# SUBJECTIVE CAUSALITY IN CHOICE[*]

ANDREW ELLIS[†] AND HEIDI CHRISTINA THYSEN[§]

ABSTRACT. An agent makes a stochastic choice from a set of lotteries. She infers the outcomes of each available lottery using a subjective causal model represented by a directed acyclic graph, and consequently may misinterpret correlation as causation. Her choices affect her inferences, which in turn affect her choices, so the two together must form a personal equilibrium. We show how an analyst can identify the agent's subjective causal model from her random choice rule. Her choices reveal the chains of causal reasoning that she undergoes, and these chains pin down her subjective causal model. In addition, we provide necessary and sufficient conditions that allow an analyst to test whether the agent's behavior is compatible with the model.

## 1. INTRODUCTION

An agent's behavior depends on her beliefs about which statistical relationships reflect causality and which reflect correlation. For instance, she may observe that duration of hospitalization is positively correlated with chance of death. This relationship may be causal if time in the hospital increases one's chance of catching an unrelated infection, or spurious if both are caused by the severity of illness. She would be more reluctant to seek treatment if she believed the former than the latter. In this paper, we develop a theory in which an analyst can use the agent's behavior, in the form of a random choice rule, to identify her subjective causal model and test whether misperceived causality explains her choices.

*Date*: October 2021.

[*] We thank Kfir Eliaz, Yusufcan Masatlioglu, Jay Lu, Pablo Schenone, Heiner Schumacher, Balazs Szentes, Chen Zhao, and especially Rani Spiegler, as well as participants at the DC Area Theory Workshop, Israeli and Hong Kong Theory Seminars, LSE, RUDC 2021, St Andrews, and UCLA for helpful comments and discussions.

[†]Department of Economics, London School of Economics and Political Science, a.ellis@lse.ac.uk.

[§]Department of Economics, Norwegian School of Economics (NHH), heidi.thysen@nhh.no.

We motivate our results with a pair of examples. Consider a doctor (the agent) prescribing medical treatments for Alzheimer's disease. These treatment can affect the patients' levels of two correlates, amyloid plaques and neurofibrillary tangles, in addition to Alzheimer's. These levels are known only after the patient undergoes a treatment. The doctor infers the effect of each treatment on Alzheimer's using data from the outcomes of her patients, and performs the one that she thinks will lead to the lowest chance of disease. A drug company (the analyst) may want to know whether a misspecified causal model can explain the doctor's reluctance to prescribe its treatment, and if so, what that model is. For instance, demonstrating that the drug decreases the chance of plaque buildup would only convince the doctor to prescribe it if she thinks that high levels of plaque *cause* Alzheimer's. It would be ineffective if the doctor believes that the correlation is spurious, as would be the case when neurofibrillary tangles cause both plaque buildup and Alzheimer's. Our results show how the analyst can answer such questions based on the frequency with which each treatment is performed in different circumstances (behavior).

For a second example, consider a firm (the agent) that hires workers from distinct pools that differ in their education, ability, and productivity. The specific attributes of a given worker are revealed to the firm only after it hires him or her. The firm uses a causal model to predict the expected productivity of workers from a given pool, and hires from each pool in proportion to predicted productivity. Since the return to education for workers depends on the firm's behavior, the Department of Education (the analyst) may want to know the firm's model. For instance, if it believes that education is purely human capital formation, i.e. that education is the only direct *cause* of productivity, then pushing for reforms that relax admission standards would not lower the return to education. If instead the firm believes that education is purely a signal of ability, then relaxed standards would lower its return. The analyst cannot directly observe the firm's causal model, and so must infer it from the frequency with which the firm hires from each pool. Our result allows the analyst to test whether the firm's hiring decisions (behavior) are consistent with a belief that education causes productivity.

This paper provides a theoretical methodology for identifying the subjective causal model of a decision maker (DM) from her behavior. Following Pearl (1995) and Spiegler (2016), we model her perception of causality using a directed acyclic graph (DAG).

Each node in the graph is a variable, such as the agent's action or outcome, and each edge represents a belief that one of the variables directly causes another. DAGs allow for a flexible and non-parametric representation of causal relationships. For example, they provide a language to distinguish between a doctor who believes that plaque causes Alzheimer's, represented by an edge between the two, and one who believes that both are caused by tangles, captured by an edge from tangles to both plaque and Alzheimer's.

We consider a DM who chooses from a set of actions, each determining a probability distribution over a vector of variables. The random choice rule representing the DM's behavior has a *subjective causality representation* if she acts as if she uses a fixed DAG to predict the consequences of her actions from data generated by her past choices, and then chooses each action with a frequency proportional to its expected utility. The choices form a personal equilibrium: how frequently she chooses each action affects her dataset, which affects her inferences about the strength of the causal relationships in her DAG, which affects her likelihood choosing each action.[1] We show how to identify the DM's DAG and preferences from her behavior. Then, we turn to the question of how to test whether a random choice rule has a subjective causality representation and provide necessary and sufficient conditions for one to exist.

The DM may mistake correlation for causation when her causal model is misspecified. In particular, she neglects the effect of her choices on the dataset from which she learns about her actions. Her choices may create a correlation between two variables that she misinterprets as a causal effect, the magnitude of which affects how likely she is to choose each action. As an example, consider a firm that thinks the pools of workers from which it hires are selected based only on education, and that education alone causes productivity, but in reality, productivity is caused only by ability. If high-ability workers in a given pool are more likely to have high education than are low-ability ones, then the perceived return to education increases with the fraction of workers hired from that pool. This incentivizes the firm to increase the frequency with which it selects that pool, reinforcing the effect. The firm's neglect of how its behavior affects the data on which it performs inferences poses technical challenges. For instance, its behavior may violate regularity, a necessary condition for a random

[1] While such a feedback effect occurs in many studies of agents with misspecified models, such as Esponda and Pouzo (2016), it has typically been absent in decision theoretic work on misspecification.

utility model (RUM). However, we show in Section 3 that it allows the model to accommodate a number of documented cognitive biases, including selection neglect, illusion of control, status-quo, and congruence biases.

Our first main result identifies the causal model that explains the DM's behavior. Within the broad class of DAGs considered, we show that the DM's perceived causal chains identify all the relevant variables and the causal relationships between them. We then reveal these chains from her behavior. Our method relies on the observation that when *every* chain passes through a set of variables, independence between those variables and the others implies indifference between all actions. For example, the doctor may believe that treatments affect the chance of plaque buildup, and that plaque buildup causes Alzheimer's. She is equally likely to choose every treatment when plaque buildup is independent of treatment, presence of tangles, and Alzheimer's. After revealing these sets, we show they can be ordered to determine the perceived causal chains.

In contrast to the large literature that empirically determines causality (e.g. Card (1999)), the result identifies an agent's *perception* of causal relationships. These perceptions may affect the agent's reaction to a policy intervention, and thereby that policy's effectiveness, even if they are not empirically valid. For instance, a firm that appears to offer minority workers a lower wage may do so because it dislikes employing minorities even though they are equally or more productive (taste-based discrimination). Alternatively, it may offer minorities a lower wage because being a minority is correlated with another attribute, such as education, that the firm thinks affects productivity (statistical discrimination, perhaps based on a wrong model and resulting in incorrect beliefs). Policies that attempt to remedy the former, such as affirmative action or awarding scholarships to minority students, may do nothing for the latter.[2]

Similarly, agents' models of the macroeconomy determine how a policy change affects their expectations about macroeconomic variables. For instance, Andre et al. (2021) show that laypeople tend to think an increase in the federal funds rate primarily affects producer costs, leading to an increase in expected inflation. However, they also show that experts tend to think that such an increase primarily affects demand, leading to a decrease in expected inflation. As inflation expectations have first-order

[2] See Lang and Kahn-Lang Spitzer (2020) for an overview of different types of race discrimination.

importance in macroeconomics, which of the two causal models prevails determines the effect of a change in the federal funds rate.

Our second main result establishes how to test whether a misspecified causal model can explain the DM's behavior. To do so, we provide necessary and sufficient conditions for a random choice rule to have a subjective causality representation. The axioms link the DM's perception of alternatives, as inferred from our first result, to her behavior. Holding her perception constant, her behavior conforms closely to Logit with an expected utility Luce index (henceforth, Logit-EU). Put another way, her choices from a pair of menus are inconsistent with Logit-EU only when the inferences she draws differ across the menus. For example, the axioms require that the DM chooses two actions with the same relative frequency whenever she makes the same prediction about their distribution of outcomes. However, her evaluation of alternatives varies across menus as how she perceives them changes. As a consequence, she may violate a number of standard axioms, including regularity.

An agent with a subjective causality representation perceives her options differently from the analyst. The result places testable restrictions on her behavior in spite of the information gap. Thus, it establishes that misspecified causality provides enough discipline on how her beliefs are distorted to be testable; without any restrictions on belief distortion, testing would be impossible. More broadly, this paper adapts decision theoretic methodology to identify and test an agent's subjective model of the world, as opposed to the usual exercise of identifying and testing her preferences with a correct, or at least an agreed upon, model of the world. We see this as a step toward providing testable implications for the growing literature studying agents with misspecified models, especially Spiegler (2016), Eliaz and Spiegler (2018), Eliaz et al. (2019), Eliaz et al. (2020), Spiegler (2020), and Schumacher and Thysen (2020), which all use versions of the subjective causality representation.[3]

We conclude by exploring how our analysis would change with different assumptions. In our analysis so far, the DM makes inferences based on her own past behavior, an important feature of the recent literature on misspecified models. However, this interpretation relies on observing the long-run steady-state distribution of choice. In

[3] Other models where misspecification leads to distorted beliefs include Esponda and Pouzo (2016), Bohren and Hauser (2018), Frick et al. (2019), He (2018), Heidhues et al. (2018), Samuelson and Mailath (2019), Montiel Olea et al. (2021), and Levy et al. (2021).

an experimental setting, it may be more convenient to instead provide subjects with an exogenously given dataset. We augment the random choice rule with such a dataset and show that our identification results naturally extend to such a setting. While random choice is widely used in experimental and empirical settings, much of the literature that adopts the DAG approach to causality focuses on deterministic choices. We adapt our identification results to this setting as well.

**Related literature.** Spiegler (2016) introduced the subjective causal representation, albeit without stochasticity and without axiomatic foundations. He shows that this can capture a number of errors in reasoning, including reversed causality and omitted variables. Taken together, our results allow us to test the underlying assumptions of existing work on the effects of causal misperception. This growing literature has been applied to monetary policy (Spiegler, 2020), political competition (Eliaz and Spiegler, 2018), communication (Eliaz et al., 2019), inference (Eliaz et al., 2020), and contracting (Schumacher and Thysen, 2020). The majority of these papers take the agents' DAGs as given, whereas our goal is to identify the DAG from behavior and test whether subjective causality explains choice. Consequently, our results increase the applicability of these papers.

Pearl (1995) argued for using and analyzing DAGs to understand causality. A large literature (e.g., Cowell et al., 1999, Koller and Friedman, 2009, Pearl, 2009) develops and applies this approach for probabilistic and causal inference. The typical exercise uses a DAG either to estimate the causal effect of a particular intervention or to infer which DAG, if any, is consistent with a given dataset.[4] Schenone (2020) introduces the DAG approach to causality into a decision theory framework. In his model, an agent expresses preferences over act-causal-intervention pairs. For instance, the DM decides which of two workers to hire; both are identical except one of them has been forced to obtain exactly 11 years of education. He provides necessary and sufficient conditions for the agent's beliefs to result from applying the "do-operator" to intervened variables for a fixed DAG and prior. The DAG is identified from the behavior. His approach is complementary to the one taken by this paper. It is mainly concerned with a normative definition of causality as a manifestation of rationality. By contrast, this paper uses DAGs to capture flaws in the reasoning of a boundedly-rational DM.

[4] Recently, Imbens (2020) contrasts this with the potential outcomes approach and discusses why these methods have attracted more attention outside of economics than within it.

More generally, our paper is related to the decision theory literature studying DMs who misperceive the world. Lipman (1999) studies a DM who may not understand all the logical implications of information provided to her. Ellis and Piccione (2017) develop a model where agents misperceive the correlation between actions. Kochov (2018) models an agent who does not accurately foresee the future consequences of her actions. Ke et al. (2020) study DMs who perceive lotteries through a neural network. Ellis and Masatlioglu (2021) consider an agent who categorizes alternatives based on the context, and the category to which it belongs affects her evaluation (or perception) of it. In all, the DM's perception of an alternative is unaffected by her behavior.

Finally, our paper also falls into the theoretical literature studying random choice. We fall between two strands. The first seeks to use choice data to identify features of otherwise rational behavior, such as Gul and Pesendorfer (2006) identifying the distribution of utility indexes, Lu (2016) identifying an agent's private information, and Apesteguia and Ballester (2018) studying comparative risk and time preferences. The second interprets randomness as a result of boundedly rational behavior in abstract environments, such as the Manzini and Mariotti (2014), Brady and Rehbeck (2016), and Cattaneo et al. (2020) models of limited attention. This paper uses random choice to identify features of explicitly boundedly-rational behavior. The only other paper of which we are aware that uses stochastic choice to capture equilibrium behavior is Chambers et al. (2021).

## 2. Model

2.1. **Setting.** Each action $a$ determines a distribution over a payoff-relevant consequence and $n$ covariates. The $i$th covariate takes a value in $\mathcal{X}_i$ and the consequence belongs to the set $\mathcal{X}_{n+1}$. For a non-empty-set $S$, let $\Delta(S)$ be the set of finite support probability distributions over $S$. Each action is a member of the set $\mathcal{X}_0 = \Delta(\prod_{i=1}^{n+1} \mathcal{X}_i)$, and it is convenient to denote $\mathcal{X}_{-0} = \prod_{i=1}^{n+1} \mathcal{X}_i$ and $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_{-0}$. We require that $\mathcal{X}_{n+1}$ is a compact subset of $\mathbb{R}$ with $|\mathcal{X}_{n+1}| \geq 2$, and take $\mathcal{X}_i = \mathbb{R}$ for simplicity.[5]

[5] We can let each $\mathcal{X}_i$, $i \leq n$, be any set with $|\mathcal{X}_i| \geq \min\{|\mathcal{X}_{n+1}|, |\mathbb{N}|\}$ and $\mathcal{X}_{n+1}$ be a compact subset of a topological space. This would increase the notational complexity but not substantively change the arguments or results.

The DM's (stochastic) choice of action determines the distribution of a random vector $X = (X_0, X_1, \dots, X_{n+1})$. If the DM chooses $a \in \mathcal{X}_0$, then $a(x_1, \dots, x_{n+1})$ is the probability that $X_i = x_i$ for every $i \in \{1, \dots, n+1\} \equiv N$. We identify the distribution over actions with the 0th random variable. The last index $n+1$ denotes the consequence. The set $\{1, \dots, n\}$ indexes the covariates. By convention, capital letters refer to variables and lowercase letters to realizations. We denote by $\operatorname{marg}_J p$ the marginal distribution of $p$ on the variables indexed by $J$. With slight abuse of notation, we sometimes identify the action $a \in \mathcal{X}_0$ with the element of $\Delta\mathcal{X}$ that has a marginal on $\mathcal{X}_{-0}$ equal to $a$ and attaches probability 1 to $X_0 = a$.

The DM decides between options in $S$, a finite subset of $\mathcal{X}_0$ where the support of the joint distribution of covariates is the product of their marginal supports for any available action. Every choice problem belongs to the set

$$\mathcal{S} = \left\{ S \subset \mathcal{X}_0 : \prod_{j=1}^{n} \operatorname{supp}(\operatorname{marg}_j a) = \operatorname{supp}(\operatorname{marg}_{\{1,\cdots,n\}} b) \text{ for all } a, b \in S \text{ and } |S| \in (1, \infty) \right\}.$$

This ensures that Bayes' rule is well-defined and can be relaxed in specific examples.

A random choice rule $\rho : \mathcal{X}_0 \times \mathcal{S} \to [0,1]$ where $\sum_{a \in S} \rho(a, S) = 1$ and $\rho(a, S) = 0$ for every $a \notin S$ describes the DM's choices. The probability she chooses $a$ from $S$ is $\rho(a, S)$. Identify $\rho^S$ with the probability distribution over $\mathcal{X}$ induced by the DM's choice probabilities, that is

$$\rho^S \in \Delta\mathcal{X} \text{ where } \rho^S(a, y) = \rho(a, S)a(y) \text{ for all } a \in \mathcal{X}_0 \text{ and } y \in \mathcal{X}_{-0}.$$

Note $\rho(\cdot, S)$ is a distribution over actions whereas $\rho^S$ is a distribution over $\mathcal{X}$.

For $p \in \Delta(B)$, the qualifier "for $p$-a.e. $z \in B$" means "for almost every $z \in B$ according to $p$," or equivalently "for every $z$ in the support of $p$" since $p$ has finite support. For a set $J \subset N$ and $x \in \mathcal{X}_{-0}$, $x_J$ denotes the event that $X_j = x_j$ for all $j \in J$. When it will not cause confusion, we sometimes write $x_j$ instead of $x_{\{j\}}$ and $x_\emptyset$ for an arbitrary constant. For $k \in \mathbb{R}$, $k_j$ denotes the event that $X_j = k$. We define the mixture between lotteries $a$ and $b$, $\alpha a + (1-\alpha)b$, in the usual way.

2.2. **Subjective Causality Representation.** A directed acyclic graph (DAG) over a set $M$ is an acyclic binary relation $R \subset M \times M$, where $iRj$ denotes $(i, j) \in R$. A

DAG $R$ over $\{0, 1, \ldots, n+1\}$ describes the DM's perception of causality. Here, $iRj$ indicates that the DM thinks that $X_i$ directly causes $X_j$ and corresponds to a directed edge in a graph with nodes $\{0, \ldots, n+1\}$. We often write $R(i)$ for the indexes of the variables that cause $X_i$ according to $R$, termed the *parents* of $i$. For a DAG $R$ and $p \in \Delta\mathcal{X}$, $p_R$ is the probability distribution so that

$$p_R(x) = \prod_{j=0}^{n+1} p\left(x_j | x_{R(j)}\right)$$

for every $x \in \mathcal{X}$. See Spiegler (2016) for further interpretation.

A DAG $R$ is *uninformed* if there is no $i \in N$ with $iR0$. That is, none of the other variables cause the DM to choose her action. A DAG $R$ is *nontrivial* if $0Ri_1Ri_2R\ldots Ri_m$ for some $i_1, i_2, \ldots i_m \in N$ with $i_m = n+1$. That is, there is a channel through which the choice of action can influence the distribution over consequences. Say that $(i, j, k)$ is an $R$-v-collider if $iRk$, $jRk$, $j \not R i$, and $i \not R j$. The DAG $R$ is *perfect* if there are no $R$-v-colliders. We focus on perfect DAGs because otherwise the perceived marginal distribution of some variable may diverge from its true distribution (Spiegler, 2017).

**Definition 1.** The random choice rule $\rho$ has a *subjective causality representation (SCR)* if there exists an uninformed, nontrivial DAG $R$ and a continuous, strictly increasing $u$ so that

$$\rho(a, S) = \frac{\exp\left(\int_{\mathcal{X}_{n+1}} u(c) d\rho_R^S(c_{n+1}|a)\right)}{\sum_{a' \in S} \exp\left(\int_{\mathcal{X}_{n+1}} u(c) d\rho_R^S(c_{n+1}|a')\right)}$$

for every $a \in S$ and $S \in \mathcal{S}$; then, we say $\rho$ has an SCR $(R, u)$ and that $R$ represents $\rho$. An SCR is perfect if its DAG is perfect.

The representation corresponds to the following "as if" procedure. The DM maximizes expected utility but with a potentially incorrect prediction of the consequence distribution resulting from her actions. She believes that taking an action only directly affects the variables caused by it, which in turn affects the variables caused by them, and so on and so forth. Each of these causal effects are calculated from the "dataset" $\rho^S$ that contains every realization of the random vector $X$ with a frequency determined by her choices. She begins by predicting the distribution of a variable caused only by her action. Then, she forms an overall prediction by recursively extending her (interim) predicted distribution to subsequent variables using her estimates of the causal effects. Ultimately, she expects to receive consequence $c$ with probability $\rho_R^S(c_{n+1}|a)$ if she takes action $a$.

For comparison, a $\rho$ has a *Logit-EU* representation if there is a continuous, increasing $u$ so that for every $a \in S$ and $S \in \mathcal{S}$,

$$\rho(a, S) = \frac{\exp\left(\int_{\mathcal{X}_{n+1}} u(c)da(c_{n+1})\right)}{\sum_{a' \in S} \exp\left(\int_{\mathcal{X}_{n+1}} u(c)da'(c_{n+1})\right)} = \frac{\exp\left(\int_{\mathcal{X}_{n+1}} u(c)d\rho^S(c_{n+1}|a)\right)}{\sum_{a' \in S} \exp\left(\int_{\mathcal{X}_{n+1}} u(c)d\rho^S(c_{n+1}|a')\right)}.$$

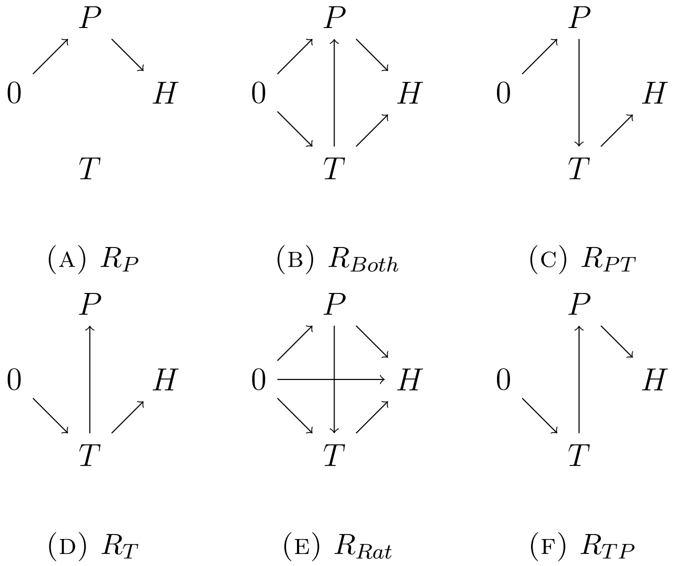
The SCR replaces the Bayesian update $\rho^S(\cdot|a)$ with the one generated by her causal model $\rho_R^S(\cdot|a)$.

An SCR is a personal equilibrium (Köszegi and Rabin, 2006): the DM maximizes expected utility given her beliefs that depend on her choices. It is easy to show that an equilibrium exists for any $S \in \mathcal{S}$ using Brouwer's fixed point theorem. For menus with more than one, we place no restrictions on which is selected.

*Remark* 1. We adopt the exponential function for concreteness and applicability. Our results adapt to any other strictly increasing and positive function. See Footnote 12.

2.3. **Running Example.** Throughout, we illustrate our framework with the following running example. The random choice rule represents frequencies with which a doctor performs medical treatments for Alzheimer's disease on observationally identical patients.[6] An analyst, e.g. a drug company, observes the doctor's choices, and wants to know whether the decisions can be explained by an SCR, and if so, what the doctor's causal model is. Each treatment leads to a realization of three variables: two correlates, amyloid plaque buildup ("plaque" for short, indexed by $P = 1$) and the presence of neurofibrillary tangles ("tangles" for short, indexed by $T = 2$), and long-term health status, i.e., whether the patient gets Alzheimer's (indexed by $H = 3$). Each of the variables is binary and takes values in $\{0, 1\}$. The vector $(0, 1, 1)$ represents a patient who does not have plaque buildup, has tangles, and is in good health (i.e., does not have Alzheimer's). The probability that a patient undergoing treatment $a$ has those characteristics is $a(0, 1, 1)$. The doctor only cares about the long-term health of her

[6] While we focus on the individual interpretation of random choice for expositional purposes, our results apply equally well to a group interpretation provided that the group has a sufficiently similar causal model. Here, the DM is a "representative" patient choosing their own treatment on the basis of their inferences from other patients' results. This would apply to vaccination, for instance.



(A) $R_P$ (B) $R_{Both}$ (C) $R_{PT}$

(D) $R_T$ (E) $R_{Rat}$ (F) $R_{TP}$

Figure 1. Possible DAGs in the Running Example

patients. When choosing the treatment, the outcome of the correlates as well as health is unknown. However, she observes the dataset of joint realizations of treatments, correlates, and health.

Figure 1 gives some possible DAGs. Each DAG represents a different theory of causation.[7] A doctor represented by $R_P$ believes that the treatment directly influences the plaque buildup, and that plaque, and plaque alone, causes Alzheimer's. By contrast, one represented by $R_T$ believes that the treatment only directly influences tangles, and that tangles cause both plaque and Alzheimer's. A doctor represented by $R_{Rat}$ always correctly predicts the outcome distribution of each action because it is a complete graph and therefore does not embed any conditional independence assumptions. However, a DM represented by any of the other DAGs in Figure 1 potentially makes incorrect predictions. For instance, if she is represented by $R_P$, then she must believe that Alzheimer's is unaffected by the treatment conditional on level of plaque, regardless of whether this holds empirically.

2.4. **Interpretations of the model.** Our main interpretation of an SCR is that it describes a DM who predicts the outcome of her action using a causal model. It may

[7] DAGs are common tools in applied health research; see Tennant et al. (2020) for a survey.

also describe a DM with limited data access (Spiegler, 2017). In this interpretation, she only considers or observes the distributions of several overlapping subsets of variables. She then extrapolates to form a distribution over all variables using the principle of insufficient reason. Formally, she uses the distribution that maximizes entropy subject to matching the marginal distributions over the subsets she considers. Identifying her DAG corresponds to identifying the considered subsets. In the running example, a doctor represented by $R_P$ only observes, or only has access to, two datasets: one that keeps track of the efficacy of the treatments on plaque and another one that tracks the impact of plaque on Alzheimer's. Recently (and controversially), the FDA approved the Alzheimer's medication aducanumab on the basis of its reduction in plaque buildup despite limited evidence of its effects on the disease itself.[8]

Alternatively, it may describe a DM learning the distribution of the random vector $X$ from the dataset $\rho^S$ with the aid of a DAG $R$. The model factorizes the distribution $\rho^S$ into a perceived distribution $\rho^S_R$, and because $R$ embeds conditional independence assumptions, it may reduce the number of moments needed to reconstruct the distribution. In the running example, the DM can store all the relevant information according to $R_P$ using only 6 parameters; for $S = \{a, b\}$, the numbers $p(a)$, $p(1_P \mid a)$, $p(1_T)$, $p(1_H \mid 1_P)$ and $p(1_H \mid 0_P)$ fully determine $p_R$. In contrast, it would require $2^4 - 1 = 15$ parameters to record the probability of each possible realization without this assumption. Such an interpretation is agnostic as to why the DM infers the overall distribution rather than just the relationship between action and consequence. She may (incorrectly) anticipate the arrival of additional information or the possibility of taking other actions. She may think it is easier or quicker to learn the stronger correlations between the covariates that make up a causal chain than to learn the weaker correlation between her action and the outcome.

For a final interpretation, we note that when $\rho$ has an SCR, $\rho^S_R$ minimizes Kullback–Liebler divergence from $\rho^S$ among all the probability distributions on $\mathcal{X}$ that are consistent with $R$. Then, $\rho$ represents a single agent Berk–Nash equilibrium (Esponda and Pouzo, 2016) with extreme-value errors. As in that model, we can interpret the behavior as the

[8] See www.theatlantic.com/health/archive/2021/07/americas-drug-approval-system-unsustainable/619422/ for a description of the controversy and fda.report/media/143503/PCNS-20201106-CombinedFDABiogenBackgrounder_0. for the evidence submitted and the FDA's evaluation thereof.

steady state of a learning process with a set of parameters (probability distributions) that do not include the "true" one.

## 3. Behavioral Implications of Misspecification

The DM's behavior may endogenously create correlations that she misinterprets as causation. Fundamentally, the DM neglects the effect of her choices on her data.[9] This leads to two key technical challenges. First, an SCR random choice rule may violate *regularity*, a property necessary for representation by a random utility model (RUM). Second, the agent's behavior may be *self-confirming*: because she chooses $a$ frequently, she thinks it is better than $b$, but would reverse her thinking if she chose $b$ more frequently. However, these features allow the model to accommodate a number of biases documented in the psychology literature, including violations of independence of irrelevant alternatives, illusion of control, status-quo bias, and congruence bias. In this section, we illustrate these challenges and relate them to the behavior permitted.

3.1. **Regularity violation.** A choice rule with an SCR may violate regularity, the requirement that $\rho(a, S) \geq \rho(a, S')$ whenever $a \in S \subseteq S'$. Consequently, the class of choice rules with an SCR and those with a RUM do not coincide. This poses a challenge for identifying the model as usual techniques do not directly apply. By contrast, many behaviors interpreted as irrational can be represented by a RUM, such as the model of limited attention due to Manzini and Mariotti (2014).

To illustrate why regularity may be violated, consider a doctor in the running example whose behavior has a SCR $(R_P, u)$. There are three treatment options, $\iota, \pi, \nu$, that are equally likely to lead to good health. Plaque and health are independent after undergoing treatment $\iota$, positively correlated under treatment $\pi$, and negatively correlated under treatment $\nu$. When the doctor decides between only $\iota$ and $\pi$, plaque buildup is necessarily positively correlated with good health. As she mistakes the correlation for causation, this makes treatments that are more likely to lead to plaque buildup more attractive. However, when she chooses between all three treatments, the patients to whom she prescribes $\nu$ may cancel out or even reverse the perceived positive

[9] Esponda and Vespa (2018) experimentally document selection neglect, and Denrell (2018) provides a recent survey of evidence for it in managers.

effect of plaque on health. When this effect is strong enough, it can lead to an increase in the probability of choosing the treatment with a lower probability of plaque buildup.

Formally, suppose that $\iota(1_P, 1_H) = \iota(1_P, 0_H) = \frac{1}{2}$, $\pi(0_P, 0_H) = \pi(1_P, 1_H) = \frac{1}{2}$, $\nu(1_P, 0_H) = \nu(0_P, 1_H) = \frac{1}{2}$, $u(L) = 0$, and $u(H) = 6$.[10] Consider menus $S = \{\iota, \pi\}$ and $S' = \{\iota, \nu, \pi\}$.[11] If $\rho(\iota, S) = z$, then

$$\rho^S(1_H|0_P) = 0 < \rho^S(1_H|1_P) = \frac{z\frac{1}{2} + (1-z)\frac{1}{2}}{z(1) + (1-z)\frac{1}{2}} = \frac{1}{1+z},$$

and since $\iota(1_P) > \pi(1_P) > 0$, we have $1 > z > \frac{1}{2}$. Then, $\frac{1}{2} < \rho^S(1_H|1_P) < \frac{2}{3}$, so $\rho^S_R(1_H|\pi) < \frac{1}{3}$, while $\rho^S_R(1_H|\iota) > \frac{1}{2}$. Hence

$$\frac{\rho(\pi, S)}{\rho(\iota, S)} < \frac{\exp[\frac{1}{3}6 + \frac{2}{3}0]}{\exp[\frac{1}{2}6 + \frac{1}{2}0]} = \exp[-1] < \frac{1}{2}$$

and $\rho(\pi, S) < \frac{1}{3}$. Because health is independent of the treatment conditional on plaque according to $R_P$, the doctor is indifferent between two treatments with the same probability of plaque buildup. Therefore, $\rho(\nu, S') = \rho(\pi, S') = \gamma$. Then,

$$\rho^{S'}(1_H|1_P) = \frac{(1-2\gamma)\frac{1}{2} + \gamma\frac{1}{2} + \gamma(0)}{(1-2\gamma) + 2\gamma\frac{1}{2}} = \frac{1}{2} = \frac{(1-2\gamma)(0) + \gamma(0) + \gamma\frac{1}{2}}{(1-2\gamma)(0) + 2\gamma\frac{1}{2}} = \rho^{S'}(1_H|0_P),$$

so $\rho(\iota, S') = \rho(\nu, S') = \rho(\pi, S') = \frac{1}{3} > \rho(\pi, S)$, violating regularity.

While the violations of regularity allow the model to accommodate phenomena like the decoy effect, the failure stems from faulty reasoning. The above doctor overestimates her ability to control events, or exhibits illusion of control (Langer, 1975). Her choices do not affect long-term health, yet she would be willing to pay a premium to choose one treatment over another. Moreover, she also exhibits "patternicity" (Shermer, 1998) in that she perceives a pattern, namely that using $\iota$ leads to better long term-health outcomes, where none exist.

3.2. **Self-confirming choices.** As illustrated above, the outcome that the DM expects to get from an action may depend on how frequently she chooses it. She may

[10] The distribution of tangles does not affect behavior, so we leave it unspecified.

[11] We note that $S, S' \notin \mathcal{S}$, so this example, and that in the next subsection, are technically outside our domain. At the cost of complicating the algebra and obscuring the logic, they can be made consistent with our assumptions by replacing each action $c$ with $c' = (1-\epsilon)c + \epsilon d$ where $\epsilon > 0$ is small enough and $d(y) = \frac{1}{8}$ for each $y \in \{0,1\}^3$.

perceive one action to be better than another only if she chooses it sufficiently frequently; this can lead to multiple personal equilibria. For instance, consider a doctor who thinks that plaque alone causes health. Action $b$ represents prescribing a drug that prevents the disease but often leads to plaque buildup. Action $a$ represents not intervening, and such patients often get the disease but rarely have plaque buildup and then only when the patient gets the disease. When the doctor usually does not prescribe the drug, low plaque buildup is mistakenly believed to increase the chance of good health. Consequently, prescribing the drug, which raises their plaque, seems like a bad idea. Symmetrically, when she usually prescribes the drug, not intervening to decrease plaque buildup seems counterproductive.

Formally, the DM has an SCR $(R_P, u)$ and chooses between two treatments, $S = \{a, b\}$, such that $b(1,1,1) = 1$, $a(0,0,1) = \frac{1}{2}q$, $a(0,0,0) = \frac{1}{2}(1-q)$, and $a(1,1,0) = \frac{1}{2}$ where $q \in (0,1)$. One can compute that

$$\rho^S\left(1_H|1_P\right) = \frac{2-2\rho(a,S)}{2-\rho(a,S)} \text{ and } \rho^S\left(1_H|0_P\right) = q.$$

That is, whether the doctor thinks that plaque has a positive or negative effect on Alzheimer's depends on the fraction who take action $a$. Setting $u(1) = \lambda$ and $u(0) = 0$ for $\lambda > 0$, we must have

$$\lambda\frac{1}{2}\left(q - \frac{2-2\rho(a,S)}{2-\rho(a,S)}\right) = \ln\rho(a,S) - \ln(1-\rho(a,S)).$$

There are multiple solutions for large enough $\lambda$. For instance when $\lambda = 30$ and $q = \frac{3}{4}$, the solutions are $\rho(a,S) \approx .34$, $\rho(a,S) \approx 0.02$, and $\rho(a,S) \approx 0.99$.

With data that contains a small enough fraction of patients who received the drug, the doctor often does not intervene because her misspecified model concludes that plaque buildups lead to good health. While this is true in that dataset, taking the drug is unambiguously superior, as she would realize if a larger fraction took it. Interpreting the choices as the steady state of a learning process, such a DM exhibits status quo bias (Samuelson and Zeckhauser, 1988), a tendency toward "maintaining one's current or previous decision," and congruence bias (Wason, 1960) by failing to test the alternative hypothesis that the drug is better than not intervening.

## 4. Revealing the Subjective Causal Model

In this section, we reveal the DM's subjective causal model from her choice behavior. In a DAG $R$, information flows along the chains of causal relations from actions to consequences. We show that we can reveal these causal chains from the DM's behavior, and that they pin down the subjective causal model.

Formally, a causal chain is an *active path from* $0$ *to* $n+1$ *in* $R$ (an $R$*-AP* for short): a finite sequence of variables $(i_0, i_1, i_2, \ldots, i_m)$ with $i_0 = 0$, $i_m = n+1$, and $i_j R i_{j+1}$ for every $j < m$. This represents a chain of causal reasoning: according to $R$, variable $0$ causes $i_1$, which in turn causes $i_2$, and so on, ending with a cause of the outcome variable. A *minimal* $R$*-AP, or* $R$*-MAP,* is an $R$-AP that cannot be made shorter. That is, $(i_0, \ldots, i_m)$ is an $R$-MAP if it is a $R$-AP and $i_j \not R i_{j'}$ whenever $j' \neq j+1$. The main result of this section shows that the $R$-MAPs suffice to identify the DAG.

**Theorem 1.** *Let* $\rho$ *have a perfect SCR* $(R, u)$ *and* $R'$ *be a perfect, uninformed, nontrivial DAG. Then,* $\rho$ *has an SCR* $(R', u')$ *if and only if the set of* $R'$*-MAPs coincides with the set of* $R$*-MAPs and there exists* $\beta$ *so that* $u(c) = u'(c) + \beta$ *for every* $c \in \mathcal{X}_{n+1}$.

The result shows that two DMs behave identically if and only if they agree on the channels through which their actions affect their payoffs and they have the same tastes. In particular, the $R$-MAPs capture the parts of the causal model relevant for predicting the expected consequence of each action. This has two immediate corollaries. First, only relationships between variables that appear in at least one $R$-MAP affect the DM's behavior. In particular, any variables caused by the consequence are inconsequential for the her behavior and can therefore be ignored. Second, the chains of causality involving such variables determine all other causal relationships. While there may be causal links not in an $R$-MAP, their existence and direction can be determined from the causal chains or are immaterial to the DM's choices.

To see the consequences of the first implication, note that the presence of tangles does not belong to any $R_P$-MAPs. A doctor represented by $R_P$ can also be represented by a DAG $R'$ that contains the links in $R_P$ and adds links to and from tangles that do not create a cycle, an $R'$-MAP, or a v-collider. For the second, a DM represented by $R_{both}$ can also be represented by a DAG that reverses the link between plaque and tangles, but no other DAG. Any representation must have the same $R_{both}$-MAPs ($(0, P, H)$ and $(0, T, H)$) and must have some link between plaque and tangles to rule out a v-collider.

The next two subsections illustrate how the analyst can reveal the $R$-MAPs from choice behavior, establishing the necessity of the condition in Theorem 1. We begin by identifying the sets of variables through which every causal chain must pass. Then, we provide an ordering on these sets that reveals the $R$-MAPs.

### 4.1. **Revealing relevant variables.**

This subsection identifies the sets of covariates that intersect every minimal causal chain from the DM's choices. These sets reveal the relevant variables for her causal model, in that any relationship involving one of the others can be dropped from her model without affecting her behavior.

Given disjoint $I, J \subseteq N$, we say that $X_I$ *is independent of* $X_J$ *within* $S \in \mathcal{S}$, written $X_I \perp_S X_J$, if $\operatorname{marg}_I a = \operatorname{marg}_I b$ for any $a, b \in S$ and $a(x_I, x_J) = a(x_I)a(x_J)$ for disjoint every $x \in \mathcal{X}_{-0}$ and every $a \in S$. If $X_I$ is independent of $X_J$ within $S$, then regardless of how the DM chooses from $S$, the random vector $X_I$ is independent of $X_J$ in the resulting joint distribution $\rho^S$.

**Definition 2.** The set $I \subseteq N$ *separates* if $\rho(a, \{a, b\}) = \frac{1}{2}$ whenever $X_I \perp_{\{a,b\}} X_{N \setminus I}$.

A set $I$ of covariates separates if the DM is indifferent between her available actions whenever $X_I$ is independent of all other variables within $\{a, b\}$. That is, she believes that there is no relationship between the outcome and action whenever there is no relationship between the variables indexed by $I$ and the others. An experimenter can create this independence by intervening to set their values without changing the others in a randomized controlled trial. Then, $I$ separates if the subject does not think she can affect the outcome after the experimenter intervenes to set the value of $X_I$. Alternatively, we explore allowing for an exogenous dataset in Section 6 where independence can be induced using unavailable actions.

**Lemma 1.** *If $\rho$ has an SCR $(R, u)$, then $I \subseteq N$ separates if and only if every $R$-MAP intersects $I$.*

To illustrate, consider a doctor who is equally likely to prescribe every treatment whenever plaque buildup is independent of the other variables, i.e. $\{P\}$ separates. In particular, she is indifferent between such treatments even when one provides a lower chance of tangles and better health. Then, she must believe that the correlation between treatments, tangles, and health is spurious. Hence, her choices reveal that she believes that every causal chain includes plaque.

In general, the set $I$ separates when every causal chain in the DM's model passes through $I$. Intuitively, making a variable independent shuts down any causal chain containing it. If $I$ intersects every chain, then the independence of $X_I$ within $S$ shuts down all the channels through which the DM thinks her choice can affect the outcome, causing her to be indifferent. Observe that $\{n+1\}$ separates, as does any set containing $n+1$. In any menu where the consequence is independent of all other variables, all causal channels to it are shut down, so the DM is equally likely to choose both actions. Whenever $I$ does not intersect some causal chain, one can construct a menu within which $X_I$ is independent and yet the DM is not indifferent between all available options. The construction of the menu is independent of $R$ and $u$, so given that $\rho$ has an SCR, it suffices to observe choice from a single menu to assess whether $I$ separates.

The set of all *minimal separators*,

$$\mathcal{A} = \{I \subset N : I \text{ is a minimal set that separates}\},$$

has a tight connection to the DM's subjective causal model. Consider an $R$-MAP $(i_0, ..., i_m)$. For any $A, B \in \mathcal{A}$, $(i_0, ..., i_m)$ must intersect both $A$ and $B$. If $A \cap B = \emptyset$, it contains exactly one member of each of $A$ and $B$. If $A \cap B \neq \emptyset$, then either it contains $k \in A \cap B$ and no other member of $A \cup B$, or exactly one member from each of $A \setminus B$ and $B \setminus A$. Moreover, the contrapositive of Lemma 1 says that if $N \setminus I$ does not separate, then some $R$-MAP involves only variables in $I$. That is, if $I$ intersects every $A \in \mathcal{A}$, then $I$ contains every covariate in some $R$-MAP.

Determining $\mathcal{A}$ immediately rules out many DAGs. In the running example, the minimal separators suffice to distinguish between populations with any of the DAGs in Figure 1 except $R_{PT}$ and $R_{TP}$. If $\rho$ has an SCR $(R, u)$, then $\mathcal{A}$ equals $\{\{P\}, \{H\}\}$ when $R = R_P$, $\{\{T\}, \{H\}\}$ when $R = R_T$, $\{\{P, T\}, \{H\}\}$ when $R = R_{Both}$, and $\{\{H\}\}$ when $R = R_{Rat}$. By contrast, $\mathcal{A} = \{\{P\}, \{T\}, \{H\}\}$ reveals that $R$ has exactly one

$R$-MAP that contains both plaque and tangles. That is, the analyst can infer that either $R = R_{PT}$ or $R = R_{TP}$ but not which.

4.2. **Revealing causal chains.** To distinguish between causal chains containing more than one covariate, we define an ordering on $\mathcal{A}$ from $\rho$ that reveals the perceived direction of causality.

**Lemma 2.** *If $\rho$ has a perfect SCR $(R, u)$, then $\mathcal{A}$ can be ordered $\mathcal{A} = \{A^*_1, \ldots, A^*_{|\mathcal{A}|}\}$ so that $(i_0, \ldots, i_m)$ is a R-MAP if and only if for every $j \in \{0, \ldots, m-1\}$, there exists $k$ so that $i_j \in A^*_k \setminus A^*_{k+1}$ and $i_{j+1} \in A^*_{k+1} \setminus A^*_k$ when $A^*_0 = \{0\}$.*

If all perceived causes of the variables indexed by $B \in \mathcal{A}$ are contained in $A$, then the DM believes any relationship between her action and the consequence is spurious whenever the variables indexed by $A$ are independent of those indexed by $B$. Intuitively, independence shuts down all the causal chains through which she thinks her action can affect the consequence. We order $\mathcal{A}$ using this observation and that the consequence must come last.

To illustrate, consider a doctor represented by $R_{PT}$. On the one hand, if tangles are independent of health, then since she believes that tangles are the only cause of Alzheimer's, the doctor thinks that each treatment is equally effective. She is equally likely to prescribe either, regardless of the true relationship between her action and health. On the other hand, the doctor may think one drug is superior when plaque is independent of Alzheimer's. For instance, when tangles occur if and only if health is bad and there is no plaque buildup, then tangles are negatively correlated with both the other variables. Then, the doctor thinks the drug that leads to a lower chance of plaque buildup is best, even if plaque is independent of health. Hence, her choices reveal that the presence of tangles causes Alzheimer's and that plaque causes tangles, so $A^*_3 = \{H\}$, $A^*_2 = \{T\}$, and $A^*_1 = \{P\}$.

We construct the general ordering recursively, starting with the last.

**Definition 3.** Let $A^*_{|\mathcal{A}|} = \{n+1\}$. Recursively define $A^*_i \in \mathcal{A} \setminus \left\{A^*_{i+1}, \ldots, A^*_{|\mathcal{A}|}\right\}$ so that $A^*_i \cap A^*_{i+1} \supseteq A \cap A^*_{i+1}$ for all $A \in \mathcal{A} \setminus \left\{A^*_{i+1}, \ldots, A^*_{|\mathcal{A}|}\right\}$ and $\rho(a, \{a, b\}) = \frac{1}{2}$ for any

$\{a,b\} \in \mathcal{S}$ so that

$$X_{A_i^* \cap A_{i+1}^*} \perp_{\{a,b\}} X_{N \setminus [A_i^* \cap A_{i+1}^*]} \;\&\; X_{A_{i+1}^* \setminus A_i^*} \perp_{\{a,b\}} X_{A_i^* \setminus A_{i+1}^*}.$$

Consider first a DM whose subjective causal model consists of a single minimal causal chain, $(j_0 = 0, j_1, \ldots, j_m = n+1)$. Then, $\mathcal{A} = \{\{j_1\}, \{j_2\}, \ldots, \{j_m\}\}$, so $A \cap B = \emptyset$ and $A \setminus B = A$ for any distinct $A, B \in \mathcal{A}$. The independence condition suffices to determine the ordering. Since $j_i$ is the only cause of $j_{i+1}$, independence of $X_{j_i}$ from $X_{j_{i+1}}$ within $\{a,b\}$ shuts down the causal chain. This leads the DM to dismiss any actual relationship between her action and the consequence as spurious. However, the DM may perceive a correlation between $X_{j_k}$ and $X_{j_{i+1}}$ for any $k < i$ even when they are actually independent, as illustrated above. Consequently, independence between $X_{j_k}$ and $X_{j_{i+1}}$ suffices for indifference if and only if $k = i$. Therefore, we must have $A_i^* = \{j_i\}$ when $A_k^* = \{j_k\}$ for every $k \geq i+1$.

Consider now a DM whose model consists of more than one chain. Then, for $A, B \in \mathcal{A}$, $A$ may not be a singleton and $A \cap B$ may not be empty. The analyst can find $A_i^*$ by applying the same logic as that in the single-chain case with three modifications. First, $A_i^*$ must have the largest intersection with $A_{i+1}^*$. If $A \in \mathcal{A} \setminus \left\{A_{i+1}^*, \ldots, A_m^*\right\}$ has $A \cap A_{i+1}^* \not\subseteq A_i^* \cap A_{i+1}^*$, then there is a cycle: for any $j \in \left(A_{i+1}^* \cap A\right) \setminus A_i^*$ and $k \in A_i^* \setminus A_{i+1}^*$, there must be an $R$-MAP $(\ldots, j, \ldots, k, j \ldots)$. Therefore, any $A, B \in \mathcal{A} \setminus \left\{A_{i+1}^*, \ldots, A_m^*\right\}$ that have the largest intersection with $A_{i+1}^*$ must satisfy $A_{i+1}^* \cap A = A_{i+1}^* \cap B = A^{**}$. Second, the actions available must shut down the chains that intersect $A^{**}$. To shut them down, we consider menus in which the variables indexed by $A^{**}$ are independent of all others. Third, if $A$ precedes $A_{i+1}^*$ then every $j \in A \setminus A^{**}$ must come before every $k \in A_{i+1}^* \setminus A^{**}$ in some $R$-MAP since each $A \in \mathcal{A}$ contains one variable from each causal chain. Therefore, the DM must be indifferent between $a$ and $b$ only if $X_{A^{**}} \perp_{\{a,b\}} X_{N \setminus A^{**}}$ and *every* variable indexed by $A_i^* \setminus A^{**}$ is independent of *every* variable indexed by $A_{i+1}^* \setminus A^{**}$ within $\{a,b\}$.

**Proposition 1.** *If $\rho$ has a perfect SCR, then $\mathcal{A}$ and its ordering as per Definition 3 can be determined by observing $\rho$ in a finite number of binary menus.*

For $I \subset N$, we can construct a binary menu $\{a,b\}$ so that the DM is indifferent between $a$ and $b$ if and only if $I$ separates. Then, $X_I \perp_{\{a,b\}} X_{N \setminus I}$ and $a$ $(b)$ almost always leads to high (low) realizations, so all the variables indexed by $N \setminus I$ are almost perfectly correlated. The DM realizes the relationship between her action and outcome if and only if $I$ does not separate. The ordering of the minimal separators is revealed recursively as above. To reveal whether $A \in \mathcal{A} \setminus \{A^*_{i+1}, \cdots, A^*_{|\mathcal{A}|}\}$ equals $A^*_i$, we construct a menu $\{a, b\}$ so that the DM strictly prefers $a$ to $b$ if and only if $A \neq A^*_i$. Combining the two observations establishes the result.

### 4.3. Revealed causes.

A perfect SCR may be represented by more than one DAG, provided that both have the same minimal causal chains (Theorem 1). We turn to the question of which links must belong to any representation of the DM's subjective causal model.

**Definition 4.** For $j, k \in \{0, 1, \ldots, n+1\}$, *$j$ is revealed to cause $k$ for $\rho$*, written $j\hat{R}^\rho k$, if $jRk$ for every SCR $(R, u)$ that represents $\rho$.

That is, $j$ is revealed to cause $k$ if $j$ causes $k$ in every representation of $\rho$. This definition is analogous to how Masatlioglu et al. (2012) define revealed preference with limited attention. When $i$ directly precedes $j$ in an $R$-MAP, then $i\hat{R}^\rho j$. However, there may be other revealed causes. Proposition 2 characterizes all of them.

**Proposition 2.** *If $\rho$ has a perfect SCR, then $j\hat{R}^\rho k$ if and only if there exists $i$ so that $j \in A^*_i$ and $k \in A^*_{i+1} \setminus A^*_i$ when $A^*_0 = \{0\}$.*

The result extends Lemma 2. The ordering of separators reveals not only the $R$-MAPs but also every causal relationship that belongs to every representation of $\rho$. Moreover, $\hat{R}^\rho$ includes but is not limited to the links in $R$-MAPs. The causal links revealed by the ordering are the only ones whose direction is uniquely determined by choice behavior. Other links, such as that between $P$ and $T$ when $\rho$ is represented by $R_{Both}$, need to be included in the DAG, but their direction is not pinned down. Determining the direction of causality between such variables requires additional data, such as choices following interventions as studied by Schenone (2020).

*Remark* 2. The DAG $\hat{R}^\rho$ may not be perfect. One can construct a perfect *revealed DAG* $R^\rho$ by setting $jR^\rho k$ if and only if either $j\hat{R}^\rho k$ or there exists $i$ so that $j, k \in A^*_{i+1} \setminus A^*_i$

and $j < k$ with $A_0^* = \{0\}$. When $\rho$ has a perfect SCR, $R^\rho$ represents $\rho$. Moreover,

$$\rho_R^S\left(x_{\cup_{i=1}^{|\mathcal{A}|} A_i^*}|a\right) = a\left(x_{A_1^*}\right)\prod_{i=1}^{|\mathcal{A}|-1}\rho^S\left(x_{A_{i+1}^*\setminus A_i^*}|x_{A_i^*}\right)$$

for any perfect DAG $R$ that represents $\rho$, any $a \in S$, and $a$-a.e. $x \in \mathcal{X}_{-0}$.

## 5. Behavioral Foundations for Subjective Causality

This section characterizes the random choice rules that have a subjective causality representation. Throughout, the properties are illustrated in the context of the running example. For these purposes, we consider a doctor for which $\mathcal{A} = \{\{P\}, \{H\}\}$, and so, using results in Section 4, can be represented by $R_P$. That is, the DM believes that the treatments can influence plaque build-up, and plaque is the sole determinant of health.

The first axiom is standard, requiring simply that the DM chooses every available action with positive probability.

**Axiom 1** (Full-support)**.** For any $S \in \mathcal{S}$ and $a \in S$, $\rho(a, S) > 0$.

The next axiom limits the perceived difference between any two options.

**Axiom 2** (Bounded Misperception)**.** The Luce ratio is bounded: $\sup_{S,a,b\in S}\frac{\rho(a,S)}{\rho(b,S)} < \infty$.

The relative frequency with which the DM takes two actions, their *Luce ratio*, indicates the strength of her preference. Since the set of consequences is compact, there is a best and worst outcome. These provide a natural limit to how much she prefers one action to another, which bounds the Luce ratio. The axiom thus bounds the size of the mistakes that the DM can make. In the running example, the doctor's Luce ratio for any two treatments is bounded above by that of the one she believes will always lead to a good health outcome and that of the one she believes will always leads to a bad health.

The third axiom ensures that the minimal separators can be ordered as in Section 4.

**Axiom 3** (Consistent Revealed Causes)**.** The set $\{n+1\} \in \mathcal{A}$, and for $i = 1, \dots, |\mathcal{A}|$, $A_i^*$ exists.

This axiom implies that the DM does not think that the realization of a variable can simultaneously influences and be influenced by the realization of another variable. The DM may still believe that the current inflation level influence the current unemployment level, and that both influence future inflation levels. However, current and future inflation levels are regarded as two distinct variables. Thus, it is possible to give a consistent account of which variables influence each other variable. As per Proposition 1, it can be falsified by observing the DM's choices from a finite number of menus.

The fourth axiom requires that if the DM predicts two actions lead to the same outcome distribution, then she chooses each with the same probability.

**Axiom 4** (Indifferent If Identical Immediate Implications, I5)**.** For $a, b \in S \in \mathcal{S}$, if $\operatorname{marg}_{A_1^*} a = \operatorname{marg}_{A_1^*} b$, then $\rho(a, S) = \rho(b, S)$.

The covariates directly caused by the DM's action are a sufficient statistic for her prediction of the outcome distribution. That is, if two actions have the same distribution over these covariates, then she believes that both lead to identical probabilities of every consequence. She is therefore indifferent between any two actions with identical immediate implications according to her subjective causal model. In the running example, it implies that the doctor is equally likely to choose any treatments with the same probability of plaque buildup, regardless of any other differences between them. Note that this axiom only considers the marginal distributions on $X_{A_1^*}$ of $a$ and $b$, remaining agnostic about their distribution on other variables and the available actions.

The following axiom relates similarities in the DM's inferences to her choices.

**Axiom 5** (Luce's Choice Axiom Given Inferences, LCI)**.**
For any $S, S_1, S_2, \dots \in \mathcal{S}$ with $a, b \in S_m \cap S$ for each $m$, if

$$\rho^{S_m}\left(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*}\right) \to \rho^{S}\left(y_{A_{i+1}^* \setminus A_i^*} | y_{A_i^*}\right)$$

for $\rho^S$-a.e. $y \in \mathcal{X}_{-0}$ and every $i = 1, \dots, |\mathcal{A}|$, then $\frac{\rho(a,S_m)}{\rho(b,S_m)} \to \frac{\rho(a,S)}{\rho(b,S)}$.

The Logit model is characterized by Luce's Choice Axiom (Luce, 1959), which requires that $\frac{\rho(a,S')}{\rho(b,S')} = \frac{\rho(a,S)}{\rho(b,S)}$ whenever $a, b \in S \cap S'$. LCI requires that the choice axiom holds when the DM's inferences about variables conditional on their revealed causes are close. That is, it implies that if $S, S' \in \mathcal{S}$, $a, b \in S \cap S'$, and $\rho^S(y_{A_{i+1}^*\setminus A_i^*}|y_{A_i^*}) = \rho^{S'}(y_{A_{i+1}^*\setminus A_i^*}|y_{A_i^*})$ for each $i$, then $\frac{\rho(a,S')}{\rho(b,S')} = \frac{\rho(a,S)}{\rho(b,S)}$. Moreover, the choice axiom is "close" to holding whenever these conditional probabilities are "close." In the running example, suppose that given the doctor's choices when facing $\{a, b\}$ and $\{a, b, c\}$, the statistical relationship between plaque and health is (almost) the same. Then, the predicted long-term health prospect from treatment $a$ or $b$ is (almost) the same across menus. Thus, the relative frequency with which she chooses each of these treatments should be (almost) the same.

The next definition identifies a set of menus for which the DM's predictions about the consequence of each action is correct.

**Definition 5.** A menu $S \in \mathcal{S}$ is *(revealed to be) correctly perceived* if

$$b\left(y_{\bigcup_{i=1}^{|\mathcal{A}|} A_i^*}\right) = b\left(y_{A_1^*}\right) \prod_{i=1}^{|\mathcal{A}|-1} a\left(y_{A_{i+1}^*\setminus A_i^*}|y_{A_i^*}\right),$$

for every $a, b \in S$ and $a$-a.e. $y \in \mathcal{X}_{-0}$.

Recall from Remark 2 that only the set of variables indexed by $\bigcup_{i=1}^{|\mathcal{A}|+1} A_i^*$ is relevant for the DM's choices and that she thinks that the variables indexed by $A_i^*$ cause those indexed by $A_{i+1}^* \setminus A_i^*$. In particular, she thinks that any variable in $A_{i+1}^* \setminus A_i^*$ is independent of any in $\bigcup_{k=1}^{i-1} A_k^*$ conditional on $A_i^*$. A correctly perceived menu satisfies this conditional independence, and her prediction of the distribution of all relevant variables, including the consequence, is correct.

The remaining axioms ensure that $\rho$ has a Logit-EU representation for menus where the DM's predictions are correct.

**Axiom 6** (Correctly Perceived Independence)**.** For any $\alpha \in (0, 1)$, if $\{\alpha p + (1 - \alpha)r, r\}, \{\beta p + (1 - \beta)r, r\} \in \mathcal{S}$ are correctly perceived, then

$$\beta \ln \frac{\rho(\alpha p + (1 - \alpha)r), \{\alpha p + (1 - \alpha)r, r\})}{\rho(r, \{\alpha p + (1 - \alpha)r, r\})} = \alpha \ln \frac{\rho(\beta p + (1 - \beta)r, \{\beta p + (1 - \beta)r, r\})}{\rho(r, \{\beta p + (1 - \beta)r, r\})}.$$

**Axiom 7** (Correctly Perceived Dominance)**.** If $\{p, q\} \in \mathcal{S}$ is correctly perceived and $\operatorname{marg}_{n+1} p$ strictly first-order stochastically dominates $\operatorname{marg}_{n+1} q$, then $\rho(p, \{p, q\}) > \frac{1}{2}$.

**Axiom 8** (Correctly Perceived Continuity)**.** If $S, \{p_1, q_1\}, \{p_2, q_2\}, \dots \in \mathcal{S}$ are correctly perceived, $p, q \in S$, $\operatorname{marg}_{n+1} p_m \to \operatorname{marg}_{n+1} p$, and $\operatorname{marg}_{n+1} q_m \to \operatorname{marg}_{n+1} q$, then

$$\frac{\rho(q_m, \{p_m, q_m\})}{\rho(p_m, \{p_m, q_m\})} \to \frac{\rho(q, S)}{\rho(p, S)}.$$

Combined, the axioms imply that the DM's behavior is suitably well-behaved when she correctly predicts the outcomes of each action. Axiom 6 guarantees that the independence axiom holds, and that the relative probability of choosing $\alpha p + (1 - \alpha) r$ to $r$ is log linear in $\alpha$ because of the Logit functional form.[12] Axioms 7 and 8 guarantee monotonicity and continuity over the distribution of consequences.

The main result of this section characterizes the rules with a perfect SCR.

**Theorem 2.** *A random choice rule $\rho$ has a perfect subjective causality representation if and only if $\rho$ satisfies Full-support, Bounded Misperception, I5, Consistent Revealed Causes, LCI, and Correctly Perceived Independence, Dominance, and Continuity.*

The result highlights the connection between SCR and the Logit-EU model. Notice that if Axioms 1 and 5-8 hold when the part of their hypotheses involving the minimal separators are dropped, the choice rule has a Logit-EU representation. Bounded Misperception gives a maximum deviation in the relative choice frequencies. The axioms thus indicate the circumstances under which the choice rule does not diverge from Logit. I5 says that two alternatives are chosen with same probability whenever they coincide on the distribution of variables that the action is revealed to cause, whereas Logit-EU requires coincidence on the consequence distribution. LCI restricts violations of Luce's Choice Axiom to when inferences about causal effects change. Correctly Perceived Independence, Dominance, and Continuity show that choice is well-behaved whenever the DM's predictions about actions match reality.

We outline the proof for sufficiency here, and defer a formal proof to the appendix. We first show that the choice rule has a Logit-EU representation when restricted to

[12] This is the only axiom that needs be adapted to replace exp with a different strictly increasing, positive function. For instance, dropping "ln" from Axiom 6 replaces it with the identity function.

correctly perceived menus. By Axiom 3, the DAG $R^\rho$ from Remark 2 is well-defined and a natural candidate to represent $\rho$. For every menu $S$, we construct a correctly perceived version of it, $S'_1$. That is, for every $a \in S$, there is an $a' \in S'_1$ so that $a'(\cdot) = \rho^S_{R^\rho}(\cdot|a)$ and $S'_1$ is correctly perceived. Our goal is to show that for any $a, b \in S$, $a$ and $b$ are chosen with the same relative frequency in $S$ as $a'$ and $b'$ are in $S'_1$. To do so, we add distinct alternatives to $S'_1$ to form a nested sequence of menus $(S'_m)_{m=1}^\infty$ while maintaining that each $S'_m$ is correctly perceived. Bounded Misperception implies that the probability of choosing anything in $S$ from $S'_m \cup S$ goes to zero as the number of alternatives in $S'_m$ goes to infinity. In particular, the inferences that the DM makes from $S'_m \cup S$ approach those she makes from $S'_1$, which are in turn equal to those she makes from $S$. LCI implies that the relative frequency with which $a'$ and $b'$ are chosen from $S'_m \cup S$ converges to that for $a'$ and $b'$ in $S'_1$. Moreover, $a$ and $a'$ (as well as $b$ and $b'$) are chosen from $S'_m \cup S$ with the same probability by I5. Applying LCI another time, we see that $a$ and $b$ are chosen with the same relative frequency in $S$ as $a'$ and $b'$ are in $S'_1$, completing the proof.

We conclude by further clarifying the relationship between SCR and Logit-EU.

**Corollary 1.** *A random choice rule $\rho$ has a Logit-EU representation if and only if $\mathcal{A} = \{\{n+1\}\}$ and $\rho$ satisfies Full-support and Correctly Perceived Independence, Dominance, and Continuity.*

If $\rho$ has an SCR $(R, u)$ and $\mathcal{A} = \{\{n+1\}\}$, then $0R(n+1)$. Moreover, every menu is correctly perceived. Consequently, the DM acts as if Logit-EU on all menus.

## 6. Discussion and Extensions

This section concludes the paper by looking at some implications of the model and considering how our modeling decisions affect our results. We compare the behavior of two DMs with nested causal models Then, we examine how to extend our analysis to eliminate stochasticity and the endogeneity of the dataset.

### 6.1. **Comparative Coarseness.**

A coarser causal model leaves out some variables or relationships relative to another. Authors often explain "irrational" behavior in situations with adverse selection via coarseness. For instance, Eyster and Rabin (2005), Jehiel and Koessler (2008), and Esponda (2008) argue that the winner's curse reflects bidders who do not fully take into account the relationship between others' actions and signals.[13] In this subsection, we compare DMs in terms of the coarseness of their model. In particular, how can an analyst separate two DMs who differ in that one's model contains more variables than the other's?

**Definition 6.** Say that $\rho_2$ *has a coarser model than* $\rho_1$ if $\rho_1(\cdot, S) = \rho_2(\cdot, S)$ whenever $X_i \perp_S X_{N\setminus\{i\}}$ for all $i \in N$ that are not in any minimal $\rho_2$-separator.

Consider DM1 represented by $\rho_1$ and DM2 represented by $\rho_2$. As revealed by Theorem 1, DM2 considers the variables that do not belong to a $\rho_2$-MAP irrelevant for determining the consequence of their action. The condition says that whenever those variables are actually irrelevant when choosing from $S$, i.e., they are independent of the other variables, then the two DMs behave in the same way. This ensures that if DM2 thinks a variable is relevant, so does DM1.

**Proposition 3.** *Let $\rho_i$ have a perfect SCR $(R_i, u_i)$ for $i = 1, 2$. If $\rho_2$ has a coarser model than $\rho_1$, then $\rho_2$ has a perfect SCR $(R_1 \cap N' \times N', u_2)$ for some $N' \subset \{0, 1, \ldots, n+1\}$ and $u_2 = u_1 + \beta$. The converse holds up to the selection of a personal equilibrium.*

The result shows that the comparison reveals when the models of two DMs are nested. Specifically, they agree on the causal relationship between any two variables that both consider relevant and on the desirability of outcomes. However, they may disagree on which variables are relevant, with DM1 considering more of them relevant than DM2.

6.2. **Welfare.** There is no general relationship between the DM's DAG and an analyst's evaluation of her choices. For instance, Section 4.3 of Spiegler (2016) shows that a DM may achieve higher objective payoff with a coarser model. Similar examples apply to this setting as well. With a perfect DAG, the DM and the analyst agree on the overall distribution of consequences. This implies that a refinement such as preferred personal equilibrium can be defined from either the agent's or analyst's perspective without changing conclusions.

[13] Section 5 of Spiegler (2016) discusses how and to what extent these models fit into the DAG framework.

6.3. **Exogenous dataset.** We can extend our results to a setting where the dataset used by the DM is exogenously given and does not depend on her behavior. This setting provides rich variation in the DM's inferences. It is particularly applicable to an experimental implementation of our result. Most of the insights from our analysis with an endogenous dataset are readily applicable. Indeed, it guarantees uniqueness of the personal equilibrium and ensures that the DM conforms to Logit holding the dataset fixed.

Formally, we consider behavior in an environment $(S, q)$ where the DM's choice set $S \in \mathcal{S}$ and the DM's dataset $q \in \Delta\mathcal{X}$ has $q(a) > 0$ for each $a \in S$ and

$$\prod_{j=0}^{n} \operatorname{supp}(\operatorname{marg}_j q) = \operatorname{supp}(\operatorname{marg}_{\{0,\cdots,n\}} q).$$

Let $\mathcal{E}$ be the set of such pairs. The DM's behavior is given by the augmented random choice rule $\rho^* : \mathcal{X}_0 \times \mathcal{E} \to [0,1]$ with $\sum_{a \in S} \rho^*(a; S, q) = 1$ and $\rho^*(a; S, q) > 0$ only if $a \in S$. The frequency that they choose $a$ in the environment $(S, q)$ is $\rho^*(a; S, q)$.

**Definition 7.** The augmented random choice rule $\rho^*$ has an *Exogenous SCR (ESCR)* if there exists an uninformed, nontrivial DAG $R$ and a continuous, nonconstant $u$ so that

$$\rho^*(a; S, q) = \frac{\exp\left(\int_{\mathcal{X}_{n+1}} u(c) dq_R(c_{n+1}|a)\right)}{\sum_{a' \in S} \exp\left(\int_{\mathcal{X}_{n+1}} u(c) dq_R(c_{n+1}|a')\right)}$$

for every $a \in S$ and $S \in \mathcal{S}$.

It is easy to adapt our identification results to this setting. Theorem 1 holds as stated. To establish that the behavior on $R$-MAPs is uniquely pinned down, we apply the conditions to the dataset directly rather than to the options in the menu. For instance, Lemma 1 says that a subset of covariates contains a $R$-AP if and only if its complement separates. The result continues to hold after we modify Definition 2 to say that $K \subset N$ separates if $\rho^*(a; \{a, b\}, q) = \frac{1}{2}$ whenever $X_K$ is independent of the other variables according to $q$: $q(x) = q(x_K)q(x_{N \setminus K})$ for $q$-a.e. $x \in \mathcal{X}$. That is, independence is required for the dataset, not the menu. If the dataset is easily manipulable, as in an experiment, then the condition may be substantially easier to test.

6.4. **Inference from multiple choice sets.** We have implicitly assumed that the modeler sees the steady-state distribution of choices from a fixed menu (or alternatively an exogenous dataset). Real-world agents face different choice sets on different occasions and learn from their decisions in all of them. Our framework generalizes to accommodate this, provided that the distribution of choice sets is fixed across time. Consider a distribution $\mu \in \Delta\mathcal{S}$ with $\mu(S)$ indicating the frequency with which the agent faces menu $S$. Then, we take as a primitive $\hat{\rho} : \mathcal{X}_0 \times \Delta(\mathcal{S}) \times \mathcal{S} \to [0,1]$ so that $\rho(\cdot; \mu, \cdot)$ is a random choice rule for fixed $\mu$. The DM chooses $a$ with probability $\rho(a, \mu, S)$ when they face menu $S$. The above analysis is the special case $\rho(\cdot, S) = \hat{\rho}(\cdot; (1, S), S)$. To take into account how learning spills over across menus, we replace the dataset $\rho^S$ with $\hat{\rho}^\mu$ where

$$\hat{\rho}^\mu(a, y) = \sum_{\mu-a.e.\ S} \mu(S)\hat{\rho}(a; \mu, S)a(y)$$

for each $(a, y) \in \mathcal{X}$. With such a substitution, our results generalize naturally. As an additional feature, the analyst can reveal the agent's causal model from its behavior when facing a single, suitably chosen distribution over menus.

6.5. **Deterministic choice.** The SCR is derived from Spiegler (2016), where choice is deterministic. We have adopted a stochastic choice framework throughout the paper. The stochastic setting is closer to that typically used in empirical and experimental work. It also deals with some technical issues. For instance, it pins down beliefs about the consequence distribution of every alternative. Moreover, it applies when only one of potentially many personal equilibria is observed. Our insights apply to a deterministic choice model, once suitably adapted. We discuss how to apply them in this subsection.

Formally, we suppose that the DM's behavior is described by a choice correspondence $c : \mathcal{S} \rightrightarrows \Delta(\mathcal{X}_0)$ where $p(S) = 1$ for all $p \in c(S)$ and $c(S) \neq \emptyset$ for each $S \in \mathcal{S}$.[14] For any $p \in \Delta(\mathcal{X}_0)$, write $p^X \in \Delta(\mathcal{X})$ for the resulting dataset, i.e. $p^X$ is the lottery so that $p^X(a, y) = p(a)a(y)$ for every $(a, y) \in \mathcal{X}$.

**Definition 8** (Spiegler (2016))**.** For $\epsilon > 0$, the lottery $p \in \Delta(B)$ is a $(R, u, \epsilon)$-*personal equilibrium* for $B \in \mathcal{S}$ if $p(a) > 0$ for all $a \in B$ and

$$p(a) \geq \epsilon \implies a \in \arg\max_{a' \in B} \int_{\mathcal{X}_{n+1}} u(c) dp_R^X(c_{n+1}|a').$$

[14] As shown in Spiegler (2016), there may not exist a personal equilibrium that does not mix.

The lottery $p \in \Delta(B)$ is a $(R, u)$-*personal equilibrium* for $B \in \mathcal{S}$ if there exists a sequence $(p_t)_{t=1}^{\infty}$ so that $p_t$ is a $(R, u, 1/t)$-personal equilibrium for $B$ and $p_t \to p$.

The choice correspondence $c$ has a *Deterministic SCR (DSCR)* if there exists a uninformed, nontrivial DAG $R$ and a nonconstant, continuous $u : \mathcal{X}_{n+1} \to \mathbb{R}$ so that for every $B \in \mathcal{S}$, $p \in c(B)$ if and only if $p$ is a $(R, u)$-personal equilibrium for $B$. A DSCR $(R, u)$ is perfect if $R$ is perfect. Observe that limiting cases of SCR are personal equilibria. Formally, let $\rho^{\lambda}$ be a random choice rule having a perfect SCR $(R, \lambda u)$ for $\lambda > 0$. If $\rho^{\lambda_n}(a, S) \to p(a)$ for every $a \in S$ and $\lambda_n \to \infty$, then $p$ is an $(R, u)$-personal equilibrium for $S$.

A version of Theorem 1 holds in this setting. For instance, Lemmas 1 and 2 require only minor alterations. Specifically, one must replace "$\rho(\{a, b\})(a) = \frac{1}{2}$" with "$c(\{a, b\}) = \Delta\{a, b\}$" in Definitions 2 and 3.

6.6. **Directions for future research.** Our results apply only to perfect subjective DAGs and settings where the agent has no access to information. Relaxing those two features is a natural direction for future research. An imperfect DAG allows for v-colliders, which provide a natural way to model an agent who theorizes the exogeneity of a variable. One direction of Theorem 1 fails for imperfect DAGs. Although any other DAG representing $\rho$ must agree on the minimal active paths from the action to the consequence, this is not sufficient for it to represent $\rho$ when $\rho$ has an imperfect SCR. They must also agree, for instance, on the v-colliders with a path to the consequence.

Informed agents provide additional challenges, not only about how a primitive should be defined, but also how to change the model itself. In the setting we studied in this paper, our uninformed, nontrivial assumption implies that $p_R(\cdot|a)$ equals the update after applying the "do-operator" of Pearl (1995). If the DM thinks signals cause actions but that she can also change her action, this no longer holds. Since the "do-operator" severs links pointing to the intervened variable, it can effectively create v-colliders. This suggests spillovers between the study of informed DMs and those with an imperfect causal model.

We also restrict attention to situations where the DM takes a single action and cares about the realization of a single variable. While without loss in a rational model, modeling multiple outcomes as a many individual variables or as a single vector affects behavior in this model. For instance, when each outcome is modeled as a separate variable, we can consider a DM who thinks the outcomes are independent given the action. By contrast, if both outcomes are modeled a single, vector-valued variable the DM considers them independent conditional on a common parent only if they actually are.

DAGs provide a tractable, non-parametric, and flexible approach to modeling misspecified causality. Many other forms of misspecification have interesting economic applications (see Footnote 3). While unrestricted misspecified beliefs are untestable, our results show that exploiting the regularities implied by a particular model of misspecification may render it testable. The techniques and setup introduced herein may prove useful for identifying and testing some of these models.

## Appendix A. Proofs from the Main Text

A.1. **Preliminaries for Proof of Theorem 1.** For a DAG $Q$, let $N^*(Q)$ be the minimal set of nodes such that $\text{marg}_{n+1}\, p_Q(\cdot \mid a) = \text{marg}_{n+1}\, p_{Q\cap N^*(Q)^2}(\cdot \mid a)$. Throughout, we say that a DAG $R$ is *well-behaved* if it is perfect, uninformed, nontrivial, and $R \subseteq N^*(R)^2$. A DAG $Q'$ is equivalent to $Q$ if and only if $p_Q = p_{Q'}$ for all $p \in \Delta(\mathcal{X})$. Let the skeleton of $Q$ be $\tilde{Q} = \{(j,k) : jQk \text{ or } kQj\}$.

**Proposition 4** (Theorem 1 of Verma and Pearl (1991))**.** *Two DAGs are equivalent if and only if they have the same skeleton and the same set of v-colliders.*

**Definition 9.** If $jGk$ for every uninformed DAG $G$ equivalent to $Q$, then the link $jGk$ is called a *fundamental link* in $Q$ and denoted by $j\hat{Q}k$.

**Proposition 5** (Proposition 6 of Schumacher and Thysen (2020))**.** *Given a perfect DAG $Q$, $N^*(Q) = \{j \in N \cup \{0\} \mid j$ is part of a $\hat{Q}$-AP$\}$.*

We will sometimes refer to the variables indexed by $N^*(Q)$ as the relevant variables for $Q$. The fundamental links are characterized in the next proposition. Before that we need another definition.

**Definition 10.** The distance between any two nodes $j, k$, denoted by $d(j,k)$, is given by the number of links in the shortest path between $j$ and $k$.

**Proposition 6** (Proposition 7 of Schumacher and Thysen (2020))**.** *Let $Q$ be a perfect, uninformed, nontrivial DAG. If $jQk$, then $j\hat{Q}k$ if and only if at least one of the following conditions is satisfied:*

(a) $d(0,j) = d(0,k) - 1$,

(b) *there exists a node $l \in N$ such that $l\hat{Q}j$ and $l \not{Q} k$.*

**Definition 11.** Let $\mathcal{C}$ be a collection of subsets of a finite set and $\mathcal{T}$ a tree with $\mathcal{C}$ as its node set. Say that $\mathcal{T}$ is a *junction tree* if for any $C_1, C_2 \in \mathcal{C}$, $C_1 \cap C_2$ is contained in every node on the unique path in $\mathcal{T}$ between $C_1$ and $C_2$.

The set $C \subseteq \{0, \ldots, n+1\}$ is a clique for $R$ if $j\tilde{R}k$ for all $j, k \in C$. By Theorem 4.6 of Cowell et al. (1999), the maximal cliques of a well-behaved DAG $R$ can be linked to form a junction tree. Call this the *maximal clique junction tree (MCJT)* for $R$.

A.2. **Proof of Theorem 1.** We present the proof as a sequence of lemmas. Lemma 1 shows that a minimal separator intersects every $R$-MAP. Lemma 3 shows that every relevant variable is in some $R$-MAP (and vice versa) and thus belongs to some $A \in \mathcal{A}$. Consequently, there is no loss in taking $R$ to relate only variables that appear in $\bigcup_{A \in \mathcal{A}} A$. Lemmas 4-7 relate the fundamental links in $R$ to $\mathcal{A}$, and Lemma 8 asserts that the ordering on $\mathcal{A}$ according to Definition 3 is well-defined and unique. Lemmas 9 and 10 characterize $R$ using this ordering. The remainder of the proof ties the Lemmas together to establish the result.

*Proof of Lemma 1.* Because $R$ is acyclic, we can relabel so that $R(i) \subset \{0, 1, \ldots, i\}$ for all $i$, where $n+1$ remains last since we can clearly drop any links from $n+1$ without changing behavior.

We first show that if $J$ intersects every $R$-MAP, then $p_R(x_J|c) = p_R(x_J)$ for $p$-a.e. $c$ whenever $X_J$ is $p$-independent of $X_{N \setminus J}$. Given our relabeling, we can take $J = \{j_1, j_2, \ldots, j_K\}$ with $j_i < j_{i+1}$. For any disjoint $E, E' \subset \{1, \ldots, j_1 - 1\}$ and $p$-a.e. $x \in \mathcal{X}_{-0}$, we have

$$\begin{aligned} p_R(x_{j_1}, x_E | c, x'_E) &= \sum_y \frac{p_R(x_E, x_{E'}, y_{R(j_1) \setminus [E \cup E']} | c)}{p_R(x_{E'}|c)} p(x_{j_1} | y_{R(j_1) \setminus [E \cup E']}, x_{R(j_1) \cap [E \cup E']}) \\ &= \sum_y \frac{p_R(x_E, x_{E'}, y_{R(j_1) \setminus [E \cup E']} | c)}{p_R(x_{E'}|c)} p(x_{j_1}) = p_R(x_E | c, x_{E'}) p_R(x_{j_1}) \end{aligned}$$

when $0 \notin R(j_1)$ by independence of $X_J$ from $X_{N \backslash J}$. The same argument holds when $0 \in R(j_1)$ after replacing $p(x_{j_1}|y_{R(j_1)\backslash[E\cup E']}, x_{R(j_1)\cap[E\cup E']})$ with $p(x_{j_1}|y_{R(j_1)\backslash[E\cup E'\cup\{0\}]}, x_{R(j_1)\cap[E\cup E']}, c)$.

If (IH) $p_R(x_{\{j_1,\dots,j_m\}}, x_E|c, x'_E) = p_R(x_E|c, x_{E'})p_R(x_{j_1,\dots,j_m})$ any disjoint $E, E' \subset \{1, \dots, j_1 - 1\}$, then for disjoint $A, A' \subset \{1, \dots, j_1 - 1\}$ and $A^* = A \cup A' \cup J$

$$\begin{aligned}
&p_R(x_{\{j_1,\dots,j_{m+1}\}}, x_A|c, x_{A'}) \\
=&\sum_y \frac{p_R(x_A, x_{A'}, x_{\{j_1,\dots,j_m\}}, y_{R(j_{m+1})\backslash A^*}|c)}{p_R(x_{A'}|c)} p(x_{j_{m+1}}|y_{R(j_{m+1})\backslash A^*}, x_{R(j_m+1)\cap A^*}) \\
=&\sum_y p_R(x_A, y_{R(j_{m+1})\backslash A^*}|c, x_{A'})p_R(x_{\{j_1,\dots,j_m\}})p(x_{j_{m+1}}|y_{R(j_{m+1})\backslash A^*}, x_{R(j_m+1)\cap A^*}) \\
=&\sum_y p_R(x_A, y_{R(j_{m+1})\backslash A^*}|c, x_{A'})p_R(x_{\{j_1,\dots,j_m\}})p(x_{j_{m+1}}|x_{R(j_{m+1})\cap J}) \\
=&p_R(x_A|c, x_{A'})p_R(x_{\{j_1,\dots,j_{m+1}\}})
\end{aligned}$$

when $0 \notin R(j_{m+1})$. The second equality uses IH for $E = A \cup [R(j_{m+1}) \backslash A^*]$ and $A' = E'$, and the third uses independence of $X_J$ from $X_{N \backslash J}$. The same arguments hold when $0 \in R(j_{m+1})$ after replacing $p(x_{j_{m+1}}|y_{R(j_{m+1})\backslash A^*}, x_{R(j_m+1)\cap A^*})$ with $p(x_{j_{m+1}}|y_{R(j_{m+1})\backslash[A^*\cup\{0\}]}, x_{R(j_m+1)\cap A^*}, c)$. The claim follows inductively, using $E = E' = \emptyset$ and that $p_R(x_{n+1}|c) = \sum_y p_R(x_{n+1}|y_J)p_R(y_J|c)$ when every $R$-AP intersects $J$.

We show the converse by contrapositive. Let $J \subset N$. Suppose that $(i_0 = 0, i_1, \dots, i_m = n + 1)$ is a $R$-MAP that does not intersect $J$, and let $I = \{i_1, \dots, i_m\}$. If $m = 1$, then $0 \; R \; (n + 1)$ and for any $\{a, b\} \in \mathcal{S}$, $c(x_{n+1}) = [\alpha a + (1 - \alpha)b]_R(x_{n+1}|c)$ for every $x \in \mathcal{X}_{n+1}$, $\alpha \in (0, 1)$, and $c \in \{a, b\}$ clearly implying that $J$ does not separate. Otherwise, consider $\{a, b\} \in \mathcal{S}$ so that $X_j \perp_{\{a,b\}} X_{N\backslash\{j\}}$ for every $j \notin I$; $\operatorname{supp} \operatorname{marg}_{i'} \frac{1}{2}a + \frac{1}{2}b = \{\bar{y}, \underline{y}\} \subset \mathcal{X}_{n+1}$ with $\bar{y} > \underline{y}$ for all $i' \in I$; for every $j \in \{1, \dots, |I|\}$, each $c \in \{a, b\}$, and $a$-a.e. $x, x' \in \mathcal{X}_{-0}$,

$$c\left(x_{i_{j+1}}|x_{i_j}, x'_{\{i_1,\dots,i_{j-1})\}}\right) = a\left(x_{i_{j+1}}|x_{i_j}\right)$$

and $a(\bar{y}_{i_{j+1}}|\bar{y}_{i_j}) > a(\bar{y}_{i_{j+1}}|\underline{y}_{i_j})$; and $a(\bar{y}_{i_1}) > b(\bar{y}_{i_1})$. For any $\alpha \in (0, 1)$, let $q \in \Delta(\mathcal{X})$ equal $\alpha a + (1 - \alpha)b$. Note that $q = q_R$, so $q_R(x_{n+1}|c) = c(x_{n+1})$ for $c = a, b$. This establishes that $\rho(a, \{a, b\}) \neq \frac{1}{2}$ so does not $J$ separate, completing the proof. □

**Lemma 3.** *Let $R$ be a perfect, uninformed, nontrivial DAG. Then, there is an $R$-MAP containing $i \in \{0, 1, \dots, n + 1\}$ if and only if $i \in N^*(R)$.*

*Proof.* First we show that if $k\hat{R}l$ and $k \neq 0$, then there exists $j \in N$ s.t. $j\hat{R}k$ and $j\not R l$. To see why, consider two nodes indexed by $k$ and $l$ so that $k\hat{R}l$ and $k \neq 0$. Assume for contradiction that if $j\hat{R}k$, then $jRl$. As this rules out condition Proposition 6.b, it must be that $d(0,k) = d(0,l) - 1$. Since $d(0,k) > 0$, there exists a node $j$ so that $jRk$ and $d(0,j) = d(0,k) - 1$. By Proposition 6.a $j\hat{R}k$, and by assumption $jRl$. But then $d(0,l) \leq d(0,j) + 1 = d(0,k)$, a contradiction.

Now we show that for any $R$-AP $(i_0, \ldots, i_m)$, if $i_j \not R i_k$ for some $k > j$, then $i_j \not R i_l$ for all $l > k$. If $i_j R i_{k+1}$ for $k > j$, then $i_j R i_k$, since we must have $i_j \tilde{R} i_k$ so that $(i_j, i_k, i_{k+1})$ is not a $R$-v-collider, and moreover $i_j R i_k$ because $i_k R i_j$ would imply that $R$ has a cycle. Inductively applying the contrapositive establishes the claim. Note that if $i \in \{0, n+1\}$, then $i$ is part of every $R$-MAP by construction. So let $i \in N^*(R) \setminus \{0, n+1\}$. By Proposition 5, there is a $\hat{R}$-AP containing $i$. By dropping nodes from this $\hat{R}$-AP if necessary, we can reduce the part of the path following $i$ to $(i_0 = i, i_1, \ldots, i_M = n+1)$ where $i_k R i_{k'}$ if and only if $k' = k+1$. We reconstruct the path preceding $i$ as follows. Let $j_2 = i$ and $j_1 = i_1$. Set $k = 2$. (*) When $j_k \neq 0$, pick a node $j_{k+1}$ so that $j_{k+1}\hat{R}j_k$ and $j_{k+1} \not\hat{R} j_{k-1}$. By the above claims, $j_{k+1}$ exists, $j_{k+1} \not R j_l$ for $l \leq k$, and $j_{k+1} \not R i_l$ for $l = 1, 2, \ldots, M$. If $j_{k+1} = 0$, then terminate. Otherwise, return to (*) with $k$ incremented by 1. This terminates after some finite number, say $K$, of iterations with $j_K = 0$. Then, $(j_K = 0, j_{K-1}, \ldots, j_2 = i, j_1 = i_1, i_2, \ldots, i_M = n+1)$ is an $R$-MAP by construction. Therefore, every $i \in N^*(R)$ is in some $R$-MAP. By Proposition 6, any $R$-MAP is a $\hat{R}$-AP, so the converse holds. □

**Lemma 4.** *If $R$ is a well-behaved DAG, then $j \in N^*(R) \setminus \{0, n+1\}$ is contained in at least two maximal cliques, and $0$ and $n+1$ each belong to exactly one maximal clique.*

*Proof.* Let $\{C_1, \cdots, C_m\}$ be the set of maximal cliques of $R$. Assume for contradiction that there exists $j \in N^*(R) \setminus \{0, n+1\}$ so that $j$ is only contained in a single maximal clique, $C_i$. Consider an $R$-MAP containing $j$, $(i_0, i_1, \cdots, i_m)$, with $j = i_{j'}$. By Assumption, $i_{j'-1}, i_{j'+1} \in C_i$. Since $R$ is acyclic, $i_{j'-1} R i_{j'+1}$, contradicting that $(i_0, i_1, \cdots, i_m)$ is a $R$-MAP.

By definition, all nodes are part of at least one clique. To see that $n+1$ is only part of one clique, let $i\tilde{R}n+1$ and $j\tilde{R}n+1$. Since $n+1 \not R j'$ for any $j' \in N^*(R)$, $iRn+1$ and $jRn+1$. By perfection, $j\tilde{R}i$.

Finally, we show that 0 is part of exactly one clique. Let $0\tilde{R}i$ and $0\tilde{R}j$. As $R$ is uninformed, this implies that $0Ri$ and $0Rj$. By Lemma 3, there exist $R$-MAPs $(i_0, i_1 = i, \ldots, i_m)$ and $(i'_0, i'_1 = j, \ldots, i'_{m'})$. Let

$$w = \min\{w'' \geq 0 : i'_z R i_{w''+1} \text{ for some } z > 0\}.$$

Since $i'_{m'-1} R i'_{m'} = i_m$, $w$ exists. If $w = 0$, then $i'_z R i$ for some $z > 0$. Since $R$ is well-behaved, $0Ri'_z$, so $z = 1$ since $(i'_0, \ldots, i'_{m'})$ is a $R$-MAP. Since $i'_1 = j$, $jRi$. If $w > 0$, let $w'$ be such that $i'_{w'} R i_{w+1}$. By perfection, $i_w \tilde{R} i'_{w'}$, and $i_w R i'_{w'}$ since $i'_{w'} R i_w$ contradicts the definition of $w$. When $w' > 1$, perfection again requires $i_w \tilde{R} i'_{w'-1}$, and definition of $w$ requires $i_w R i'_{w'-1}$. Inductively, $i_w R i'_{w''}$ for $w'' = w', w'-1, \ldots, 1$, and in particular, $i_w R j$. Since $R$ is well-behaved and $0Rj$, $0Ri_w$. As $(i_0, \ldots, i_m)$ is a $R$-MAP, this implies that $w = 1$, and so $iRj$. Conclude there is exactly one maximal clique containing 0. □

If $C_i$ is adjacent to exactly one other clique $C_j$, then $k \in C_i \setminus C_j$ is in no other clique, so by Lemma 4, $k = 0$ or $k = n + 1$ ($k$ exists since $C_i$ is maximal). Therefore, a MCJT for $R$ consists of a single path between $C_1 \ni 0$ and $C_m \ni n + 1$. We denote it by $(C_1, \ldots, C_m)$ where $C_i$ is adjacent to $C_{i+1}$ for each $i$.

**Lemma 5.** *Let $(C_1, \ldots, C_m)$ be a MCJT for a well-behaved DAG R. If $A_0 = \{0\}$, $A_m = \{n + 1\}$, and $A_i = C_i \cap C_{i+1}$ for $i = 1, \ldots, m - 1$, then $C_i = A_{i-1} \cup A_i$.*

*Proof.* First note that this clearly holds for $i = 1, m$. Thus, pick an $i \in \{2, \cdots, m - 1\}$ and $j \in C_i$. By Lemma 4, there exists $i' \neq i$ so that $j \in C_{i'}$. If $i' > i$, $j \in C_{i+1}$ and hence $j \in A_{i+1}$. If $i' < i$, $j \in C_{i-1}$ and hence $j \in A_{i-1}$. □

**Lemma 6.** *If $(C_1, \ldots, C_m)$ is a MCJT for a well-behaved DAG R, then $k\hat{R}j$ if and only if there exists i so that $k \in C_i \cap C_{i-1}$ and $j \in C_i \setminus C_{i-1}$ where we take $C_0 = \{0\}$ and $C_{m+1} = \{n + 1\}$.*

*Proof.* First we show necessity. Let (*) be the assertion that "if $k \in C_i \cap C_{i-1}$ and $j \in C_i \setminus C_{i-1}$, then $k\hat{R}j$ ." We prove this inductively. For $i = 1$, $k \in C_0 \cap C_1$ iff $k = 0$. For all $j \in C_1 \setminus C_0$, $0\tilde{R}j$ , and since $R$ is uninformed, $0\hat{R}j$. Hence (*) holds for $i = 1$.

Assume (IH) that (*) holds for all $i' \in \{1, \cdots, i - 1\}$ where $i \geq 2$. Take any $k \in C_i \cap C_{i-1}$ and $j \in C_i \setminus C_{i-1}$. Since $i \geq 2$, $k \neq 0$, and there exists $i' \leq i - 1$ for which

$k \in C_{i'} \setminus C_{i'-1}$ and also $l \in C_{i'} \setminus C_{i'+1}$ since $C_{i'} \not\subseteq C_{i'+1}$. By Lemma 5, $l \in C_{i'} \cap C_{i'-1}$, and by IH, $l\hat{R}k$. Since $j \notin C_{i''}$ for any $i'' < i$ and $l \notin C_{i''}$ for any $i'' > i'$, $l\,\tilde{\not R}j$, so by perfection, $j\,\not R k$. Then, $kRj$ since $k, j \in C_i$, and $k\hat{R}j$ by Proposition 6.b.

To complete the proof, we show that “if $j \in C_i \setminus C_{i-1}$ and $k\hat{R}j$, then $k \in C_{i-1} \cap C_i$.” For $i \geq 1$, let $j \in C_i \setminus C_{i-1}$ and $k\hat{R}j$. Take any $j' \in C_i \setminus C_{i+1}$, noting $j' \in C_{i-1} \cap C_i$ by Lemma 5. By necessity, $j'\hat{R}j$. If $j' = k$, then we are done. If not, then $k\tilde{R}j'$ by perfection, so $k \in C_{i'}$ for some $i' \leq i$. Similarly, $k\tilde{R}j$ so $k \in C_{i''}$ for some $i'' \geq i$. Because $(C_1, \ldots, C_m)$ is a MCJT, $k \in C_i$. To see $k \in C_{i-1} \cap C_i$, suppose not, so $k \notin C_{i-1}$. Let $i^* \in C_i \cap C_{i-1}$ be closest to 0. Then, $d(0, j), d(0, k) \leq d(0, i^*) + 1$ since $i^*\hat{R}j$ and $i^*\hat{R}k$ by necessity. Any path from 0 to $j$ or $k$ has some node in $C_{i-1} \cap C_i$, so $d(0, j) = d(0, k) = d(0, i^*) + 1$. Hence by Proposition 6, there is $l$ so that $l\hat{R}k$ and $l\,\not R j$. By the arguments above, replacing $j$ with $k$, $l \in C_i$, so $jRl$. But then $jRlRkRj$, a contradiction. Hence $k \in C_{i-1} \cap C_i$. □

**Lemma 7.** *Let $\rho$ have a perfect SCR $(R \subseteq N^*(R)^2, u)$. Then, $I \in \mathcal{A} \setminus \{\{n+1\}\}$ if and only if $I = C_i \cap C_{i+1}$ for some $i$ where $(C_1, \ldots, C_m)$ is a MCJT for $R$.*

*Proof.* Let $\rho$ have a perfect SCR $(R \subseteq N^*(R)^2, u)$, $(C_1, \ldots, C_m)$ be a MCJT for $R$, $A_i = C_i \cap C_{i+1}$ for each $i$, and $B_i = \cup_{j=1}^{i} C_j$ for each $i$.

Pick any $i \geq 0$. We first show that every $R$-AP intersects $A_{i+1}$. Take any $R$-AP. It contains an $R$-MAP $(i_0, \ldots, i_M)$. Let $i_k$ be first index so that $i_k \notin B_i$. In particular, $i_k \in C_j \setminus C_{j-1}$ for some $j \geq i + 1$. Then, $i_{k-1} \in B_{j-1}$ by Lemma 6, so $j - 1 = i$ and $i_k \in A_{i+1}$. By Lemma 1, $A_{i+1}$ separates.

Let $I \in \mathcal{A} \setminus \{\{n+1\}\}$. By Lemma 1, $I$ intersects every $R$-AP. By Theorem 4.4 of Cowell et al. (1999), $I$ is a clique, so $I \subset C_i$ for some $i$. We show that either $I = A_i$ or $I = A_{i-1}$. Since both $A_i$ and $A_{i-1}$ separate and $I$ is minimal, it suffices to show that either $A_{i-1} \subseteq I$ or $A_i \subseteq I$. If $A_i \cap A_{i-1} \not\subseteq I$, then there exists $j \in A_i \cap A_{i-1} \setminus I$. Let $i' < i - 1$ be such that $j \in C_{i'} \setminus C_{i'-1}$. Then, there exist $j' \in C_{i'} \setminus C_{i'+1} \subseteq C_{i'} \cap C_{i'-1}$ and $l \in C_{i+1} \setminus C_i$. By Lemma 6, $j'\hat{R}j$ and $j\hat{R}l$, and Lemma 3 implies the existence of a $R$-AP that does not intersect $I$. Therefore, $A_i \cap A_{i-1} \subseteq I$. If $A_{i-1} \not\subseteq I$, then there is $j \in A_{i-1} \setminus I$. For any $k \in A_i \setminus A_{i-1}$, $j\hat{R}k$, so Lemma 3 requires that $k \in I$; since $k$ was arbitrary, $A_i \subseteq I$. If $A_i \not\subseteq I$, then there is $k \in A_i \setminus I \subset A_i \setminus A_{i-1}$. For any $j \in A_{i-1}$, $j\hat{R}k$, so Lemma 3 requires that $j \in I$; since $j$ was arbitrary, $A_{i-1} \subseteq I$.

It remains to be shown that $A_i \not\subseteq A_j$ for any $j \neq i$. Suppose not, so $A_i \subseteq A_j$ for $j \neq i$. Consider $j > i$; similar arguments apply when $j < i$. By Lemma 5, $x \in A_i \cap A_j$ implies that $x \in C_{i+1} \cap C_{j+1}$. Since $(C_1, \ldots, C_m)$ is a MCJT, $x \in C_{i+2}$. But then every $x \in A_i$ is also in $A_{i+1}$, and by Lemma 5 , $C_{i+1} = A_i \cup A_{i+1} = A_{i+1} \subseteq C_{i+2}$, contradicting that $C_{i+1}$ is maximal. □

**Lemma 8.** *If $\rho$ has a perfect SCR, then $A_i^*$ exists and is unique for each $i = 1, \ldots, |\mathcal{A}|$.*

*Proof.* Let $\rho$ have a perfect SCR $(R, u)$ and $(C_1, \ldots, C_m)$ be a MCJT for $R$. Denote $A_i = C_i \cap C_{i+1}$ and $B_i = \cup_{j=1}^{i} C_j$ for $i = 1, \cdots, m-1$ as well as $A_m = \{n+1\}$. Note $\mathcal{A} = \{A_i : i = 1, \ldots, m\}$ by Lemma 7.

Clearly, $A_m^* = A_m$. Suppose that $(A_{i+1}^*, \ldots, A_m^*) = (A_{i+1}, \ldots, A_m)$ for $1 \leq i < m$. Theorem 1 of Spiegler (2017), combined with Lemma 5, gives that

$$p_R(x_{N^*(R)}) = p(x_{A_1 \cup \{0\}}) \prod_{j=2}^{m} p(x_{A_j \setminus A_{j-1}} | x_{A_{j-1}}). \tag{1}$$

Since $(C_1, \ldots, C_m)$ is a MCJT, $A_{i+1}^* \cap A_i \supseteq A_j \cap A_{i+1}^*$ for any $j < i$. For any $S = \{a, b\}$ so that $X_{A_{i+1}^* \cap A_i} \perp_S X_{N \setminus [A_{i+1}^* \cap A_i]}$ and $X_{A_{i+1}^* \setminus A_i} \perp_S X_{A_i \setminus A_{i+1}^*}$,

$$\begin{aligned}
\rho_R^S(x_{n+1}|c) &= \sum_{y \in \mathcal{X}_{A_i \cup A_{i+1}}} \rho_R^S(y_{A_i}|c) \rho^S(y_{A_{i+1} \setminus A_i} | y_{A_i \cap A_{i+1}}, y_{A_i \setminus A_{i+1}}) \rho_R^S(x_{n+1}|y_{A_{i+1}}) \\
&= \sum_{y \in \mathcal{X}_{A_i \cup A_{i+1}}} \rho_R^S(y_{A_i \cap A_{i+1}}) \rho_R^S(y_{A_i \setminus A_{i+1}}|c) \rho^S(y_{A_{i+1} \setminus A_i} | y_{A_i \setminus A_{i+1}}) \rho_R^S(x_{n+1}|y_{A_{i+1}}) \\
&= \sum_{y \in \mathcal{X}_{A_{i+1}}} \rho_R^S(y_{A_i \cap A_{i+1}}) \rho^S(y_{A_{i+1} \setminus A_i}) \rho_R^S(x_{n+1}|y_{A_{i+1}})
\end{aligned}$$

where the first equality is just Eq (1), the second comes from $X_{A_{i+1}^* \cap A_i} \perp_S X_{N \setminus [A_{i+1}^* \cap A_i]}$, and the third from $X_{A_{i+1}^* \setminus A_i} \perp_S X_{A_i \setminus A_{i+1}^*}$. In particular, $A_i$ satisfies the condition to be $A_i^*$.

We show that no $A_{\hat{i}}$ satisfies the condition for $\hat{i} < i$. If $A_{i+1} \cap A_{\hat{i}} \neq A_{i+1} \cap A_i$ for all $\hat{i} < i$, then the only candidate for $A_i^*$ is $A_i$. If $A_{i+1} \cap A_{\hat{i}} = A_{i+1} \cap A_i$ for some $\hat{i} < i$, then observe that

$$A_{\hat{i}} \setminus [A_i \cup A_{i+1}^*] = A_{\hat{i}} \setminus C_{i+1} \supset A_{\hat{i}} \setminus C_{\hat{i}+2} \neq \emptyset$$

by Lemma 5 and that $(C_1, \ldots, C_m)$ is a MCJT. Pick $j \in A_{\hat{i}} \setminus [A^*_{i+1} \cup A_i]$, $k \in A^*_{i+1} \setminus A_{\hat{i}}$, and $I' \subset \bigcup_{l=1}^{i} A_l \setminus A^*_{i+1}$ so that $|I' \cap A_l| = 1$ for every $l < i+1$ and $j \in I'$. Set

$$J^- = I' \cap \{A_{i'} : i' < i+1 \ \& \ A_{i'} \cap A^*_{i+1} = A_i \cap A^*_{i+1}\} \setminus \{j\},$$

noting $|J^-| \geq 1$. For $1 < l' \leq |\mathcal{A}| - i$, pick $i_{l'} \in A^*_{i+l'} \setminus A^*_{i+l'-1}$. By Lemma 6, $i_{l'} R i_{l'+1}$ and that for every $i' \in I'$, there exist unique $i_-, i_+ \in I' \cup \{0, k\}$ so that $i_- R i' R_+$. Let $I = I' \cup \{k, i_2, \ldots, i_{|\mathcal{A}|-i}\}$, $k^* = |I'| + 1$, and $\pi$ be a bijection from $\{0, \ldots, |I|\}$ to $I \cup \{0\}$ with $\pi(0) = 0$, $\pi(j') \in I' \setminus J^-$ for $1 \leq j' < j^*$, $\pi(j^*) = j$, $\pi(j') \in J^-$ when $j' \in (j^*, k^*)$, $\pi(k^*) = k$, and $\pi(j') = i_{j'-k^*}$ for $j' > k^*$. Consider $\{a^\pi, b^\pi\} \in \mathcal{S}$ so that:

(i) $\operatorname{supp} \operatorname{marg}_{i'} \frac{1}{2} a^\pi + \frac{1}{2} b^\pi = \{\bar{y}, \underline{y}\} \subset \mathcal{X}_{n+1}$ with $\bar{y} > \underline{y}$ for all $i' \in I \cup \{n+1\}$;

(ii) for all $i' \notin [j^*, k^*)$, each $c \in \{a, b\}$, and $a^\pi$-a.e. $x, x' \in \mathcal{X}_{-0}$,

$$c^\pi \left( x_{\pi(i+1)} | x_{\pi(i)}, x'_{\{1, \ldots, \pi(i-1)\}} \right) = a^\pi \left( x_{\pi(i+1)} | x_{\pi(i)} \right);$$

(iii) for all $i' \notin [j^*, k^*)$, $a^\pi(\bar{y}_{\pi(i'+1)} | \bar{y}_{\pi(i')}) > a^\pi(\bar{y}_{\pi(i'+1)} | \underline{y}_{\pi(i')})$;

(iv) $X_{\{k\}} \perp_S X_{\{\pi(1), \ldots, \pi(j^*)\}}$ and $X_{N \setminus I} \perp_{\{a^\pi, b^\pi\}} X_I$; and

(v) $a^\pi(\bar{y}_{\pi(1)}) > b^\pi(\bar{y}_{\pi(1)})$, for all $i' \in (j^*, k^*)$

$$a^\pi \left( \bar{y}_{\pi(i')} | x_{\pi(k^*)}, x_{\pi(j^*)} \right) = \begin{cases} \frac{3}{4} & if \ (x_{\{\pi(k^*), \pi(j^*)\}}) = (\bar{y}, \bar{y}) \\ \frac{1}{4} & otherwise \end{cases}$$

and

$$c \left( x_{\pi(i')} | x_{\pi(k^*)}, x_{\{\pi(1), \ldots, \pi(i'-1)\}} \right) = a^\pi \left( x_{\pi(k^*-1)} | x_{\pi(k^*)}, x_{\pi(j^*)} \right)$$

for each $c \in \{a^\pi, b^\pi\}$ and $a^\pi$-a.e. $x \in \mathcal{X}_{-0}$.

By (iv), $X_{A^*_{i+1} \cap A_{\hat{i}}} \perp_{\{a^\pi, b^\pi\}} X_{N \setminus [A^*_{i+1} \cap A_{\hat{i}}]}$ and $X_{A_{\hat{i}} \setminus A^*_{i+1}} \perp_{\{a^\pi, b^\pi\}} X_{A^*_{i+1} \setminus A_{\hat{i}}}$.

For any $\alpha \in (0,1)$, let $p = \alpha a^\pi + (1-\alpha) b^\pi$ and $\pi'$ be the index for $I$ that agrees with $R$, noting that $\pi'(j') = \pi(j')$ for all $j' \geq k^*$, that each $j' \in I' \cup \{k\}$ has exactly one parent in $I' \cup \{0\}$ by Lemma 6, and that $\pi'(j') \in [j^*, k^*)$ for all $j' \in \{j\} \cup J^-$. We show that $p_R(\bar{y}_{n+1} | a^\pi) > p_R(\bar{y}_{n+1} | b^\pi)$ whenever $j \notin A_i$. Observe that

$$p_R(x_{n+1} | c^\pi) = \sum_{y \in \{\bar{y}, \underline{y}\}^I} c^\pi(y_{\pi'(1)}) \prod_{j'=1}^{|I|-2} p(y_{\pi'(j'+1)} | y_{\pi'(j')}) p(x_{n+1} | y_{\pi'(|I|-1)}).$$

By (ii) and (iii),

$$p(\bar{y}_{\pi'(j'+1)} | \bar{y}_{\pi'(j')}) > p(\bar{y}_{\pi'(j'+1)} | \underline{y}_{\pi'(j')}) \tag{2}$$

for all $j' \notin [j^*, k^*)$. Note that $\pi'(k^*) = k$ and that $\pi'(k^* - 1) \in J^-$ when $j \notin A_i$. For any $j' \in J^-$, by (v) we have

$$p(\bar{y}_{\pi'(k^*)}|\bar{y}_{j'}) = \frac{p(\bar{y}_j)p(\bar{y}_k)\frac{3}{4}}{p(\bar{y}_j)p(\bar{y}_k)\frac{1}{2} + \frac{1}{4}} > \frac{p(\bar{y}_j)p(\bar{y}_k)\frac{1}{4}}{\frac{3}{4} - p(\bar{y}_j)p(\bar{y}_k)\frac{1}{2}} = p(\bar{y}_{\pi'(k^*)}|\underline{y}_{j'}),$$

so Eq. (2) also holds for $j' = k^* - 1$. Similarly, one can verify using (v) and similar arguments to the above that Eq. (2) holds for all $j' \in [j^*, k^* - 2]$. Now, $a^\pi(\bar{y}_{\pi'(1)}) > b^\pi(\bar{y}_{\pi'(1)})$ by (iii) and $a^\pi(\bar{y}_{\pi(1)}) > b^\pi(\bar{y}_{\pi(1)})$. Successively applying Eq. (2), we have $p_R(\bar{y}_{n+1}|a^\pi) > p_R(\bar{y}_{n+1}|b^\pi)$. Applying to $p = \rho^{\{a^\pi, b^\pi\}}$, we see that $\rho_R^{\{a^\pi,b^\pi\}}(\bar{y}_{n+1}|a^\pi) > \rho_R^{\{a^\pi,b^\pi\}}(\bar{y}_{n+1}|b^\pi)$. Therefore, $\rho(a^\pi, \{a^\pi, b^\pi\}) > \frac{1}{2}$. □

Combining Lemmas 5, 6, and 8, we have the following.

**Lemma 9.** *If $\rho$ have a perfect SCR $(R, u)$, then $j\hat{R}k$ if and only if there exists $i$ so that $j \in A_i^*$ and $k \in A_{i+1}^* \setminus A_i^*$ where $A_0^* = \{0\}$.*

**Lemma 10.** *For a well-behaved DAG $R$ with $j, k \in N^*(R)$, if $jRk$ and $j \not\hat{R} k$, then there exists $l$ such that $j\hat{R}l$ and $k\hat{R}l$.*

*Proof.* Pick any $j, k \in N^*(R)$ so that $jRk$ and $j \not\hat{R} k$. Let $(C_1, \dots, C_m)$ be a MCJT for $R \cap N^*(R)^2$, noting that $j, k \in C_i$ for some $i$. If $j \notin C_{i+1}$, then $k \notin C_{i+1}$ since $j \not\hat{R} k$, so by Lemma 4, we can pick $i$ so that $j, k \in C_i \cap C_{i+1}$. There exists $l \in C_{i+1} \setminus C_i$, so $j\hat{R}l$ and $k\hat{R}l$ by Lemma 6. □

*Proof of Lemma 2.* Lemma 9, combined with the observation that the ordering for the minimal separators according to Definition 3 is unique by Lemma 8, immediately implies the result. □

**Necessity:** Immediately follows from Lemma 2.

**Sufficiency:** Suppose that $\rho$ has a perfect SCR $(R, u)$. Let $R'$ be a well-behaved DAG so that $(i_1, \dots, i_m)$ is a $R'$-MAP if and only if it is also a $R$-MAP. WLOG, $R' \subset N^*(R')^2 = N^*(R)^2$ by Lemma 3. By Lemmas 8 and 9, $i\hat{R}j$ if and only if $i\hat{R}'j$. By Lemma 10, $(i, j) \in R' \setminus \hat{R}'$ if and only if $i\hat{R}'k$ and $j\hat{R}'k$ for some $k$. Since $\hat{R}' = \hat{R}$ and $R$ is perfect, either $iRj$ or $jRi$. Hence, $R^* = R \cap N^*(R)^2$ and $R'$ have the same skeleton and v-colliders. By Proposition 4, $\rho_{R^*}^S = \rho_{R'}^S$, so $\rho$ has an SCR $(R', u)$ if and only if it has an SCR $(R^*, u)$. By definition of $N^*(R)$ and hypothesis, $\rho$ has an SCR $(R^*, u)$.

Uniqueness of $u$ follows from the uniqueness results for Logit and EU since $\rho$ has a Logit-EU representation on correctly perceived menus. □

*Proof of Proposition 1.* Clearly, if $n+1 \in J$, then $J$ separates. For any $J \not\ni n+1$, we construct a menu $S_J = \{a_J, b_J\}$ so that $J$ separates if and only if $\rho(a_J, S_J) = \frac{1}{2}$. Pick $\bar{y}, \underline{y} \in \mathcal{X}_{n+1}$ with $\bar{y} > \underline{y}$, and let $\underline{x}, \bar{x} \in \mathcal{X}_{-0}$ be such that $\underline{x}_i = \underline{y}$ and $\bar{x}_i = \bar{y}$ for all $i$. Let $a_J(\bar{x}_J, \bar{x}_{N\setminus J}) = 1-\epsilon = b_J(\bar{x}_J, \underline{x}_{N\setminus J})$, $c_J(\bar{x}_J, x_{N\setminus J}) = \epsilon\kappa$ for any $x \in \mathcal{X}_{-0} \setminus \{\underline{x}, \bar{x}\}$ with $x_{n+1} = \underline{y}$, and $b_J(\bar{x}_J, \bar{x}_{N\setminus J}) = \epsilon\kappa = a_J(\bar{x}_J, \underline{x}_{N\setminus J})$ where $\kappa = (2^{n+1-|J|-1} - 2)^{-1}$ and $(1-\epsilon)^{n+1} > \frac{1}{2}$. If $J$ separates, then clearly $\rho(a_J, S_J) = \frac{1}{2}$.

Suppose $J$ does not separate. Consider $p = \frac{1}{2}a_J + \frac{1}{2}b_J$. By Equation (1),

$$p_R(\bar{y}_{n+1}|a) \geq a(\bar{x}_{A_1\setminus J}) \prod_{j=2}^{|\mathcal{A}|} p(\bar{x}_{A_j\setminus[A_{j-1}\cup J]}|\bar{x}_{A_{j-1}\setminus J}).$$

Since $J$ does not separate, $J \not\supset A_j, A_{j-1}$ for each $j$, and we have

$$p(\bar{x}_{A_j\setminus[A_{j-1}\cup J]}|\bar{x}_{A_{j-1}\setminus J}) = \frac{\frac{1}{2}(1-\epsilon)}{\frac{1}{2}(1-\epsilon) + \frac{1}{2}\epsilon\kappa} > 1-\epsilon$$

Since there are no more than $n+1$ minimal separators, $p_R(\bar{y}_{n+1}|a) \geq (1-\epsilon)^{n+1} > \frac{1}{2}$. Symmetrically, $p_R(\underline{y}_{n+1}|b) \geq (1-\epsilon)^{n+1} > \frac{1}{2}$, so $\operatorname{marg}_{n+1} p_R(\cdot|a)$ first order dominates $\operatorname{marg}_{n+1} p_R(\cdot|b)$. If $\rho(a_J, S_J) = \frac{1}{2}$, then $\rho^S = p$, so conclude that $\rho(a_J, S_J) \neq \frac{1}{2}$. Consequently, we can determine $\mathcal{A}$ from the finite set $\{\rho(\cdot, S_J) : J \subset N, |J| \geq 1\}$.

To find $A_i^*$, we can use the proof of Lemma 8. Clearly, $\{n+1\} = A_m^*$ when $m = |\mathcal{A}|$, so assume we have found $A_{i+1}^*, \ldots, A_m^*$. Let $\mathcal{A}_i^*$ be the members of $\mathcal{A} \setminus \{A_{i+1}^*, \ldots, A_m^*\}$ with the largest intersections with $A_{i+1}^*$. If $A, B \in \mathcal{A}_i^*$, then $A \cap A_{i+1}^* = B \cap A_{i+1}^*$. If $|\mathcal{A}_i^*| = 1$, then $A_i^*$ its unique member. Otherwise, notice that $A_i \setminus [A_{i+1}^* \cup A_{i-1}] \neq \emptyset$, so there exists $j^* \in A_i \setminus A_{i+1}^*$ for which $j^* \notin A_{i'}$ for all $i' < i$ using properties of a junction tree and Lemma 5. Then, $A_i^*$ is $A^* \in \mathcal{A}_i^*$ if and only if there exists $j^* \in A^* \setminus A_{i+1}^*$ so that $j^* \notin A$ for all $A \in \mathcal{A}_i^* \setminus A^*$ and $\rho(a^\pi, \{a^\pi, b^\pi\}) = \frac{1}{2}$ where $\{a^\pi, b^\pi\}$ is constructed as in the proof of Lemma 8 using $j = j^*$. When $A^* \neq A_i$, $j^* \notin A_i$ so Lemma 8 gives that $\rho(a^\pi, \{a^\pi, b^\pi\}) > \frac{1}{2}$. When $A^* = A_i$, the first part of Lemma 8 gives that

$\rho(a^\pi, \{a^\pi, b^\pi\}) = \frac{1}{2}$. Hence, we can determine $A_i^*$ from $\rho$ evaluated at one such menu for each $A \in \mathcal{A}_i^*$. □

*Proof of Proposition 2.* Suppose that $\rho$ has a perfect SCR. Let $j\hat{R}^\rho k$ and $R$ represent $\rho$. Then, Lemma 9 gives that $jRk$. Since $R$ was arbitrary, $j$ is revealed to cause $k$. Now, let $j$ be revealed to cause $k$ and $R$ represent $\rho$. If $j\hat{R}k$, then $j\hat{R}^\rho k$ by Lemma 9. If $j\not\hat{R}k$, then per definition, there is a perfect, uninformed, nontrivial DAG $R'$ equivalent to $R$ for which $kR'j$. Then, $\rho$ is also represented by $R'$, a contradiction. Hence, $j$ is revealed to cause $k$ if and only if $j\hat{R}^\rho k$. □

### A.3. **Proof of Theorem 2.**

**Lemma 11.** *If $\rho$ satisfies Axioms 1, 3, 5, 6, 7, and 8, then there exists a continuous, strictly-increasing $u : \mathcal{X}_{n+1} \to \mathbb{R}$ so that*

$$\rho(a, S) = \frac{\exp\left(\int_{\mathcal{X}_{n+1}} u(c)da(c_{n+1})\right)}{\sum_{b\in S}\exp\left(\int_{\mathcal{X}_{n+1}} u(c)db(c_{n+1})\right)}$$

*for every correctly perceived $S \in \mathcal{S}$, and $u$ is unique up to adding a constant.*

*Proof of Lemma 11.* Say that $\rho$ has a Luce representation $u$ on a subset of menus $\Sigma \subset \mathcal{S}$ if for every $S \in \Sigma$, $\rho(a, S) = u(a)/\sum_{b\in S} u(b)$ for every $a \in S$.

For any finite $Y \subset \mathcal{X}_{n+1}$, let $P(Y, \epsilon) = \{p \in \Delta\mathcal{X}_{n+1} : p(Y) = 1, p(y) \geq \epsilon \forall y \in Y\}$ for $\epsilon \in (0, \frac{1}{M})$ with $M = |Y|$. For any $p_1, \dots, p_m \in P(Y, \epsilon)$, there is an $R^\rho$-Markov menu $S = \{a_1, \dots, a_m\}$ so that $\operatorname{marg}_{n+1} a_i = p_i$. To see why, let $(i_0 = 0, i_1, \dots, i_k = n + 1)$ be a $R^\rho$-MAP, and label $Y = \{y_1, \dots, y_M\}$. Take $A(Y, \eta) \subset \Delta\mathcal{X}$ to be the lotteries so that (i) $X_{i_1}$ takes values in $\{1, \dots, M\}$; (ii) for $j = 1, \dots, k - 2$, $X_{i_{j+1}} = X_{i_j}$ with probability $(1 - \eta)$ and equals every other value in $\{1, \dots, M\}$ with equal probability; (iii) $X_{n+1} = y_i$ with probability 1 whenever $X_{i_{k-1}} = i$; and (iv) $X_j = 0$ with probability 1 for every $j \notin \{i_0, \dots, i_k\}$. For each $i = 1, \dots, M$, consider $a \in A(Y, \eta)$ so that $a(i_1) = 1 - \gamma$, $a(j_1) = \gamma/(M - 1)$ for $j \neq i$. Then, $a(X_{n+1} = y_i) \geq (1 - \gamma)(1 - \eta)^k$, which approaches 1 as $\eta, \gamma \to 0$. Moreover $a(X_{n+1} = y_j) = a(X_{n+1} = y_{j'})$ for all $j, j' \neq i$, and $a(X_{n+1} = y_j) \to 0$ as $\eta, \gamma \to 0$. Observe that $A(Y, \eta)$ is convex, so given $\epsilon > 0$, there exists $\eta > 0$ so that for any $p_i \in P(Y, \epsilon)$ there is an $a_i \in A(Y, \eta)$ so that $\operatorname{marg}_{n+1} a_i = p_i$.

Let $P(Y) = \cup_{\epsilon>0} P(Y,\epsilon)$. Since any $S' \in \mathcal{S} \cap A(Y,\eta)$ is correctly perceived, Axiom 5 and standard results imply there is a Luce representation when restricted to subsets of $A(Y,\eta)$ for each $\eta > 0$; let $u_\eta$ be its index. By Axiom 8, $u_\eta(a)/u_\eta(b) = u_{\eta'}(a')/u_{\eta'}(b')$ whenever $\operatorname{marg}_{n+1} a = \operatorname{marg}_{n+1} a'$, $\operatorname{marg}_{n+1} b = \operatorname{marg}_{n+1} b'$, $a, b \in A(Y,\eta)$, and $a', b' \in A(Y,\eta')$. Pick $\eta^*$ and $a \in A(Y,\eta^*)$. Normalize $u_{\eta'}$ for each $\eta' < \eta^*$ so that $u_{\eta'}(a) = u_{\eta^*}(a)$. Since there is one degree of freedom, $u_{\eta'}(b) = u_{\eta''}(b)$ for any $b \in A(Y,\eta') \cap A(Y,\eta'')$, so $\hat{u}_Y$ can be defined unambiguously via $\hat{u}_Y(p) = u_{\eta'}(p)$ for any $\eta'$ such that $p \in A(Y,\eta')$. By Axiom 8, $\hat{u}_Y$ is continuous and $\hat{u}_Y(a) = \hat{u}_Y(b)$ whenever $\operatorname{marg}_{n+1} a = \operatorname{marg}_{n+1} b$, so we can decompose $\hat{u}_Y = \dot{u}_Y \circ \operatorname{marg}_{n+1}$.

For each $Y$, let $p_Y = \left(\frac{1}{|Y|}, y\right)_{y\in Y}$. Extend $\dot{u}_Y$ to $q \in \Delta(Y)$ via the formula

$$u_Y(q) = \exp[2 \ln \dot{u}_Y(\frac{1}{2}q + \frac{1}{2}p_Y) - \ln \dot{u}_Y(p_Y)].$$

Axiom 6 gives that $u_Y(q) = \dot{u}_Y(q)$ when $q \in P(Y)$. We show $u_Y \circ \operatorname{marg}_{n+1}$ is a Luce representation of $\rho$ on

$$\mathcal{S}_Y = \{S \in \mathcal{S} : \operatorname{marg}_{n+1} a \in \Delta(Y) \ \forall a \in S \text{ and } S \text{ is correctly perceived}\}.$$

Take any $S \in S_Y$. Pick any $p^*, q^* \in S$. Let $r = \operatorname{marg}_{n+1} r^* = r$ for $r = p, q$ and $\{p_m, q_m\} \in \mathcal{S} \cap A(Y, \frac{1}{m|Y|+1})$ have $\operatorname{marg}_{n+1} r_m = \frac{1}{m} p_Y + \frac{m-1}{m} r$ for $r = p, q$ and $m = 1, 2, \ldots$. By Axiom 6, for $m > 2$ we have

$$\frac{\rho(p_m, \{p_m, q_m\})}{\rho(q_m, \{p_m, q_m\})} = \frac{\exp[\frac{2m-2}{m} \ln u_Y(\frac{1}{2}p + \frac{1}{2}p_Y) + \frac{2-m}{m} \ln u_Y(p_Y)]}{\exp[\frac{2m-2}{m} \ln u_Y(\frac{1}{2}q + \frac{1}{2}p_Y) + \frac{2-m}{m} \ln u_Y(p_Y)]} = \frac{u_Y(\operatorname{marg}_{n+1} p_m)}{u_Y(\operatorname{marg}_{n+1} q_m)}.$$

By Axiom 8, $\frac{\rho(p_m,\{p_m,q_m\})}{\rho(q_m,\{p_m,q_m\})} \to \frac{\rho(p^*,S)}{\rho(q^*,S)} = \frac{u_Y(p)}{u_Y(q)}$. Since $S$, $p^*$ and $q^*$ were arbitrary, this establishes the claim.

Pick any finite $Y^* \subset \mathcal{X}_{n+1}$. Define $u(q) = u_{Y^*}(q)$ for any $q$ with support in $Y^*$. For any $Y \supset Y^*$ and $q \in P(Y)$, define $u(q) = \lambda u_Y(q)$ where $\lambda = u(p_{Y^*})/u_Y(p_{Y^*})$. We claim that $\rho$ has a Luce representation $U$ on the correctly perceived menus. Pick any correctly perceived $S$ and any $p, q \in S$. Let $Y \supset Y^*$ be so that $\operatorname{supp} r \subset Y$ for $r = p, q$. By the above, $\frac{\rho(p,S)}{\rho(q,S)} = \frac{u_Y(\operatorname{marg}_{n+1} p)}{u_Y(\operatorname{marg}_{n+1} q)} = \frac{u(\operatorname{marg}_{n+1} p)}{u(\operatorname{marg}_{n+1} q)}$, and by usual uniqueness results $u$ is well-defined. Conclude $U = u \circ \operatorname{marg}_{n+1}$ represents $\rho$ on the $R^\rho$-Markov menus.

Now, let $V = \ln u$. Observe that $V$ is affine (by Axiom 6), strictly-increasing in FOSD (by Axiom 7), continuous (by Axiom 8), and ranks every lottery in $\Delta(\mathcal{X}_{n+1})$.

Conclude there exists a continuous strictly-increasing $v : \mathcal{X}_{n+1} \to \mathbb{R}$ so that $V(p) = \int v(c)dp(c)$ for any $p \in \Delta\mathcal{X}_{n+1}$, completing the proof. □

**Necessity:** Suppose that $\rho$ has a perfect SCR. Theorem 1 implies that $\rho$ has an SCR $(R^\rho, u)$ and that Axiom 3 holds. Since for any DAG $R$

$$\int u(c)dp_R(c_{n+1}|a') \in \left[\min_{x\in\mathcal{X}_{n+1}} u(x), \max_{x\in\mathcal{X}_{n+1}} u(x)\right],$$

Axiom 2 holds. Axiom 5 follows from continuity of the expected utility functional and that the hypothesis implies $\rho^{S_m}_{R^\rho}(\cdot|a) \to \rho^{S}_{R^\rho}(\cdot|a)$ for every $a \in S$. For a correctly perceived $S$,

$$\ln\frac{\rho(a,S)}{\rho(b,S)} = \int_{\mathcal{X}_{n+1}} u(c)da(c_{n+1}) - \int_{\mathcal{X}_{n+1}} u(c)db(c_{n+1}),$$

so when $a = \alpha a' + (1-\alpha)b$ and $\{a, b\} = S$ is correctly perceived,

$$\ln\frac{\rho(a,S)}{\rho(b,S)} = \alpha\int_{\mathcal{X}_{n+1}} u(c)d[a'-b](c_{n+1}),$$

so Axiom 6 holds. Axioms 1, 4, 7, and 8 are easily seen to be necessary.

**Sufficiency:** Suppose that $\rho$ satisfies the axioms. Axiom 3 implies $A^*_1, \dots, A^*_{|\mathcal{A}|}$ exist, and so $R^\rho$ is a perfect, uninformed, nontrivial DAG. Given Lemma 11, we approximate choice in every menu by a sequence of correctly menus. We show that

$$\frac{\rho(a,S)}{\rho(b,S)} = \frac{\exp[\int_{\mathcal{X}_{n+1}} u(c)d\rho^S_{R^\rho}(c_{n+1}|a)]}{\exp[\int_{\mathcal{X}_{n+1}} u(c)d\rho^S_{R^\rho}(c_{n+1}|b)]} \tag{3}$$

for any $a, b \in S$ and any $S \in \mathcal{S}$. Then, $\rho$ has an SCR $(R^\rho, u)$ since $\sum_{a\in S}\rho(a,S) = 1$.

Pick any $S \in \mathcal{S}$ and any $a, b \in S$. Let $a'(y) = \rho^S_{R^\rho}(y|a)$ and $b'(y) = \rho^S_{R^\rho}(y|b)$ for every $y \in \mathcal{X}_{-0}$. Since $\{a', b'\}$ is correctly perceived, $\rho(a', \{a', b'\})/\rho(b', \{a', b'\})$ has the desired form by Lemma 11. If $a' = b'$, then $\mathrm{marg}_{A^*_1} a = \mathrm{marg}_{A^*_1} b$, so $\rho(a, S) = \rho(b, S)$ by Axiom 4, and Equation (3) holds.

Otherwise, let $S_1 = \{a', b'\}$ and recursively define $S_m = S_{m-1} \cup \{\frac{1}{m}a' + \frac{m-1}{m}b'\}$. Each $S_m$ is correctly perceived by construction, and each has $m+1$ distinct alternatives. By Axiom 2, there exists $K > 0$ so that for any $a'', b'' \in S'' \in \mathcal{S}$, $\frac{\rho(a'',S'')}{\rho(b'',S'')} \le K$. In particular, for $S_m \setminus S = \{s_1, \dots, s_{M(m)}\}$ (noting $M(m) \ge m + 1 - |S|$), $a'' \in S$, and

$i \leq M(m)$, we have $\rho(s_i, S_m \cup S) \geq K^{-1}\rho(a'', S_m \cup S)$. Then,

$$1 \geq \sum_{i \leq M(m)} \rho(s_i, S_m \cup S) + \rho(a'', S_m \cup S) \geq [M(m)K^{-1} + 1]\rho(a'', S_m \cup S)$$

so $\rho(a'', S_m \cup S) \leq \frac{K}{m+1-|S|+K} \to 0$ as $m \to \infty$.

For $p_m = \rho^{S_m \cup S}$, arbitrary $i \leq |\mathcal{A}|$, and $E = A^*_{i+1} \setminus A^*_i$, we have $p_m(x_E|x_{A^*_i})$ equals

$$\frac{1}{p_m(x_{A^*_i})}\left[\sum_{a'' \in S} p_m(a'')p_m(x_{A^*_i}|a'')a''(x_E|x_{A^*_i}) + p_m(S_m)p_m(x_{A^*_i}|x_0 \in S_m)a'(x_E|x_{A^*_i})\right]$$

for $p$-a.e. $x \in \mathcal{X}_{-0}$ since $\hat{a}(x_E|x_{A^*_i}) = a'(x_E|x_{A^*_i})$ for all $\hat{a} \in S_m$. This converges to $\rho^{S_1}(x_E|x_{A^*_i}) = a'(x_E|x_{A^*_i})$ because $p_m(a'') \to 0$ for all $a'' \in S$. Since $i$ was arbitrary, $\rho^{S_m \cup S}(x_{A^*_{i+1} \setminus A^*_i}|x_{A^*_i}) \to \rho^{S_1}(x_{A^*_{i+1} \setminus A^*_i}|x_{A^*_i})$ for every $i$.

Axiom 4 gives that $\rho(a, S_m \cup S) = \rho(a', S_m \cup S)$ and $\rho(b, S_m \cup S) = \rho(b', S_m \cup S)$. Axiom 5 implies that

$$\frac{\rho(a', S_m \cup S)}{\rho(b', S_m \cup S)} = \frac{\rho(a, S_m \cup S)}{\rho(b, S_m \cup S)} \to \frac{\rho(a', S_1)}{\rho(b', S_1)}$$

and that

$$\frac{\rho(a, S_m \cup S)}{\rho(b, S_m \cup S)} = \frac{\rho(a', S_m \cup S)}{\rho(b', S_m \cup S)} \to \frac{\rho(a, S)}{\rho(b, S)}.$$

Therefore, $\frac{\rho(a', S_1)}{\rho(b', S_1)} = \frac{\rho(a, S)}{\rho(b, S)}$ and Equation (3) holds for $a, b$. Since $a$, $b$, and $S$ were arbitrary, $\rho$ has a perfect SCR $(R^\rho, u)$. □

### A.4. Proof of Proposition 3.

Suppose that $\rho_i$ has a perfect SCR $(R_i, u_i)$ for $i = 1, 2$ and that $\rho_2$ has a coarser model than $\rho_1$. Let $(i_0, \ldots, i_m)$ be a $R_2$-MAP, $I = \{i_0, \ldots, i_m\}$, and $S_I = \{S \in \mathcal{S} : X_i \perp_S X_{N \setminus i}$ for all $i \in N \setminus I\}$. Restricted to menus in $S_I$, $\rho_i$ has an SCR $(R'_i, u_i)$ where $R'_i = R_i \cap I \times I$, and by construction $\rho_1(\cdot, S) = \rho_2(\cdot, S)$ for all $S \in S_I$. Applying Theorem 1 gives that $(i_0, \ldots, i_m)$ is a $R_1$-MAP. Similar arguments show that any $R_1$-MAP with covariates contained in $N^*(R_2)$ is also a $R_2$-MAP. Letting $R^*_2 = R_1 \cap [N^*(R_2) \times N^*(R_2)]$, the set of $R^*_2$-MAPs coincides with the set of $R_2$-MAPs, so applying Theorem 1 establishes the result.

Conversely, let $\rho_i$ have a perfect SCR $(R_i, u_i)$ for $i = 1, 2$ $u_2 = u_1 + \beta$ and $R_2 = R_1 \cap [N' \times N']$ for some $N' \subset \{0, \ldots, n+1\}$. Pick any $S \in \mathcal{S}$ so that $X_i \perp_S X_{N \setminus \{i\}}$ for all $i \notin N^*(R_2)$ and any $p \in co(S) \subset \Delta\mathcal{X}$ with full support. Since $N' \supset N^*(R_2)$, we also have $X_i \perp_S X_{N\setminus\{i\}}$ for all $i \notin N'$, so for every $i$ and $p$-a.e. $x \in \mathcal{X}$,

$$p\left(x_i|x_{R_1(i)}\right) = p\left(x_i|x_{R_1(i)\cap N'}, x_{R_1(i)\setminus N'}\right) = p\left(x_i|x_{R_1(i)\cap N'}\right) = p\left(x_i|x_{R_2(i)}\right).$$

Hence, $p_{R_1} = p_{R_2}$, and the set of $R_1$-personal equilibriums for $S$ equals the set of $R_2$-personal equilibriums for $S$. □

## References

Andre, P., Pizzinelli, C., Roth, C., and Wohlfart, J. (2021). Subjective models of the macroeconomy: Evidence from experts and a representative sample. *Available at SSRN 3355356*.

Apesteguia, J. and Ballester, M. A. (2018). Monotone Stochastic Choice Models: The Case of Risk and Time Preferences. *Journal of Political Economy*, 126(1):74–106.

Bohren, J. A. and Hauser, D. (2018). Social learning with model misspecification: A framework and a robustness result. *working paper*.

Brady, R. L. and Rehbeck, J. (2016). Menu-dependent stochastic feasibility. *Econometrica*, 84(3):1203–1223.

Card, D. (1999). The causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pages 1801–1863. Elsevier.

Cattaneo, M., Ma, X., Masatlioglu, Y., and Suleymanov, E. (2020). A random attention model. *Journal of Political Economy*, 128.

Chambers, C. P., Cuhadaroglu, T., and Masatlioglu, Y. (2021). Behavioral influence. *working paper*.

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

Denrell, J. (2018). Sampling biases explain decision biases. pages 49–95. Oxford University Press.

Eliaz, K. and Spiegler, R. (2018). A model of competing narratives.

Eliaz, K., Spiegler, R., and Thysen, H. C. (2019). On persuasion with endogenous misspecified beliefs.

Eliaz, K., Spiegler, R., and Weiss, Y. (2020). Cheating with models. *American Economic Review: Insights*.

Ellis, A. and Masatlioglu, Y. (2021). Choice with Endogenous Categorization. *The Review of Economic Studies*, Forthcoming.

Ellis, A. and Piccione, M. (2017). Correlation misperception in choice. *American Economic Review*, 107(4):1264–92.

Esponda, I. (2008). Behavioral equilibrium in economies with adverse selection. *American Economic Review*, 98(4):1269–91.

Esponda, I. and Pouzo, D. (2016). Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84(3):1093–1130.

Esponda, I. and Vespa, E. (2018). Endogenous sample selection: A laboratory study. *Quantitative Economics*, 9(1):183–216.

Eyster, E. and Rabin, M. (2005). Cursed equilibrium. *Econometrica*, 73(5):1623–1672.

Frick, M., Iijima, R., and Ishii, Y. (2019). Misinterpreting others and the fragility of social learning. *working paper*.

Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, 74(1):121–146.

He, K. (2018). Mislearning from censored data: The gambler's fallacy in optimal-stopping problems. *working paper*.

Heidhues, P., Koszegi, B., and Strack, P. (2018). Unrealistic expectations and misguided learning. *Econometrica*, 86(4):1159–1214.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79.

Jehiel, P. and Koessler, F. (2008). Revisiting games of incomplete information with analogy-based expectations. *Games and Economic Behavior*, 62(2):533–557.

Ke, S., Zhao, C., Wang, Z., and Hsieh, S.-L. (2020). Behavioral neural networks. *Working paper*.

Kochov, A. (2018). A behavioral definition of unforeseen contingencies. *Journal of Economic Theory*, 175:265–290.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Köszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, 121(4):1133–1165.

Lang, K. and Kahn-Lang Spitzer, A. (2020). Race discrimination: An economic perspective. *Journal of Economic Perspectives*, 34(2):68–89.

Langer, E. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32:311–328.

Levy, G., Razin, R., and Young, A. (2021). Misspecified politics and the recurrence of populism. *working paper*.

Lipman, B. L. (1999). Decision theory without logical omniscience: Toward an axiomatic framework for bounded rationality. *The Review of Economic Studies*, 66(2):pp. 339–361.

Lu, J. (2016). Random choice and private information. *Econometrica*, 84(6):1983–2027.

Luce, R. D. (1959). Individual choice behavior.

Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *American Economic Review*, 102(5):2183–2205.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M., and Prat, A. (2021). Competing models. *working paper*.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

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

Samuelson, L. and Mailath, G. (2019). Learning under diverse world views: Model based inference. *American Economic Review*.

Samuelson, W. and Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1:7–59.

Schenone, P. (2020). Causality: A decision theoretic foundation. Technical report.

Schumacher, H. and Thysen, H. C. (2020). Equilibrium contracts and boundedly rational expectations.

Shermer, M. (1998). *Why people believe weird things: pseudoscience, superstition, and other confusions of our time*. Freeman & Co.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics*, 131(3):1243–1290.

Spiegler, R. (2017). “data monkeys”: a procedural model of extrapolation from partial statistics. *The Review of Economic Studies*, 84(4):1818–1841.

Spiegler, R. (2020). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association*, 18(2):583–617.

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., and Ellison, G. T. H. (2020). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, 50(2):620–632.

Verma, T. S. and Pearl, J. (1991). Equivalence and synthesis of causal models. Technical report.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.