conditional-independence properties. For example, the DAG $1 \rightarrow 3 \leftarrow 2$ represents the property $x_2 \perp x_1$, while the DAG $1 \rightarrow 3 \rightarrow 2$ represents the property $x_2 \perp x_1 \mid x_3$. (Not every consistent list of conditional-independence properties has a DAG representation.)

The DM's subjective distribution over y conditional on a is defined as usual,

$$p_R(y \mid a) = \frac{p_R(a,y)}{p_R(a)} = \frac{p_R(a,y)}{\sum_{y'} p_R(a,y')} \tag{4}$$

as long as $p_R(a) > 0$.

The interpretation of R

I regard the DM's DAG R as a *subjective causal model*: for every i, $R(i)$ represents the collection of variables that the DM perceives as immediate causes of x_i (alternative interpretations are discussed in Section 6). The subjective belief p_R is the outcome of the DM's attempt to fit his causal model to long-run data generated by p. The DM does not have any preconception regarding the sign or magnitude of causal relations - he infers those from the data; his causal model merely postulates causal links and their direction. The causal interpretation justifies the inclusion of the DM's action in the description of a state - the DM's subjective model establishes causal relations among all variables, including the action.

The following image makes the causal interpretation more concrete. The DM has access to a rich "historical database" consisting of many observations of joint realizations of a and y, independently drawn from p. He poses a sequence of n questions to the database, where question i is: "What is the distribution over x_i conditional on $x_{R(i)}$?" The DM poses these particular questions because he looks for the correlations that are required to complete the specification of his causal model. The data does not "speak for itself"; extracting correlations from it requires an effort, which the DM exerts only if it serves the identification of his model. In particular, he does not look for additional correlations that could test whether his model is misspecified. The DM forms his belief by taking the product of the measured conditional distributions $p(x_i \mid x_{R(i)})$, thus quantifying his causal model.

Equivalent DAGs

A given p can be consistent with multiple DAGs, even when they do not add links to one another. For instance, the DAGs $1 \rightarrow 2$ and $2 \rightarrow 1$ are both consistent with rational expectations, due to the basic identity $p(x_1, x_2) \equiv p(x_1)p(x_2 \mid x_1) \equiv p(x_2)p(x_1 \mid x_2)$. This suggests a natural equivalence relation: two DAGs are equivalent if they represent the same mapping from objective distributions to subjective beliefs.

Definition 1 *Two DAGs R and Q are* ***equivalent*** *if $p_R(x) \equiv p_Q(x)$ for every $p \in \Delta(X)$.*

Thus, two different causal models can be indistinguishable in terms of the statistical regularities they are consistent with. In particular, a DAG that involves intuitive causal relations can be equivalent to a DAG that makes little sense as a causal model (e.g., it postulates that the DM's action is caused by his final payoff).

The following characterization of equivalent DAGs will be useful in the sequel. it relies on two definitions. First, let $\tilde{R}$ be the undirected version, or *skeleton* of R - that is, $i\tilde{R}j$ if and only if iRj or jRi. Second, define the *v-structure* of a DAG R to be the set of all ordered triples of nodes (i, j, k) such that iRk, jRk, $i\not{R}j$ and $j\not{R}i$ (that is, R contains links from i and j into k, yet i and j are not linked to each other).

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

To illustrate this result, all fully connected DAGs have the same skeleton and a vacuous v-structure; hence they are all equivalent (indeed, they all induce rational expectations because they reduce (2) to a textbook chain rule). In contrast, the DAGs $1 \rightarrow 2 \rightarrow 3$ and $1 \rightarrow 2 \leftarrow 3$ are not equivalent because they have identical skeletons but different v-structures (vacuous in the former case, and consisting of the triple $(1, 3, 2)$ in the latter).

True and subjective DAGs

I will often restrict the domain of possible objective distributions p to be those that are consistent with some DAG R^*. In this case, I will simply say that the "*true DAG*" is R^*. Such domain restrictions arise naturally when reality has an underlying causal structure. A fully connected DAG corresponds to an unrestricted domain of objective distributions.

In the applications, the DM's subjective DAG R will be obtained from R^* via one of the following simple operations: *inverting, removing, reorienting or adding links*. These operations intuitively correspond to basic errors of causal attribution. We saw that inverting a link captures reverse causation.[8] Let us briefly discuss the others:

(i) Removing a link captures *coarseness*. For example, if $R^* : 1 \rightarrow 3 \leftarrow 2$ and $R : 1 \rightarrow 3 \quad 2$, then R captures a coarse perception of the causes of x_3: in reality, x_3 is a function of both x_1 and x_2, yet R acknowledges only x_1. Thus, while in reality $p(x) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$, the DM's subjective belief is $p_R(x) = p(x_1)p(x_2)p(x_3 \mid x_1)$.

(ii) Changing the origin of a link captures *misattribution*. For instance, if $R^* : 1 \quad 3 \leftarrow 2$ and $R : 1 \rightarrow 3 \quad 2$, then R errs by attributing x_3 to the wrong cause (x_1 instead of x_2). Compare this example to the one used to illustrate coarseness: the same R can capture different errors, depending on its exact relation to R^*.

(iii) Adding a link captures *spurious direct causation* - i.e., the DM postulates a direct causal relation that does not exist in reality. However, recall that if p is consistent with R^* then it is also consistent with any DAG that adds links to R^*. It follows that incorrect subjective beliefs that arise from spurious direct causation are outside the framework's scope.[9]

Throughout the paper, I will assume that $R^*(1) = R(1) = \varnothing$ - i.e. the

[8] Because inversion and reorientation seem to capture distinct errors, I consider them "primitive", even though they can be decomposed into addition and removal of links.

[9] The operations need not preserve equivalence between DAGs. Let $N = \{1, 2, 3\}$ and omit the link $2 \rightarrow 3$ from two linear orderings that contain it - one in which $1 \rightarrow 2$ and another in which $3 \rightarrow 1$. The resulting DAGs are not equivalent: $2 \leftarrow 1 \rightarrow 3$ and $2 \rightarrow 1 \leftarrow 3$.

node that represents the DM's action is *ancestral* in both true and subjective DAGs. This restriction is unnecessary: a DAG in which 1 is ancestral can be equivalent to a DAG in which it is not. Nevertheless, I will adhere to it for several reasons. First, it fits the causal interpretation: arguably, the DM's action cannot be caused by something he is not informed of. Second, the restriction simplifies the expression for (4):

$$p_R(x_2, ..., x_n \mid x_1) = \prod_{i=2}^{n} p(x_i \mid x_{R(i)})$$

Finally, the restriction suggests a natural definition of endogenous variables: x_i is *endogenous* according to R if i is a descendant of 1 (i.e., if there is a directed path from 1 to i).

2.2 Decisions

Let us turn to decision making under the DAG representation of subjective beliefs. Our DM is an expected utility maximizer, with a vNM utility function $u : X \to \mathbb{R}$. Recall that p represents a long-run joint distribution over all variables. We will require the DM's long-run behavior (given by the marginal of p over a) to be optimal with respect to his $p_R(y \mid a)$ (i.e., the perceived stochastic mapping from a to y). The belief distortion inherent in p_R allows $p_R(y \mid a)$ to vary with $(p(a))_a$, which means that the marginal long-run distribution over actions can influence the DM's evaluation of each course of action. Therefore, we are led to define individual choice as an *equilibrium* notion.

To motivate the definition of equilibrium, suppose the above-mentioned "historical database" is created by a long sequence of short-lived agents facing the decision problem. The distribution p represents the historical joint distribution over the agents' actions and all other variables. Each agent forms a subjective (possibly distorted) view of historical data given by p_R, and takes an action that maximizes his expected payoff according to this subjective view. Equilibrium will correspond to a steady state of this dynamic.

As usual, equilibrium will require the DM to optimize with respect to his

subjective belief. It would be conventional to treat $(p(a))_a$ as the object of the definition and take $(p(y \mid a))_{a,y}$ (i.e., the true stochastic mapping from a to y) as given. Instead, I will treat the *entire joint distribution* p as the object of the formal definition. This is a contrivance that simplifies notation (in applications, I will fix $(p(y \mid a))_{a,y}$ and find $(p(a))_a$).

The need for an equilibrium definition requires us to consider off-equilibrium actions. I address this concern with a conventional "trembling hand" criterion. We say that p' is a *perturbation* of p if $p'(y \mid a) \equiv p(y \mid a)$ and the marginal of p' on A has full support. A perturbation fixes every aspect of p except the DM's behavior, such that every action is played with positive probability.

Definition 2 (Personal equilibrium) *Fix an arbitrary DAG R. A distribution $p \in \Delta(X)$ with full support on A is an ε-**perturbed personal equilibrium** if*

$$a \in \arg\max_{a'} \sum_y p_R(y \mid a') u(a', y)$$

*whenever $p(a) > \varepsilon$. A distribution $p^* \in \Delta(X)$ is a **personal equilibrium** if there exists a sequence $p^k \to p^*$ of perturbations of p^*, as well as a sequence $\varepsilon^k \to 0$, such that p^k is an ε^k-perturbed personal equilibrium for every k.*

The concept of ε-perturbed personal equilibrium allows the DM to experiment with (subjectively) sub-optimal actions with probability ε at most. To use the historical-database metaphor, in order for its empirical distribution p to be an ε-perturbed equilibrium, it must be the case that if the frequency of an action in the database is greater than ε, then the DM finds it subjectively optimal with respect to p_R. Personal equilibrium simply takes the $\varepsilon \to 0$ limit.

Proposition 2 *Fix an arbitrary DAG R. For every $(p(y \mid a))_{a,y}$, there exists $(p(a))_a$ such that $p = ((p(a)_a, (p(y \mid a))_{a,y})$ is a personal equilibrium.*

As we shall see in Section 3, "pure" personal equilibria (where the marginal of p over a is degenerate) need not exist. Note that if we restrict the

domain of possible objective distributions to those that are consistent with a true DAG R^*, then all personal equilibria are consistent with this DAG.

3 Illustrations

This section analyzes personal equilibria for various specifications of true and subjective DAGs R and R^*. In each example, R is obtained from R^* by one of the basic operations discussed in Section 2. Each sub-section presents the material in a different concrete economic context, thus illustrating the framework's scope of applications. The analyses follow a two-step "recipe". First, I provide a basic characterization of the DM's personal-equilibrium behavior, involving a formula for $p_R(y \mid a)$ that is based entirely on the structure of R. Certain properties of this characterization can be gleaned from this formula (aided by knowledge of the true DAG R^* and the payoff-relevant variables). These properties hold under *any* parameterization of the true process that is consistent with R^*. In the second step, I impose parametric assumptions on u and $p(y \mid a)$ that fit the economic scenario, and use them to obtain closed-form expressions for the conditional-probability terms in the formula for $p_R(y \mid a)$. This enables me to complete the characterization of personal equilibria.[10]

3.1 Reversing Causation: Health and Lifestyle Choices

In this section I formally develop the example of the Dieter's Dilemma, described in the Introduction. The three variables, a, h, c, represent the DM's nutritional choice, health outcome and chemical level. The DM is uninformed of c and h at the time he chooses a. The variables a and h are statistically independent, yet c is potentially correlated with both. Thus, every possible objective distribution p can be written as $p(a, h, c) = p(a)p(h)p(c \mid a, h)$ - i.e., the true DAG is $R^* : a \rightarrow c \leftarrow h$. If the DM had rational expectations,

[10]Throughout this section, the trembling-hand criterion in the definition of personal equilibrium merely ensures that equilibria are well-defined, but the equilibria do not rely on the selection of the sequence of perturbations. Trembles play a more interesting role in the example analyzed in Appendix C.

he would choose a to maximize

$$\sum_h \sum_c p(h)p(c \mid a, h)u(a, h, c)$$

Suppose that the DM's subjective DAG is $R : a \to c \to h$, such that $p_R(a, h, c) = p(a)p(c \mid a)p(h \mid c)$. Thus, relative to the true DAG R^*, R *inverts the direction of the causal link between health and the chemical level.* If p is a personal equilibrium, then for every a' for which $p(a') > 0$,

$$a' \in \arg\max_a \sum_h \sum_c p(c \mid a)p(h \mid c)u(a, h, c)$$

According to the true DAG R^*, h is *not* necessarily independent of a conditional on c. The conditional probability $p(h \mid c)$ implicitly involves summing over the DM's actions, where the weights are affected by the marginal of p over a; if $(p(a))_a$ were to change, so could $p(h \mid c)$, and so could the DM's subjectively optimal action. Thus, the equilibrium aspect of the DM's choice is not redundant.

The following concrete example is approximated by this description. Observational studies revealed that low levels of Vitamin D are associated with certain adverse health conditions. The common practice of prescribing Vitamin D pills is justified by the interpretation of this observed correlation as a causal effect of Vitamin D deficiency on health outcomes. Clinical tests that directly tested for this effect came later. In a systematic literature review (which has admittedly generated controversy), Autier et al. (2014) argued that these studies showed no effect for many of the measured health indicators, and suggested that a leading explanation for the null effect is reverse causation.[11]

The argument that a popularly held belief exhibits reverse causation can

[11]The debate over SSRI anti-depressants has similar contours - see http://www.webmd.com/depression/features/serotonin for a popular description. In both cases, the true causal mechanism is not known to medical researchers. To the extent that drug intake and health are statistically independent, this regularity is merely *consistent* with the reverse–causation explanation. However, this consistency is all we need to assume for the present exercise.

be found in other contexts. Harris (1998) claimed that psychologists' tendency to attribute children's personality traits to their parents' behavior may be a reverse-causation fallacy (e.g., children with a mild disposition may cause parents to behave mildly). In macroeconomics, theories of the "Phillips Curve" sometimes differ in the direction of causation between inflation and unemployment that they posit. Debates over the causal interpretation of the correlation between GDP growth and income inequality or public debt are another case in point.

For the rest of this sub-section, I impose additional structure. All variables take values in $\{0, 1\}$. The DM's payoff is purely a function of a and h: $u(a, h) = h - \kappa a$. The interpretation is that $a = 1$ represents an action referred to as "dieting" (taking a food supplement, abstaining from a favorite type of food, etc.) the cost of which is κ, and $h = 1$ represents a good health outcome. Let $c = 0$ represent a normal chemical level. The true stochastic process is as follows: the probability of $h = 1$ is $\frac{1}{2}$, independently of a. The value of c is a deterministic function of a and h, given by $c = (1 - a)(1 - h)$. Thus, the chemical level is abnormal if and only if the DM's health state is poor and he does not diet. Under rational expectations, the DM would choose $a = 0$ with certainty.

Let us now characterize personal equilibria under the DM's subjective DAG R. Denote $p(a = 0) = \beta$. We have specified $p(h, c \mid a)$; hence it remains to find β. Before stating the result, let us calculate a few relevant conditional probabilities: $p(c = 0 \mid a = 1) = 1$, $p(c = 0 \mid a = 0) = \frac{1}{2}$, $p(h = 1 \mid c = 1) = 0$ and

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

The fact that β appears in the last expression demonstrates our earlier observation that it is *not* true that $h \perp_{R^*} a \mid c$. The intuition is simple. If we learn that the DM is not dieting ($a = 0$), we can predict perfectly negative correlation between c and h. In contrast, if we learn that the DM is dieting ($a = 1$), health is equally likely to be good or bad, independently of the chemical level. The lower the long-run frequency of dieting, the stronger the

estimated (negative) correlation between health and the chemical level. The example thus exhibits "strategic substitutability", in the sense that a higher steady-state frequency of dieting leads to a smaller perceived effect of c on h, and this in turn weakens the DM's tendency to diet.

Proposition 3 *Given the specification of R, u and $(p(h,c \mid a))_{a,h,c}$ in this sub-section, there is a unique personal equilibrium, in which*

$$\beta = \begin{cases} 0 & if \quad \kappa \leq \frac{1}{4} \\ 2 - \frac{1}{2\kappa} & if \quad \kappa \in (\frac{1}{4}, \frac{1}{2}) \\ 1 & if \quad \kappa \geq \frac{1}{2} \end{cases}$$

Proof. (Proofs of later results appear in Appendix A.) The DM's evaluation of $a = 0$ given β is

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

The DM's evaluation of $a = 1$ given β is

$$\begin{aligned} p_R(h &= 1 \mid a = 1) - \kappa = \\ p(c &= 0 \mid a = 1) \cdot p(h = 1 \mid c = 0) + p(c = 1 \mid a = 1) \cdot p(h = 1 \mid c = 1)] - \kappa = \\ & 1 \cdot \frac{1}{2 - \beta} + 0 \cdot 0 - \kappa \end{aligned}$$

Note that when the DM evaluates an action, he takes β *as given*, as required by the notion of personal equilibrium. In ε-perturbed personal equilibrium, $\beta > \varepsilon$ ($\beta < 1 - \varepsilon$) only if $a = 0$ (1) attains the highest evaluation. Taking the $\varepsilon \to 0$ limit gives the result.[12] ■

Thus, although dieting is unambiguously sub-optimal under rational expectations, it is played with positive probability in personal equilibrium given

[12] The trembling-hand aspect of the equilibrium concept is a mere formal nicety in all the examples of this section. Therefore, I will skip the step that examines ε-perturbed equilibrium in proofs of later results in this section.

the DM's subjective DAG, as long as it is not too costly. The intuition is as follows. Suppose the DM plays $a = 0$ in a putative equilibrium. Then, he learns a perfectly negative correlation between c and h. He correctly grasps the effect of his own action on the chemical level. And since he misperceives the correlation between c and h as a causal effect of the former on the latter, he erroneously concludes that the normal chemical level attained thanks to dieting will lead to good health. This is not the usual logic of self-confirming expectations (Fudenberg and Levine (1993)): the DM's reasoning does *not* rest on off-equilibrium beliefs, but on incorrect causal inference from statistical regularities *on* the equilibrium path.

The possibility of a "mixed" unique equilibrium demonstrates that individual choice in this model is fundamentally an equilibrium notion; such an effect would be impossible under conventional expected-utility maximization.

3.2 Coarseness I: Demand for Education

In this sub-section I present an example in which the DM's subjective DAG commits a "coarseness" error of ignoring an exogenous confounding variable. The example is couched in terms of demand for education. The DM is a parent who chooses how much to invest in his child's education (I use a female pronoun for the child). There are four variables, denoted a, θ, s, w, representing the parent's investment, the child's innate ability, her school performance and her labor-market outcome (measured by her wage). The parent is uninformed of θ, s and w at the time of choice. The true DAG R^* is

$$\begin{array}{ccccc} a & \rightarrow & s & \leftarrow & \theta \\ & & & \searrow & \downarrow \\ & & & & w \end{array} \tag{5}$$

Note that θ is an exogenous "confounder": it causes variation in both s and w. The parent's utility u is purely a function of a and w. Therefore, if the parent had rational expectations, he would choose a to maximize

$$\sum_{\theta} p(\theta) \sum_{s} p(s \mid a, \theta) \sum_{w} p(w \mid \theta, s) u(a, w)$$

Now suppose the parent's subjective DAG removes the links of θ in R^*. Because θ is payoff-irrelevant, we can equivalently assume that R omits it altogether - i.e., $R : a \rightarrow s \rightarrow w$. One interpretation is that the parent neglects θ because it is *unobservable*. In personal equilibrium, the parent will choose a to maximize

$$\sum_s p(s \mid a) \sum_w p(w \mid s) u(a, w)$$

This expression can be elaborated into

$$\sum_s \left(\sum_\theta p(\theta) p(s \mid a, \theta) \right) \sum_w \left(\sum_{\theta'} p(\theta' \mid s) p(w \mid \theta', s) \right) u(a, w)$$

Thus, the parent's objective function involves two implicit summations over the latent variable. The term $p(\theta' \mid s)$ is sensitive to $(p(a))_a$ - e.g., if the child performs well at school despite low parental investment, then she probably has high ability. It follows that individual choice in this example is fundamentally an equilibrium notion.

In the rest of this sub-section I impose additional structure. Assume a can take any value in $[0, 1]$, whereas all other variables have two possible realizations: high (denoted 1) or low (denoted 0).[13] Let $u(a, w) = w - \kappa(a)$, where κ is a twice-differentiable, increasing and weakly convex cost function, with $\kappa'(0) = 0$, and $\kappa'(1) \geq 1$. Finally, assume $p(s = 1 \mid a, \theta) = a\theta$ and $p(w = 1 \mid \theta, s) = \theta\beta_s$, where $\beta_1 > \beta_0$. Thus, high ability is necessary for both school and labor-market success; conditional on high ability, success at school increases the probability of a high wage. Denote $p(\theta = 1) = \delta > 0$. If the parent had rational expectations, he would choose a to maximize $\delta[a\beta_1 + (1 - a)\beta_0] - \kappa(a)$. The optimal action a^* would be given by the first-order condition, $\kappa'(a^*) = \delta[\beta_1 - \beta_0]$. The following result characterizes personal equilibria under the parent's subjective DAG.

[13] Although the formal analysis so far has assumed finite action sets, the extension to a continuum of actions is straightforward.

Proposition 4 *Given the specification of R, u and $(p(\theta, s, w \mid a))_{a,\theta,s,w}$ in this sub-section, every personal equilibrium has the following structure: the parent assigns probability one to some action a^{**} that solves the equation*

$$\kappa'(a^{**}) = \delta \left[\beta_1 - \beta_0 \cdot \frac{\delta(1 - a^{**})}{\delta(1 - a^{**}) + 1 - \delta} \right] \tag{6}$$

If κ' is either weakly convex or weakly concave, a^{**} is unique.

According to Equation (6), the parent over-invests relative to the rational-expectations benchmark. The reason is that he interprets the positive correlation between s and w as a pure *causal effect* of s on w, whereas in reality the correlation is partly due to the confounder θ. This generates an upward bias in the parent's perceived marginal benefit from education (given by the R.H.S), thus leading him to over-invest.

In addition, the perceived marginal benefit is a function of the equilibrium investment a^{**}. The reason is that while the parent's subjective DAG postulates that $w \perp_R a \mid s$, the true DAG violates this property; as a result, the perceived causal effect of s on w is sensitive to the parent's equilibrium behavior. In particular, higher long-run investment *raises* the perceived marginal benefit of education. To see why, note that success at school implies high latent ability, and the two jointly imply high expected wage. In contrast, poor school performance is a strong indicator of low ability only when parental investment is high. Therefore, the measured gap $E(w \mid s = 1) - E(w \mid s = 0)$ increases with long-run investment. This "strategic complementarity" effect can lead to multiple personal equilibria, depending on how the curvature of κ' behaves.[14]

Comment: Direct measurement of $a - w$ correlation
In this example, the DM ends up mispredicting the effect of education on wages - i.e., $p_R(w \mid a) \neq p(w \mid a)$ for some a, w (likewise in the Dieter's Dilemma, $p_R(h \mid a) \neq p(h \mid a)$ for some a, h). This begs the question: if the

[14]The multiplicative form of $p(s \mid a, \theta)$ and $p(w \mid \theta, s)$ magnifies the strategic complementarity effect, but it is not necessary for it. A similar effect would appear under an additive specification of these conditional probabilities, for sufficiently high values of a.

DM's payoff is purely a function of a and w, why does he not test directly for the effect of a on w?

There are several possible answers. First, my modeling approach retains the traditional separation between figuring out the feasible set and choosing from it. Our DM first attempts to fully specify his causal model in order to understand what is feasible. Only then does he proceed to the choice stage (much as an econometrician estimates a given model, independently of the particular decision problem that will be subsequently faced). Second, processing data carries an implicit cost, and our DM is willing to incur it only if he thinks it might change his beliefs. If the DM puts sufficient faith in his causal model, he will not find it worthwhile to test for $a-w$ correlation. Furthermore, this test may simply be infeasible at the time of choice. For example, in the context of the Dieter's Dilemma, it is realistic to assume that direct evidence regarding the effect of nutrition on a target health variable becomes available long after the accumulation of observational data regarding the correlation between the chemical level and health.

3.3 Coarseness II: Public Policy

The previous sub-section examined a DM who ignores the causal effects of an exogenous confounding variable. Now I turn to a DM who misunderstands the role of an *endogenous* variable in the causal chain from his action to some target variable. Failure to account for endogeneity is a common target of economists' criticism of public policy. When a government evaluates tax or tariff reforms, it may take certain consumption or investment quantities as given, whereas in fact they are endogenous variables that respond to changes in policy. Likewise, when a higher-education regulator considers changing the minimal accreditation requirement of some degree, he may neglect the possible effect on the composition of the applicant pool, and therefore on graduates' ultimate quality.

In macroeconomics, neglect of the response of private-sector expectations to policy changes was a primary object of the Lucas Critique (Lucas (1976)). Sargent (2001) modeled a central bank that commits this "sin": it evaluates

policy according to a classical Phillips curve that ignores private-sector expectations (thus implicitly holding them fixed), whereas the true process is given by a Phillips curve that incorporates rational private-sector expectations. The simple example analyzed in this sub-section is based on Sargent's model: it distills its underlying causal misperception, using a different parameterization that generates new insights.

Formally, there are four variables, a, y, e, z, where a represents the government's policy; y and z represent two different "macro" variables; and e represents the private sector's expectation of y. Assume that u is purely a function of y and z. The true DAG R^* is

$$\begin{array}{ccccc} a & \rightarrow & y & \rightarrow & z \\ & \searrow & & \nearrow & \\ & & e & & \end{array} \tag{7}$$

If the government had rational expectations, it would choose a to maximize

$$\sum_y \sum_e p(y \mid a) p(e \mid a) \sum_z p(z \mid y, e) u(y, z)$$

Now suppose that the government's subjective DAG R differs from R^* by removing at least one of the two links of e (or by eliminating this node and its links altogether). In personal equilibrium, if $p(a) > 0$, then a maximizes

$$\begin{aligned} \sum_y p(y \mid a) \sum_z p(z \mid y) u(y, z) \qquad (8) \\ = \sum_y p(y \mid a) \sum_z \left(\sum_{a'} \sum_e p(a' \mid y) p(e \mid a') p(z \mid y, e) \right) u(y, z) \end{aligned}$$

The government's failure to fully account for the causal channel that passes through e means that when it calculates $p(z \mid y)$, it effectively sums over a and e, weighted according to the government's long-run behavior.

Let us impose additional structure. The variables a and y take values in $\{0, 1\}$; $p(y = 1 \mid a) = a\beta$, where $\beta > 0$ is a parameter that captures the government's ability to control y. The private sector has rational expecta-

tions, such that for every a, $e = E(y \mid a) = a\beta$ with probability one. The variable z is a deterministic function of y and e, given by $z = y + \delta e$, where $\delta \neq 0$. Finally, $u(y, z) = z - \kappa y$, where $\kappa > 0$ is the government's rate of substitution between the two macro variables. If the government had rational expectations, it would choose $a = 1$ whenever $\kappa < 1 + \delta$.

The parameter δ captures a distinction that turns out to be important for our analysis. When $\delta > 0$, private-sector expectations have a *reinforcing* effect. For example, $a = 1$ represents an intervention that is meant to prevent currency depreciation; y is the direct effect of this action. Private-sector expectations respond to this intervention and boost demand for the currency, creating a "multiplier effect". Conversely, when $\delta < 0$, private-sector expectations have a *countervailing* effect. For instance, $a = 1$ represents monetary expansion; y represents inflation and z represents real output. As in Sargent (2001), the government is averse to inflation, but regards it as a possible means for increasing real output via a Phillips effect; in reality, this effect exists only to the extent that inflation is not anticipated by the private sector.

Proposition 5 *Given the specification of R, u and $(p(y, e, z \mid a))_{a,y,e,z}$ in this sub-section, the set of personal equilibria is as follows. Denote $p(a = 1) = \alpha$. (i) When $\kappa \leq 1$, there is an equilibrium in which $\alpha = 1$. (ii) When $\kappa \geq 1+\beta\delta$, there is an equilibrium in which $\alpha = 0$. (iii) When κ is between 1 and $1+\beta\delta$, there is an equilibrium with*

$$\alpha = \frac{\beta\delta + 1 - \kappa}{\beta\delta + \beta(1 - \kappa)}$$

There exist no other equilibria.

Thus, as expected, the government's neglect of a reinforcing (countervailing) effect biases its behavior toward $a = 0$ ($a = 1$). A less obvious qualitative difference concerns equilibrium multiplicity: when $\delta > 0$ there is a unique personal equilibrium, whereas multiple equilibria are possible under $\delta < 0$. The intuition is as follows. The government's causal model implies that the distribution over z is independent of its action *conditional* on y. This is

correct under $y = 1$, a realization that can only occur if the government has played $a = 1$. However, it is *incorrect* under the realization $y = 0$. This leads the government to form a biased estimate of $E(z \mid a = 0)$. According to the true model, $a = 0$ implies $z = 0$ with certainty, whereas

$$E_R(z \mid a = 0) = E(z \mid y = 0) = p(a = 1 \mid y = 0) \cdot \beta\delta \tag{9}$$

The bias is positive (negative) when private-sector expectations have a reinforcing (countervailing) effect.

Imagine that α goes up. Since $p(a = 1 \mid y = 0)$ increases with α, the bias given by (9) becomes more severe. When $\delta > 0$, this means that the government's overvaluation of $a = 0$ worsens; hence its perceived incentive to play $a = 1$ weakens. Thus, when private-sector expectations have a *reinforcing* effect, the model exhibits strategic *substitutability*: the pressure to play $a = 1$ decreases with the long-run frequency of this action. This in turn implies equilibrium uniqueness. In contrast, the case of $\delta < 0$ generates "strategic complementarity"; hence the possibility of multiple equilibria.

Another noteworthy feature is that while the parameter β plays no role in the rational-expectations case, it matters for personal equilibrium given the government's subjective DAG. The reason is that a change in β affects the government's biased estimate of $E(z \mid y = 0)$.

4 General Analysis

In this section I characterize rationality properties of personal equilibrium, using elementary tools from the Bayesian-networks literature. Results are stated in terms of structural features of true and subjective DAGs, involving no parametric restrictions on u and $p(y \mid a)$. For expositional ease, I present the simplest versions of the results. In Sections 4.1 and 4.2, I use x_1 (rather than a) to denote the DM's action. I employ two graph-theoretic definitions. A subset $M \subseteq N$ is a *clique* in R if $i\tilde{R}j$ for every distinct $i, j \in M$. A clique M is *ancestral* if $R(i) \subset M$ for every $i \in M$.

4.1 Consequentialist Rationality

In the illustrations, individual optimization under a misspecified subjective DAG led to genuine equilibrium effects (mixed unique equilibrium, multiple equilibria). Equilibrium effects would not arise if the DM's perception of the mapping from actions to consequences were invariant to long-run action frequencies. I refer to such invariance as "consequentialist| rationality".

Definition 3 *A DAG R is **consequentialistically rational** with respect to a true DAG R^* if the following holds for every pair of objective distributions p, q that are consistent with R^*: if $p(x_2, ..., x_n \mid x_1) = q(x_2, ..., x_n \mid x_1)$ for every x, then $p_R(x_2, ..., x_n \mid x_1) = q_R(x_2, ..., x_n \mid x_1)$ for every x.*

Consequentialistic rationality requires that if we modify an objective distribution p that is consistent with R^* only by changing its marginal over x_1 - without changing the stochastic mapping from x_1 to $x_2, ..., x_n$ - then the DM's perception of this mapping should remain unchanged as well. When R is consequentialistically rational with respect to R^*, we can rewrite the definition of personal equilibrium as a maximization problem, because $p_R(x_2, ..., x_n \mid x_1)$ is invariant to $(p(x_1))_{x_1}$. When consequentialist| rationality is violated, individual behavior will exhibit equilibrium effects for *some* specifications of p, u. Clearly, any DAG is consequentialistically rational with respect to itself. From now on, I will take it for granted that $R \neq R^*$.

Proposition 6 *The subjective DAG R is consequentialistically rational with respect to R^* if and only if for every $i > 1$, $1 \notin R(i)$ implies $x_i \perp_{R^*} x_1 \mid x_{R(i)}$.*

Thus, a necessary and sufficient condition for consequentialist| rationality is the following: whenever R fails to include the DM's action as an immediate cause of some other variable x_i, it must be the case that for *every* distribution that is consistent with R^*, x_i is independent of x_1 conditional on $x_{R(i)}$. The proof consists of simply writing down the explicit formula for $p_R(x_2, ..., x_n \mid x_1)$ and checking its individual terms. When R^* is fully connected - i.e.,

when the domain of p is unrestricted - the condition for consequentialist| rationality becomes $1 \in R(i)$ for every $i > 1$.

To illustrate this result, recall the Dieter's Dilemma, where true and subjective DAGs were $R^* : 1 \to 3 \leftarrow 2$ and $R : 1 \to 3 \to 2$ (using variable indices rather than intuitive labels). Since $1 \in R(3)$ and $1 \notin R(2) = \{3\}$, Proposition 6 implies that we only need to check whether $x_2 \perp_{R^*} x_1 \mid x_3$. As observed in Section 3.1, this property does not hold; hence consequentialist| rationality is violated.

Alternatively, for the same $R^* : 1 \to 3 \leftarrow 2$, suppose the DM's subjective DAG is $R : 1 \to 3 \quad 2$. This DM is "fully coarse/cursed" in the sense of Eyster and Rabin (2005) and Jehiel and Koessler (2008): he fails to perceive the effect of the exogenous state x_2 on the final consequence x_3. The DM will choose x_1 to maximize

$$\sum_{x_2,x_3} p_R(x_2, x_3 \mid x_1)u(x_1, x_2, x_3) = \sum_{x_2}\sum_{x_3} p(x_2)p(x_3 \mid x_1)u(x_1, x_2, x_3)$$

If the DM had rational expectations, $p(x_3 \mid x_1)$ would be replaced with $p(x_3 \mid x_1, x_2)$ in this expression. To see why consequentialist| rationality holds, note that $R(2) = \varnothing$ and $R(3) = \{1\}$; by Proposition 6 we only need to check that $x_2 \perp_{R^*} x_1$, which clearly holds.

When R and R^* are large, checking the conditional-independence conditions of Proposition 6 can be a daunting task. However, it is greatly facilitated by a Bayesian-networks tool called *d-separation*, which provides a linear-time algorithm for checking whether a conditional independence property is satisfied by all the distributions that are consistent with a given DAG. Appendix B presents the tool and illustrates its applicability in the present context.

4.2 Behavioral Rationality

Consequentialistic rationality is a weak rationality requirement, which allows the DM to choose an objectively sub-optimal action. In this sub-section I look for a structural property of the DM's subjective DAG that will ensure

fully rational behavior in terms of objective payoffs. I impose no restriction on the set of possible objective distributions - i.e., the true DAG R^* is fully connected. Instead, I restrict the set of possible utility functions: there exists a strict subset $M \subset N$, $1 \in M$, such that u is purely a function of x_M (i.e., it is constant in x_{N-M}). All the examples in Section 3 had this feature.

Definition 4 *A DAG R is **behaviorally rational** if in every personal equilibrium p, $p(x_1) > 0$ implies $x_1 \in \arg\max_{x_1'} \sum_{x_{-1}} p(x_{-1} \mid x_1')u(x_1', x_{-1})$.*

As a first step toward characterizing behavioral rationality, I examine the following question: when is the DM's subjective marginal distribution over some collection of variables guaranteed to be unbiased, despite his misspecified subjective DAG? By definition, if R is not fully connected, then there exists an objective p such that $p_R \neq p$. However, p_R may agree with p on some projections. The next result characterizes which ones.

Proposition 7 (Spiegler (2015)) *Let R be a DAG and let $S \subset N$. Then, $p_R(x_S) \equiv p(x_S)$ for every p if and only if S is an ancestral clique in some DAG in the equivalence class of R.*

For instance, let $n = 3$ and $R : 1 \to 2 \leftarrow 3$. Then, $p_R(x_2)$ is biased for some p, because the node 2 is not ancestral in R (and no other DAG is equivalent to R, by Proposition 1). In contrast, when $R : 1 \to 2 \to 3$, $p_R(x_3)$ coincides with $p(x_3)$ because 3 is ancestral in the equivalent DAG $R' : 3 \to 2 \to 1$. In both examples, $p_R(x_1, x_3)$ does not coincide with $p(x_1, x_3)$ for every p, because the nodes 1 and 3 are not linked and therefore cannot form a clique (let alone an ancestral one) in any equivalent DAG.

Proposition 8 *The DM is behaviorally rational if and only if M is an ancestral clique in some DAG in the equivalence class of R.*

Thus, when all payoff-relevant variables are causally linked and have no other cause (according to some DAG in the equivalence class of the DM's subjective DAG), the DM is behaviorally rational. Otherwise, there are specifications of p, u for which his behavior is inconsistent with rational expectations.

Application: When can coarseness lead to sub-optimal behavior?
The modeling framework enables us to formulate the following question: When does a specific error of causal attribution (captured by a basic operation on the true DAG) violate behavioral rationality? The following is an example of such an exercise. Let $M = \{1, n\}$. Recall that the true DAG R^* is fully connected, and assume that n is a terminal node in R^* (i.e., $n \notin R^*(i)$ for all $i < n$). The interpretation is that the variable x_n is an ultimate consequence, such that the DM's payoff depends only on his action and the ultimate consequence. Now suppose that the DM's subjective DAG R differs from R^* only by omitting a *single* link.

Proposition 9 *Suppose that the DM's subjective DAG R departs from the true, fully connected DAG R^* by omitting one link $i \to j$. Then, the DM is behaviorally rational if and only if $j = n$ and $i \neq 1$.*

Thus, even if the DM neglects the direct causal effect of some intermediate variable x_i ($i \neq 1, n$) on the ultimate consequence x_n, he is behaviorally rational. In any other case, there are specifications of $p(x_2, ..., x_n \mid x_1)$ and u for which the DM's error will have payoff implications.

To illustrate this result, let $N = \{1, 2, 3\}$. When $R : 1 \to 3 \leftarrow 2$ (omitting the link $1 \to 2$ from R^*), the DM regards x_1 and x_2 as independent causes of x_3; if, however, x_2 is in fact a deterministic function of x_1, the DM may err by "double-counting" the effect of x_1 on x_3. When $R : 1 \to 2 \to 3$ (omitting the link $1 \to 3$ from R^*), the DM regards x_3 as independent of x_1 conditional on x_2; if, however, x_2 is in fact an independent variable, the DM will fail to perceive any effect of x_1 on x_3. Finally, when $R : 2 \leftarrow 1 \to 3$ (omitting the link $2 \to 3$ from R^*), the DM correctly estimates the total causal effect of x_1 on x_3, even though he fails to realize that it consists of direct and indirect effects (the latter runs through x_2).

4.3 Payoff Ranking of DAGs

A more complete subjective DAG represents a more thorough understanding of correlation structures; hence it intuitively captures "more rational" expectations. Does this mean that it will always lead to better objective performance? The following example shows the answer to be negative. Let $n = 4$, and suppose R is fully connected and contains the links $1 \to 2 \to 3 \to 4$. Suppose further that u is purely a function of x_1 and x_4. Obtain the DAG R' by removing the link $2 \to 3$ from R. By Proposition 9, R' is not behaviorally rational - i.e., it is weakly dominated by R in terms of expected performance. Now obtain R'' by removing the link $2 \to 4$ from R'. It is easy to verify that R'' is equivalent to a DAG in which $\{1, 4\}$ is an ancestral clique. By Proposition 8, a DM whose subjective DAG is R'' performs exactly like a DM whose subjective DAG is R. Thus, removing a link from the DM's subjective DAG can result in better performance for some u, p.[15]

I now examine the question with some generality. I return to the notation $x_1 = a$, $x_{-1} = y$. For expositional simplicity, suppose that 1 is an isolated node in all relevant true and subjective DAGs. This guarantees consequentialist rationality.

Definition 5 *Let R, R' be two DAGs in which the node that corresponds to the DM's action is isolated. We say that R is* ***more rational*** *than R' if for every p, u, a, a', the pair of inequalities*

$$\begin{aligned} \sum_y p_R(y)u(a,y) &> \sum_y p_R(y)u(a',y) \\ \sum_y p_{R'}(y)u(a',y) &> \sum_y p_{R'}(y)u(a,y) \end{aligned}$$

implies

$$\sum_y p(y)u(a,y) > \sum_y p(y)u(a',y)$$

[15] Eyster and Piccione (2013) made an observation in the same spirit, in the context of their model of competitive asset markets in which traders hold diversely coarse theories.

That is, if R ranks a above a' and R' ranks a' over a, then the rational-expectations payoff ranking of the two actions necessarily sides with R. When R is fully connected and R' is not, the property holds trivially. The following result shows that this is the only case in which two DAGs can be unambiguously ranked in terms of their expected performance.

Proposition 10 *Let R, R' be two DAGs that are not fully connected (and in which the node that corresponds to the DM's action is isolated). Then, neither DAG is more rational than the other.*

The proof of this result is a simple application of Farkas' Lemma. Domination implies a linear relation between p_R and $p_{R'}$, which can only mean that R and R' are equivalent.

5 Variations and Relation to Other Concepts

I begin the section with two extensions of the modeling framework.

A "Mixed" DAG representation
Define the DM by a *probability distribution* λ over DAGs, such that his subjective belief is given by the following extension of the DAG representation:

$$p_\lambda(x) = \sum_R \lambda(R) p_R(x) \tag{10}$$

This representation captures *magnitudes* of belief errors. When λ mixes between the true DAG R^* and some other DAG R, the magnitude of the DM's error increases with $\lambda(R)$. An example of a mixed representation is "partial cursedness" (Eyster and Rabin (2005)).

Definition 5 is extendible to mixed DAG representations. Let λ^* be a distribution that assigns probability one to fully connected DAGs. Consider two distributions λ, λ' that do not assign probability one to fully connected DAGs, and satisfy $\lambda = \alpha\lambda^* + (1-\alpha)\lambda'$, where $\alpha \in (0,1)$. It is easy to see that λ is more rational than λ'. Thus, mixed DAG representations enable us to rank some types according to their performance.

DAG heterogeneity

The motivation for personal equilibrium was that the DM confronts a "historical database" that includes actions taken by other DMs who faced the decision problem. Now relax the assumption that *all* DMs in this population share the same subjective causal model. Let $\beta(R)$ denote the fraction of DMs whose subjective DAG is R, and let $p(a \mid R)$ be the probability that they play the action a. Then, $p(a) = \sum_R \beta(R)p(a \mid R)$, while $p(y \mid a)$ is the same for all DM types. The extended definition of personal equilibrium requires that for every R for which $\beta(R) > 0$, if $p(a \mid R) > 0$ then a maximizes $\sum_y p_R(y \mid a)u(a, y)$. This extension has interesting implications for the Dieter's Dilemma. Assume that a fraction of the DM population holds the correct causal model $d \rightarrow c \leftarrow h$. This lowers the overall probability of dieting, thus strengthening the empirical $c - h$ correlation. As a result, DMs with the incorrect DAG will have a stronger tendency to diet.

Let us now turn to the relation between the Bayesian-network framework and existing equilibrium models with non-rational expectations. Most of this literature proceeded by postulating game-theoretic solution concepts that capture different aspects of limited understanding of correlations. I briefly discuss the relation between some of these concepts (as defined for *static* games) and the Bayesian-network formalism. For expositional simplicity, I consider two-player games and examine the behavior of one player, to whom I refer as the DM. Throughout the discussion, θ represents a state of Nature that is known by the opponent, and z denotes the game's outcome (induced by the two players' actions).

Analogy-based expectations

This concept was introduced by Jehiel (2005) in the context of extensive-form games with complete information, and was later adapted to static Bayesian games by Jehiel and Koessler (2008). The latter can be translated into the Bayesian-networks language, as the following example illustrates. The true DAG R^* is $a \rightarrow z \leftarrow \theta \rightarrow e$, where e represents the "analogy class" to which θ belongs. For a DM with rational expectations, e is irrelevant and can be omitted from his subjective model altogether. The DM's subjective DAG R

is $a \to z \leftarrow e \leftarrow \theta$ - i.e., R differs from R^* only by changing the origin of the link that goes into z, from θ to e. Thus, the DM interprets the opponent's equilibrium behavior as if it were a measurable function of the analogy class, rather than the actual state of Nature. To use the terms of Section 2, R exhibits a "misattribution error" relative to R^*. The definition of Analogy-Based Expectations Equilibrium (ABEE) requires the DM's action to be subjectively optimal with respect to p_R. Thus, we can reduce individual best-replying under ABEE in *static* Bayesian games to subjective optimization under the misspecified DAG R.[16]

Naive behavioral equilibrium

Esponda (2008) introduced a solution concept called "naive behavioral equilibrium", which can be described as a *refinement* of personal equilibrium for a suitable specification of R and R^*. Let R^* be

$$\begin{array}{ccccccc} a & \to & z & \leftarrow & \theta \\ & & & \searrow & \downarrow \\ & & & & f \end{array} \tag{11}$$

where f is a "learning feedback" variable.[17] The DM's utility can be written as a function of a, z, θ. The DM's subjective DAG R differs from R^* by removing the link $\theta \to z$. Like personal equilibrium in the present framework, naive behavioral equilibrium requires the DM's action to be subjectively optimal with respect to p_R. However, it goes further by requiring $p_R(f) = p(f)$ - i.e., the subjective marginal distribution over the feedback variable should be unbiased. Because f is not an ancestral node in any DAG in the equivalence class of R, Proposition 7 implies that Esponda's additional requirement is not vacuous. If there exists no personal equilibrium p for which $p_R(f) \equiv p(f)$, the set of naive behavioral equilibria will be empty. In fact, if we substitute $z = s$ and $f = w$, then (11) is none other than the

[16] Esyter and Piccione's (2013) formalism can be similarly translated. Suppose that u is purely a function of x_M, $M \subset N$. The DM's subjective DAG omits some of the nodes outside M. This DAG corresponds to Eyster and Piccione's notion of an incomplete theory. The DM in their model best-replies to the subjective belief p_R.

[17] Esponda (2008) employs multiple feedback variables. My simplification does not violate the spirit of Esponda's approach.

true DAG of Section 3.2; under the parameterization of that example, naive behavioral equilibrium fails to exist.

For other parametrizations, Esponda's additional requirement is satisfied and naive behavioral equilibrium coincides with personal equilibrium. 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. This example can be translated as follows: a, θ, z and f represent the buyer's offer, the seller's valuation, the final allocation and the buyer's gross payoff from it ($f = 0$ when there is no trade, and $f = \theta + b$ when trade occurs, where b is a constant). The buyer fails to realize that the probability of trade is affected by the seller's valuation. Personal equilibrium in this example has the feature that $p_R(f)$ is unbiased; hence it coincides with naive behavioral equilibrium.[18]

Naive behavioral equilibrium is a concept in the tradition of self-confirming equilibrium (Fudenberg and Levine (1993)). In the present context, self-confirming equilibrium would require the DM to play a best-reply to a subjective belief q that satisfies $q(f) = p(f)$ - that is, the only restriction on the DM's subjective belief is that it is consistent with the feedback. The present framework takes a different approach to modeling feedback: rather than adding explicit feedback variables, it implicitly assumes that the DM's feedback consists of the correlations that identify his causal model.

Partial cursedness

Suppose the true DAG is $R^* : a \rightarrow z \leftarrow \theta$. Let R differ from R^* only by removing the link from θ into z. The DM is characterized by a *mixed* DAG representation (10) that assigns weight χ to R and weight $1 - \chi$ to R^*. This DM is "partially cursed" in the sense of Eyster and Rabin (2005); the parameter χ captures the extent to which the DM neglects the relationship between his opponent's action and information. The discussion at the end of Section 4.3 implies that DMs with different values of χ can be unambiguously ordered in terms of their objective payoff performance: a DM with a lower χ is "more rational".

[18] Esponda and Pouzo (2014a) study an electoral model, in which individual voters' behavior can be translated into an informed-DM variation on the current specification.

$S(K)$ equilibrium

Osborne and Rubinstein (1998) presented a solution concept in which each player postulates a direct mapping from his action to the game's payoff-relevant outcome, without forming an explicit belief regarding additional variables. The player estimates this mapping by sampling each action K times against the opponents' distribution, and selecting the action that performs best in his sample. The player's misperception can be described in terms of the Bayesian-network formalism. For instance, suppose that the true DAG R^* is given by (11) and that u is purely a function of f. The DM's subjective DAG is $R : a \to f$. In our model, the DM's behavior would be consistent with rational expectations. However, this is because he perfectly learns $p(f \mid a)$, whereas in Osborne and Rubinstein (1998), he uses a *finite sample* to estimate it and naively neglects the sampling error.

The organizing principle behind most concepts in this literature is that the DM's beliefs are statistically correct with respect to partial feedback about specific marginal or conditional probabilities; different concepts assume different types of feedback. The Bayesian-network formalism systematizes and generalizes this principle, via the requirement that the DM correctly estimates the conditional distributions that identify his causal model.

Thus, the Bayesian-network representation of non-rational expectations can be viewed as a *unifying* framework. What is the value of this unification? First, it deepens our understanding of the relation between concepts. Second, as the Dieter's Dilemma taught us, the framework accommodates novel belief distortions that previous concepts did not address: despite having an infinite amount of data, the dieter ends up believing in a correlation that does not exist in reality.[19] Third, the framework provides new tools for analyzing implications of existing concepts: the analysis in Section 4.2 enables us to see the reason that naive behavioral equilibrium may fail to exist, and the analysis in Section 4.3 implies that DMs with different analogy partitions as in Jehiel and Koessler (2008) cannot be ranked in terms of performance. As

[19] Most existing concepts capture *underestimation* of correlations due to coarseness. The $S(K)$ model generates spurious correlations, due to neglect of sampling error in *small* samples.

we will see in the concluding section, the framework has additional potential benefits in multi-agent models.

6 Concluding Remarks

The modeling framework developed in this paper enables us to analyze behavioral implications of systematic errors in statistical and causal reasoning - mistaking correlation for causation, ignoring confounding variables, etc. Statistics teachers invest considerable effort to "cure" students of such fallacies. Indeed, one motivation behind the Bayesian-network literature has been to systematize *correct* causal reasoning. Instead, this paper employed the tool *descriptively*, to capture the very errors statisticians warn us against.

Throughout this paper, I interpreted the DM's subjective DAG as an explicit causal model. I now discuss several alternative interpretations.

Objective data limitations. Certain DAGs can be interpreted as representations of objective feedback limitations that are faced by the DM as he tries to learn p. Imagine that the DM only manages to learn the marginals of p over some collections of variables, and that he wishes to extend these marginals to a fully specified distribution over X. A result by Hajek et al. (1992) implies that when $R(i)$ is a clique for every $i \in N$, there exists $\mathcal{S} \subset 2^N$ such that p_R is the *maximal-entropy extension* of the known marginals of p over x_S, $S \in \mathcal{S}$. Moreover, $\mathcal{S}$ is the set of maximal cliques in R. In this sense, the DAG representation (2) can be justified as the outcome of systematic extrapolation from limited data. In Spiegler (2015), I further develop the limited-feedback foundation for the DAG representation, by considering a more "behaviorally motivated" method of extrapolation.

Limited ability to perceive statistical patterns. In Section 2, I used a "Q&A story" to illustrate the causal interpretation of R. However, the questions can be interpreted differently, as an attempt to find statistical patterns in the data. The DM is unable to grasp the multivariate distribution p in its totality, and poses a sequence of partial queries that examine slices of p. The DM behaves *as if* he has an explicit subjective causal model, but the

essence of this bounded rationality is that he *fails to ask the right questions* about p. The DAG R captures the questions that he can think of. For instance, in the example of Section 3.2, observable variables are realized in fixed chronological order: parental investment comes first, then the school outcome is realized, followed by the wage. It is therefore natural for the parent to pose questions that track this chronological order: What is the likelihood of school success as a function of parental investment? How do wages vary with school performance? These are the very questions that identify the parent's subjective DAG.[20]

Bayesian learning with misspecified priors. Esponda and Pouzo (2014b - EP henceforth) proposed a game-theoretic framework, where each player is characterized by a "subjective model" - a set of stochastic mappings from his action a to a primitive set of payoff-relevant consequences y (for simplicity, assume players are uninformed). EP's definition of equilibrium requires the player's subjective distribution over y to be the closest (in the set implied by his subjective model) to the true equilibrium distribution; distance is measured by a weighted Kullback-Leibler divergence. EP justify this concept as the steady state of a process of Bayesian learning by forward-looking agents, extending classical results on Bayesian learning with misspecified priors (Berk (1966)). Personal equilibrium in the present paper can be viewed as an EP equilibrium in single-player games, where the player's subjective model is the set of all conditional probabilities $(p_R(y \mid a))_{a,y}$ that are consistent with the DAG R.[21]

Although this paper has focused entirely on individual choice, the Bayesian-network representation allows us to capture *interactive* situations in which different agents view the same interaction through the prisms of different subjective causal models. Existing notions of non-rational-expectations equilibrium are typically presented as distinct solution concepts in some class of games. The current framework reduces the element of non-rational expec-

[20] Esponda and Vespa (2014) interpret a pivotal-voting experiment in this spirit.

[21] Relatedly, there is a strand in the literature that views boundedly rational agents as "time-series econometricians" who work with a misspecified model (e.g., Bray (1982), Cho, Sargent and Williams (2002), Rabin and Vayanos (2010)).

tations to individual agents' *types* (their subjective DAGs) within a single modeling framework. This reduction has several potential advantages. First, it enables us to analyze interactions among agents who commit different kinds of errors. Second, a model in which agents are characterized by distinct DAGs exhibits "structured belief heterogeneity", because agents' beliefs are different deterministic transformations of the same objective distribution (e.g., when the latter is consistent with the empty DAG, all agents have correct beliefs). Third, since the DAG representation is not tied to a particular economic model, it can be incorporated into diverse classes of models (games, competitive markets). Finally, it provides a language for studying "high-order" reasoning about boundedly rational expectations. By incorporating one agent's DAG as a variable in another agent's causal model, we can express statements such as "player i does not understand the correlation between player j's understanding of correlations and his information". This element is beyond the reach of current approaches, and I plan to explore it in future research.

Appendix A: Proofs

Proposition 2

Fix $(p(y \mid a))_{a,y}$. For a fixed $\varepsilon \in (0,1)$, let Q^ε be the set of distributions $q \in \Delta(A)$ such that $q(a) \geq \varepsilon$ for every a. Denote $p = (q, (p(y \mid a))_{a,y})$, and define p_R accordingly. Define

$$BR(p) = \arg\max_{q \in Q^\varepsilon} \sum_a q(a) \sum_y p_R(y \mid a) u(a,y)$$

If q is an ε-perturbed personal equilibrium, then $q \in BR(q)$. Because $p_R(y \mid a)$ is continuous in q, BR is continuous as well. Also, the target function in the definition of BR is linear in q; hence $BR(p)$ is a convex set. Since the set Q^ε is compact and convex, BR has a fixed point, by Kakutani's theorem. Therefore, an ε-perturbed personal equilibrium exists for any $\varepsilon > 0$. By standard arguments, there is a convergent sequence of ε-perturbed personal equilibria.

Proposition 4

The parent's objective function is $\sum_{s=0,1} p(s \mid a)p(w=1 \mid s) - \kappa(a)$. By our assumptions on p, $\theta = 0$ implies that $s = w = 0$ with certainty, whereas $\theta = 1$ implies that $s = 1$ with probability a. Let μ denote the parent's probability measure over actions a. This implies the following conditional probabilities: $p(s=1 \mid a) = \delta a$, $p(w=1 \mid s=1) = \beta_1$, and

$$p(w=1 \mid s=0) = \frac{\delta\beta_0 \int_{a'} d\mu(a')(1-a')}{1-\delta+\delta \int_{a'} d\mu(a')(1-a')} = \gamma\beta_0$$

The derivative of the parent's objective function is $\delta(\beta_1 - \gamma\beta_0) - \kappa'(a)$. The first term, which represents the parent's perceived marginal benefit from education, lies in $(0,1)$. The parent takes it as given when choosing a. The second term is continuous and strictly increasing, with $\kappa'(0) = 0$ and $\kappa'(1) > 1$. Therefore, there is a unique best-reply a^{**}, which means that μ assigns probability one to a^{**}. In equilibrium, a^{**} solves the first-order condition $\kappa'(a^{**}) = \delta(\beta_1 - \gamma\beta_0)$, where

$$\gamma = \frac{\delta(1-a^{**})}{1-\delta+\delta(1-a^{**})} \tag{12}$$

This gives the equation (6). Observe that $\kappa'(a^{**})$ is continuous in a^{**}, with $\kappa'(0) = 0$, $\kappa'(1) \geq 1$, whereas $\delta(\beta_1 - \gamma\beta_0)$ is a strictly convex and increasing function of a^{**}, which attains values strictly between 0 and 1. Therefore, the two functions must cross at least once, such that (6) has a solution. When κ' is either weakly convex or weakly concave, the two functions cross exactly once, such that the solution is unique.

Proposition 5

Denote $p(a=1) = \alpha$. To calculate personal equilibria, we need to compute the conditional probabilities that appear in (8). All of them were defined up-front, except $p(a \mid y)$, which is given by $p(a=1 \mid y=1) = 1$ and

$$p(a=0 \mid y=0) = \frac{1-\alpha}{\alpha(1-\beta)+1-\alpha} = \gamma$$

Fix α. The government's evaluation of each action a is

$$\sum_y p(y \mid a)[E(z \mid y) - \kappa y]$$

To calculate this expression for each a, let us first derive $E(z \mid y)$. Consider the case of $y = 1$ first. Because $p(a = 1 \mid y = 1) = 1$,

$$E(z \mid y = 1) = \sum_e p(e \mid a = 1)(1 + \delta e) = 1 + \delta\beta$$

Now consider the case of $y = 0$:

$$E(z \mid y = 0) = \gamma \cdot 0 + (1 - \gamma) \cdot \sum_e p(e \mid a = 1) \cdot \delta e = (1 - \gamma)\delta\beta$$

Recall that $p(y = 1 \mid a) = \beta a$. Then, the government's evaluation of $a = 0$ is $(1 - \gamma)\delta\beta$, and its evaluation of $a = 1$ is

$$\beta \cdot [1 + \delta\beta - \kappa] + (1 - \beta) \cdot (1 - \gamma)\delta\beta$$

By the definition of personal equilibrium, $\alpha > 0$ ($\alpha < 0$) only if the government's evaluation of $a = 1$ is weakly above its evaluation of $a = 0$. Plugging the expressions for these evaluations and the definition of γ, we obtain the result.

Proposition 6

The conditional probability $p_R(x_2, ..., x_n \mid x_1)$ can be written as

$$\frac{p(x_1) \cdot \prod_{i=2}^{n} p(x_i \mid x_{R(i)})}{\sum_{x'_2,...,x'_n} p(x_1) \cdot \prod_{i=2}^{n} p(x'_i \mid x_{R(i)\cap\{1\}}, x'_{R(i)-\{1\}})}$$

and the term $p(x_1)$ cancels out. Recall that we are considering a modification of p that changes the marginal of p on x_1, while leaving $p(x_2, ..., x_n \mid x_1)$ intact for all x. We need to check whether the term $p(x'_i \mid x_{R(i)\cap\{1\}}, x'_{R(i)-\{1\}})$ is affected, for $i = 2, ..., n$. If $1 \in R(i)$, the term is clearly unchanged. In

contrast, when $1 \notin R(i)$, the term can be written as

$$p(x'_i \mid x'_{R(i)}) = \sum_{x''_1} p(x''_1) p(x'_i \mid x''_1, x'_{R(i)})$$

The term $p(x'_i \mid x''_1, x'_{R(i)})$ is unaffected by the modification of p. If it is not constant in x''_1, we can find a modification of $p(x'_1)$ for some values of x'_1 such that the expression for $p_R(x_2, ..., x_n \mid x_1)$ will change. In contrast, if the probability is constant in x''_1 (i.e., $p(x'_i \mid x''_1, x'_{R(i)}) = p(x'_i \mid x'_{R(i)})$, the expression for $p_R(x_2, ..., x_n \mid x_1)$ is necessarily unchanged.

Proposition 8

By assumption, 1 is an ancestral node in R. By Proposition 7, $p_R(x_1) \equiv p(x_1)$ for every p. We can write $p_R(x_{M-\{1\}} \mid x_1) = p_R(x_M)/p_R(x_1)$. Therefore, $p_R(x_{M-\{1\}} \mid x_1) \equiv p(x_{M-\{1\}} \mid x_1)$ if and only if $p_R(x_M) \equiv p(x_M)$. By Proposition 7, the latter holds if and only if M is an ancestral clique in a DAG in the equivalence class of R. If $p_R(x_M) \equiv p(x_M)$, the definition of behavioral rationality is trivially satisfied. If $p_R(x_M) \neq p(x_M)$ for some p and x, then we can easily construct u such that the DM will strictly prefer an action that is objectively sub-optimal.

Proposition 9

(i) If $i = 1$ and $j = n$, then $1 \not R n$; hence $\{1, n\}$ cannot be a clique (let alone an ancestral one) in any DAG in the equivalence class of R.

(ii) If $i, j \neq n$, then iRn, jRn and yet i and j are not linked. By Proposition 1, these properties must hold in any DAG in the equivalence class of R, which means that $\{1, n\}$ cannot be an ancestral clique in any such DAG.

(iii) If $j = n$ and $i \neq 1$, then R is a perfect DAG - i.e., $R(k)$ is a clique for every $k \in N$. It is well-known that every clique in a perfect DAG is ancestral in some DAG in its equivalence class (see Spiegler (2015) for details). Note that $R(n) \cup \{n\}$ is a clique because the only link that was removed from the original fully connected DAG was $i \to n$. Therefore, $\{1, n\}$ is an ancestral clique in some DAG in the equivalence class of R.

The result follows from $(i) - (iii)$, by Proposition 8.

Proposition 10

If R and R' are equivalent, the claim holds trivially. Now assume there exist non-equivalent R, R' that are not fully connected, such that R is more rational than R'. Fix p and denote $q = (p_R(y))_y$, $r = (p_{R'}(y))_y$. Both q and r are probability vectors of length $n-1$. I use p^i, q^i, r^i to denote the i-th component of the $(n-1)$ vectors p, q, r. Define the $(n-1)$-vector z as follows: for each y, $z^y = u(a, y) - u(a', y)$. Consider the $(n-1) \times 3$ matrix

$$D = \begin{matrix} r^1 & -q^1 & -p^1 \\ \vdots & \vdots & \vdots \\ r^{n-1} & -q^{n-1} & -p^{n-1} \end{matrix}$$

Let $b = (-\varepsilon, -\varepsilon, -\varepsilon)$ be a vector in $\mathbb{R}^3$, where $\varepsilon > 0$ is arbitrarily small. The assumption that R is more rational than R' thus implies that there exists no z that satisfies the inequality $Dz > b^T$. By Farkas's Lemma, this means that there is a vector $a > 0$ in $\mathbb{R}^3$, such that there $D^T a = 0$ (and since $a > 0$, $ba^T < 0$). Thus, $r^i = \frac{a^2}{a^1} q^i + \frac{a^3}{a^1} p^i$ for every $i = 1, ..., n-1$. Since $\sum_{i=1}^{n-1} r^i = \sum_{i=1}^{n-1} q^i = \sum_{i=1}^{n-1} p^i = 1$ by assumption, $a^1 = a^2 + a^3$, such that the claim holds with $\alpha = a^3/(a^2 + a^3)$.

We have thus established that for any p, we can find $\alpha \in (0, 1)$ such that $p_R = \alpha p + (1-\alpha) p_{R'}$. In particular, for any p that is consistent with R, $p_R = p$ and so the equation reduces to $p_R = p_{R'}$. Likewise, for any p that is consistent with R', $p_{R'} = p$ and again we obtain $p_R = p_{R'}$. It follows that the sets of distributions that are consistent with R and R' are identical, contradicting the assumption that R and R' are not equivalent.

Appendix B: d-Separation

In this appendix I present the concept of d-separation, which is useful for applying the characterization of consequentialist| rationality in Section 4.1. A *path* in a DAG R is a sequence of directly connected nodes in R, ignoring the links' directions.

Definition 6 (Blocking a path) *A subset $D \subset N$ **blocks** a path in R if either of the following two conditions holds:* (1) *the path contains a segment*

of the form $i \to m \to j$ *or* $i \leftarrow m \to j$ *such that* $m \in D$*;* (2) *the path contains a segment of the form* $i \to m \leftarrow j$ *such that neither* m *nor any of its descendants are in* D*.*

To illustrate this definition, consider the DAG $R : 1 \to 2 \leftarrow 3 \to 4$. The path between nodes 1 and 4 is blocked by $\{3\}$ - either because it contains the segment $2 \leftarrow 3 \to 4$ (thus satisfying condition (1)) or because it contains the segment $1 \to 2 \leftarrow 3$ (thus satisfying condition (2), as the node 2 has no descendants and $2 \notin \{3\}$). However, the path between 1 and 4 is not blocked by $\{2\}$, because it does not contain a segment of the form $i \to 2 \to j$ or $i \leftarrow 2 \to j$, and the only segment of the form $i \to m \leftarrow j$ that it contains satisfies $m = 2$.

Definition 7 (*d*-separation) *Let* B, C, D *be disjoint subsets of* N*. We say that* B *and* C *are d-separated by* D *(in a DAG* R*) if* D *blocks every path between any node in* B *and any node in* C*.*

Proposition 11 (Verma and Pearl (1990)) *Let* B, C, D *be disjoint subsets of* N*. Then,* $x_B \perp_R x_C \mid x_D$ *if and only if* B *and* C *are d-separated by* D *in* R*.*

Thus, *d*-separation provides a convenient rule for checking whether a conditional independence property is satisfied by all the distributions that are consistent with a DAG. Moreover, the rule is *computationally simple*: Geiger et al. (1990) presented a linear-time algorithm for checking *d*-separation. Armed with this result, I illustrate Proposition 6, using two specifications from Section 3.

Ignoring a confounder (Section 3.2). The true DAG R^* is given by (5), and $R : a \to s \to w$. Because $a \notin R(w) = \{s\}$, Proposition 6 requires us to check whether $w \perp_{R^*} a \mid s$. To see why this condition fails, observe that $a \to s \leftarrow \theta \to w$ is a path in R^* that connects a and w. This path is *not* blocked by s (as we saw in the example that illustrated blocking), and therefore a and w are not *d*-separated by s.

Ignoring an endogenous effect (Section 3.3). The true DAG R^* is given by (7) and $R : a \to y \to z$. Since $a \in R(y)$ and $a \notin R(z) = \{y\}$, we only need to check that $z \perp_{R^*} a \mid y$ - i.e., that a and z are d-separated by y in R^*. This condition is violated, because R^* contains the path $a \to e \to z$, which is not blocked by y. It follows that consequentialist| rationality is violated.

Appendix C: The Case of an Informed DM

In this appendix I extend the decision model to the case in which the DM receives a signal $x_0 \in X_0$ prior to making his decision. Thus, $x = (x_0, x_1, ..., x_n)$. I use the notations x_0 and t interchangeably, and often write $x = (t, a, y)$. All DAGs are now defined over $N = \{0, 1, 2..., n\}$, such that

$$p_R(x) = \prod_{i=0}^{n} p(x_i \mid x_{R(i)})$$

Assume that in all DAGs, whether true or subjective, the DM's signal is the sole direct cause of his action - i.e., $R(1) = R^*(1) = \{0\}$.

The extension of the definition of personal equilibrium is straightforward. Let p have full support on T. We say that p' is a *perturbation* of p if $p'(t) \equiv p(t)$, $p'(y \mid t, a) \equiv p(y \mid t, a)$, and p' has full support on $T \times A$.

Definition 8 *A distribution $p \in \Delta(X)$ with full support on $T \times A$ is an ε-**perturbed personal equilibrium** if*

$$a \in \arg\max_{a'} \sum_{y} p_R(y \mid t, a') u(t, a', y)$$

*for every t, a for which $p(a \mid t) > \varepsilon$. A distribution $p^* \in \Delta(X)$ with full support on T is a **personal equilibrium** if there exists a sequence $p^k \to p^*$ of perturbations of p^*, as well as a sequence $\varepsilon^k \to 0$, such for every k, p^k is an ε^k-perturbed personal equilibrium.*

As in the basic model of Section 2, the object of the definition of personal equilibrium is the entire joint distribution p. An alternative approach

would fix $p(t)$ and $p(y \mid t, a)$ and regard $p(a \mid t)$ as the definition's object. However, this would give an impression of an extensive-game-like chain of causation from t to y via a, an impression that would be misleading in many applications.

The existence result given by Proposition 2 extends to this case.

An example: Illusion of control

As in Section 3.2, consider a parent who makes a decision regarding his child's education. There are three variables, denoted a, s, v, representing the parent's investment decision, the child's school performance, and the parent's valuation of success at school. The parent is informed of v before making his choice. The true DAG $R^* : a \leftarrow v \rightarrow s$. Thus, s is independent of a conditional on v. One story behind the causal link $v \rightarrow s$ is that the parent's values are imbued in the child and affect her attitude to learning. If the parent had rational expectations, he would choose a to maximize $\sum_s p(s \mid v)u(v, a, s)$.

Now suppose that the parent's subjective DAG is $R : v \rightarrow a \rightarrow s$. Thus, R departs from R^* by changing the origin of the link that goes into s. This link reorientation captures a misattribution error often referred to as "*illusion of control*" (Langer (1975)): in reality, s is caused by the exogenous variable v, yet the parent attributes s to his own action. As in Section 3.2, the parent's error is that he mishandles a confounding variable. The difference is that while in the previous example the parent neglected the confounder altogether, here he is aware of it (indeed, he conditions his action on it), yet he fails to perceive its role as a confounder. The parent interprets any correlation between a and s as a causal effect of a on s, whereas in reality the correlation is due to the confounder. In personal equilibrium, whenever $p(a' \mid v) > 0$,

$$a' \in \arg\max_a \sum_s p(s \mid a)u(v, a, s) = \sum_v p(v \mid a) \sum_s p(s \mid v)u(v, a, s) \quad (13)$$

As the term $p(v \mid a)$ indicates, the equilibrium aspect of individual behavior is fundamental in this example.

Let us add structure to the example. All variables take values in $\{0,1\}$. Assume $p(v=1)=\beta$ and $p(s=1 \mid v)=v$, where $\beta \in (0,1)$. Finally, let $u(v,a,s)=vs-\kappa a$, where $\kappa \in (0,1)$ is a constant. Thus, the child succeeds at school if and only if her family thinks it is important. The action $a=1$ is a useless investment in the child's education, and κ is its cost. If the parent had rational expectations, he would choose $a=0$ for every v, because the costly action does not affect the child's school performance. The following result characterizes personal equilibria under the parent's subjective DAG.

Proposition 12 *There are multiple personal equilibria. In one equilibrium,* $p(a=1 \mid v)=0$ *for all* v*. In another,* $p(a=1 \mid v)=v$*. Finally, if* $\kappa > 1-\beta$*, there is a third equilibrium, in which*

$$p(a=1 \mid v)=v \cdot \frac{\beta+\kappa-1}{\beta\kappa}$$

Proof. When the parent's information is $v=0$, playing $a=0$ is optimal regardless of his beliefs. Therefore, in any personal equilibrium, $p(a=1 \mid v=0)=0$. Let us try to sustain $p(a=1 \mid v=1)=0$ in equilibrium. Because the action $a=1$ is never taken in this putative equilibrium, we need to check whether there is a sequence of ε-perturbed personal equilibria that converges to p. For every $\varepsilon>0$, define $p^\varepsilon(a=1 \mid v)=\varepsilon$ for all v, $p^\varepsilon(v) \equiv p(v)$ and $p^\varepsilon(s \mid v) \equiv p(s \mid v)$. Then, $p^\varepsilon(s=1 \mid a=0) \equiv p^\varepsilon(s=1 \mid a=1)$; hence playing $a=0$ is subjective optimal, which is consistent with the definition of ε-perturbed personal equilibrium.

Now, let us try to sustain personal equilibria in which $p(a=1 \mid v=1)=\sigma>0$. Then, $p(v=1 \mid a=1)=1$ and

$$p(v=1 \mid a=0)=\frac{\beta(1-\sigma)}{1-\beta+\beta(1-\sigma)}=\frac{\beta-\beta\sigma}{1-\beta\sigma}$$

It follows that when the parent's information is $v=1$, his evaluation of $a=1$ is $1-\kappa$, whereas his evaluation of $a=0$ is $(\beta-\beta\sigma)/(1-\beta\sigma)$. Therefore, $\sigma>0$ if and only if $1-\kappa \geq (\beta-\beta\sigma)/(1-\beta\sigma)$. This inequality is binding if $\sigma \in (0,1)$. Since $\kappa<1$, the inequality holds for $\sigma=1$. If $\kappa>1-\beta$, the

inequality can hold bindingly with $\sigma = (\beta + \kappa - 1)/\beta\kappa$. ■

This result demonstrates several effects. At the substantive level, it is possible to sustain a sub-optimal equilibrium, in which parents make a useless investment in their children's education if and only if they care about school performance. Thanks to the positive correlation between v and s, this behavioral pattern is consistent: the parent mistakenly attributes success at school to the investment; and since only high-valuation parents make the investment, a and s will be positively correlated. The parents' erroneous causal interpretation of this correlation aligns it with their incentives. At the methodological level, Proposition 12 highlights the role of "trembles" in sustaining personal equilibria: the rational-expectations outcome is sustainable because the parent's off-equilibrium experimentation is uncorrelated with his information, in which case he is not led to attribute variations in school performance to variations in investment.

Consequentialist rationality

Let us extend the notion of consequentialist| rationality to the case of an informed DM.

Definition 9 *A DAG R is **consequentialistically rational** with respect to the true DAG R^* if for every pair of distributions p, q that are consistent with R^*, if $p(t) = q(t)$ and $p(y \mid t, a) = q(y \mid t, a)$ for every t, a, y, then $p_R(y \mid t, a) = q_R(y \mid t, a)$ for every t, a, y.*

Proposition 13 *The subjective DAG R is consequentialistically rational with respect to the true DAG R^* if and only if:*

1. For every $i > 1$, if $1 \notin R(i)$, then $x_i \perp_{R^*} x_1 \mid x_{R(i)\cup\{0\}}$.

2. For every $i > 1$, if $0 \notin R(i)$, then $x_i \perp_{R^*} x_0 \mid x_{R(i)\cup\{1\}}$.

3. If $R(0) \neq \varnothing$, then $x_1 \perp_{R^*} x_{R(0)} \mid x_0$.

Proof. Recall that $t = x_0$, $a = x_1$, $y = (x_2, ..., x_n)$. By assumption, $R(1) = R^*(1) = \{0\}$. Hence, by the asymmetry of R and R^*, $1 \notin R(0), R^*(0)$. We

can thus write $p_R(y \mid a, t) = p_R(x_2, ..., x_n \mid x_0, x_1)$ as

$$= \frac{\prod_{i=0}^{n} p(x_i \mid x_{R(i)})}{\sum_{x'_2,...,x'_n} p(x_0 \mid x'_{R(0)}) \cdot p(x_1 \mid x_0) \cdot \prod_{i=2}^{n} p(x'_i \mid x_{R(i)\cap\{0,1\}}, x'_{R(i)-\{0,1\}})}$$

$$= \frac{\prod_{i\neq 1} p(x_i \mid x_{R(i)})}{\sum_{x'_2,...,x'_n} p(x_0 \mid x'_{R(0)}) \cdot \prod_{i=2}^{n} p(x'_i \mid x_{R(i)\cap\{0,1\}}, x'_{R(i)-\{0,1\}})} \tag{14}$$

Recall that we are considering modifications of p that change $p(x_1 \mid x_0)$ for some x_0, x_1, while leaving the marginal of p on x_0 and the conditional distributions $p(x_2, ..., x_n \mid x_0, x_1)$ intact. Both p and its modification are required to be consistent with R^*. Let us now examine each of the terms in (14). If all terms are invariant to any eligible modification of p, so will be (14) itself; otherwise, there is an eligible modification of p that changes (14).

(i) For any $x_{R(0)}, x_0$, the term $p(x_0 \mid x_{R(0)})$ can be written as follows:

$$p(x_0 \mid x_{R(0)}) = \frac{p(x_0) \sum_{x'_1} p(x'_1 \mid x_0) p(x_{R(0)} \mid x'_1, x_0)}{\sum_{x'_0} p(x'_0) \sum_{x'_1} p(x'_1 \mid x'_0) p(x_{R(0)} \mid x'_1, x'_0)}$$

The term $p(x_0)$ is by definition invariant to the modification of p. As to $p(x_{R(0)} \mid x'_1, x_0)$, we saw that $R(0) \subseteq \{2, ..., n\}$. By definition, $p(x_2, ..., x_n \mid x_1, x_0)$ is invariant to the modification of p. Since

$$p(x_{R(0)} \mid x_1, x_0) = \sum_{x_{\{2,...,n\}-R(0)}} p(x_2, ..., x_n \mid x_1, x_0)$$

It follows that $p(x_{R(0)} \mid x'_1, x_0)$ is invariant to the modification of p. Suppose that $p(x_{R(0)} \mid x'_1, x_0)$ is not constant in x_1. Then, we can find a distribution p that is consistent with R^* and an eligible modification of p that will change $p(x_0 \mid x_{R(0)})$. In particular, let all variables x_i, $i \notin R(0)$, be independently distributed under p. The only term in (14) that will change as a result of

the modification of p is $p(x_0 \mid x_{R(0)})$, and therefore (14) will change, too. In contrast, if $p(x_{R(0)} \mid x_1', x_0)$ is constant in x_1, then $p(x_0 \mid x_{R(0)})$ is invariant to any eligible modification of p.

(ii) For any $i > 0$ and any $x_i, x_{R(i)}$, if $1 \notin R(i)$ and $0 \in R(i)$, then the term $p(x_i \mid x_{R(i)})$ can be written as follows

$$\begin{aligned} p(x_i \quad \mid \quad x_{R(i)}) &= \sum_{x_1} p(x_1 \mid x_{R(i)}) p(x_i \mid x_1, x_{R(i)}) \\ &= \sum_{x_1} \left(\frac{p(x_0) p(x_1 \mid x_0) p(x_B \mid x_1, x_0)}{\sum_{x_1'} p(x_0) p(x_1' \mid x_0) p(x_B \mid x_1', x_0)} \right) p(x_i \mid x_1, x_0, x_B) \end{aligned}$$

where $B = R(i) - \{0\}$. By definition, the terms $p(x_0)$, $p(x_B \mid x_1, x_0)$ and $p(x_i \mid x_1, x_0, x_B)$ are invariant to the modification of p. Suppose that $p(x_i \mid x_1, x_0, x_B)$ is not constant in x_1. Then, we can find a distribution p that is consistent with R^* and an eligible modification of p that will change $p(x_i \mid x_{R(i)})$. In particular, let all variables x_j, $j > 0$, $j \neq i$, be independently distributed under p. The only term in (14) that will change as a result of the modification of p is $p(x_i \mid x_{R(i)})$, and therefore (14) will change, too. In contrast, if $p(x_i \mid x_1, x_0, x_B)$ is constant in x_1, then $p(x_i \mid x_{R(i)})$ is invariant to any eligible modification of p.

(iii) For any $i > 0$ and any $x_i, x_{R(i)}$, if $0 \notin R(i)$ and $1 \in R(i)$, then the term $p(x_i \mid x_{R(i)})$ can be written as follows

$$p(x_i \mid x_{R(i)}) = \sum_{x_0} \left(\frac{p(x_0) p(x_1 \mid x_0) p(x_C \mid x_1, x_0)}{\sum_{x_0'} p(x_0') p(x_1 \mid x_0') p(x_C \mid x_1, x_0')} \right) p(x_i \mid x_1, x_0, x_C)$$

where $C = R(i) - \{1\}$. By definition, the terms $p(x_0)$, $p(x_C \mid x_1, x_0)$ and $p(x_i \mid x_1, x_0, x_C)$ are invariant to the modification of p. Using essentially the same argument as in (ii), we can show that if $p(x_i \mid x_1, x_0, x_C)$ is constant in x_0, then $p(x_i \mid x_{R(i)})$ is invariant to any eligible modification of p; and if $p(x_i \mid x_1, x_0, x_C)$ is not constant in x_0, we can find a distribution p that is consistent with R^* and an eligible modification of p that will change $p(x_i \mid x_{R(i)})$.

(iv) For any $i > 0$ and any $x_i, x_{R(i)}$, if $1, 0 \notin R(i)$, then the term $p(x_i \mid x_{R(i)})$ can be written as follows

$$p(x_i \mid x_{R(i)}) = p(x_0 \mid x_{R(i)})p(x_1 \mid x_0, x_{R(i)})p(x_i \mid x_1, x_0, x_{R(i)})$$

where

$$p(x_0 \mid x_{R(i)}) = \frac{\sum_{x_1'} p(x_0)p(x_1' \mid x_0)p(x_{R(i)} \mid x_1', x_0)}{\sum_{x_0'}\sum_{x_1'} p(x_0')p(x_1' \mid x_0')p(x_{R(i)} \mid x_1', x_0')}$$
$$p(x_1 \mid x_0, x_{R(i)}) = \frac{p(x_0)p(x_1 \mid x_0)p(x_{R(i)} \mid x_1, x_0)}{\sum_{x_1'} p(x_0)p(x_1' \mid x_0)p(x_{R(i)} \mid x_1', x_0)}$$

By definition, the terms $p(x_0)$, $p(x_{R(i)} \mid x_1, x_0)$ and $p(x_i \mid x_1, x_0, x_{R(i)})$ are invariant to the modification of p. Using essentially the same argument as in (ii), we can show that if $p(x_i \mid x_1, x_0, x_{R(i)})$ is constant in x_1, x_0, then $p(x_i \mid x_{R(i)})$ is invariant to any eligible modification of p; and if $p(x_i \mid x_1, x_0, x_{R(i)})$ is not constant in x_1, x_0, we can find a distribution p that is consistent with R^* and an eligible modification of p that will change $p(x_i \mid x_{R(i)})$. ■

The first condition in Proposition 13 is a minor variation on the condition for consequentialist| rationality in the uninformed-DM case. The second is an analogous condition for the new variable x_0. The third condition is that under the true DAG, the DM's action is independent of the signal's immediate (subjective) causes conditional on the signal. Let us illustrate this result using two specifications.

Cursedness. The true DAG R^* is $a \leftarrow t_a \leftarrow \theta \rightarrow t_b \rightarrow b$, where θ represents a state of Nature, t_a and t_b represents the signals obtained by the DM and his opponent, and a and b represent their actions. The subjective DAG R differs from R^* by changing the origin of the link that goes into b, from t_b to t_a. This specification captures a "fully cursed", partially informed DM, as in Eyster and Rabin (2005). Condition (1) in Proposition 13 holds because the node t_a blocks any path in R^* between the node a and any other node. Condition (2) holds vacuously because t_a is the sole cause of b under R. Finally, Condition (3) holds because $a \perp_{R^*} \theta \mid t_a$. Therefore, consequentialist| rationality holds.

Illusion of control. The true DAG is $R^* : a \leftarrow v \rightarrow s$, while the subjective DAG is $R : v \rightarrow a \rightarrow s$. Since $v \notin R(s) = \{a\}$, the second condition for consequentialist| rationality requires that $s \perp_{R^*} v \mid a$, which is clearly false.

References

[1] Autier, P., M. Boniol, C. Pizot and P. Mullie (2014), "Vitamin D Status and Ill health: A Systematic Review," *The Lancet Diabetes & Endocrinology* 2, 76-89.

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

[3] Bray, M. (1982), "Learning, Estimation, and the Stability of Rational Expectations," *Journal of Economic Theory,* 26, 318-339.

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

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

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

[7] Esponda, I. and D. Pouzo (2014a), "Conditional Retrospective Voting in Large Elections," mimeo.

[8] Esponda, I. and D. Pouzo (2014b), "An Equilibrium Framework for Modeling Bounded Rationality," mimeo.

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

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

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

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

[13] Fudenberg, D. and D. Levine (1993), "Self-confirming equilibrium," *Econometrica*, 61, 523-545.

[14] Geanakoplos, J., D. Pearce and E. Stacchetti (1989), "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60-79.

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

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

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

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

[19] Kearns, M., M. Littman and S. Singh (2001), "Graphical models for game theory," *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.

[20] Koller, D. and B. Milch (2003), "Multi-Agent Influence Diagrams for Representing and Solving Games," *Games and Economic Behavior* 45, 181-221.

[21] Koski, T. and J. Noble (2009), *Bayesian Networks: An Introduction,* John Wiley and Sons.

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

[23] Kőszegi, B. and M. Rabin (2006), "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics* 121, 1133-1165.

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

[25] Lucas, R. (1976), "Econometric Policy Evaluation: A Critique," Carnegie-Rochester Conference Series on Public Policy, Vol. 1, North-Holland, 1976.

[26] Madarasz, K. (2012), "Information projection: Model and Applications," *Review of Economic Studies* 79, 961-985.

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

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

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

[30] 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.

[31] Rabin, M. and D. Vayanos (2010), "The gambler's and Hot-Hand Fallacies: Theory and Applications," *Review of Economic Studies*, 77, 730-778.

[32] Samuelson, W. and M. Bazerman (1985), "The Winner's Curse in Bilateral Negotiations," in *Research in Experimental Economics*, Vol. 3, ed. by V. Smith, 105–137. JAI Press.

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

[34] Schwartzstein, J. (2014), "Selective Attention and Learning," *Journal of European Economic Association,* 12, 1423-1452.

[35] Sloman, S. (2009), Causal Models: How People Think about the World and its Alternatives, Oxford University Press.

[36] Spiegler, R. (2015), "Bayesian Networks and Missing-Data Imputation," mimeo.

[37] Verma, T. and J. Pearl (1990), "Causal Networks: Semantics and Expressiveness," *Uncertainty in Artificial Intelligence,* 4, 69-76.

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

[39] White, H. and K. Chalak (2009), "Settable Systems: an Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning," *Journal of Machine Learning Research* 10, 1759-1799.

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