SUBJECTIVE CAUSALITY IN CHOICE*

ANDREW ELLIS† AND HEIDI CHRISTINA THYSEN§

Abstract. When making a decision based on observational data, a person's choice depends on her beliefs about which correlations reflect causality and which do not. We model an agent who predicts the outcome of each available action from observational data using a subjective causal model represented by a directed acyclic graph (DAG). An analyst can identify the agent's DAG from her random choice rule. Her choices reveal the chains of causal reasoning that she undertakes and the confounding variables she adjusts for, and these objects pin down her model. When her choices determine the data available, her behavior affects her inferences, which in turn affect her choices. We provide necessary and sufficient conditions for testing whether such an agent's behavior is compatible with the model.

1. Introduction

When making a decision based on observational data, a person's choice depends on her beliefs about which correlations reflect causality and which do not. While correlation is unambiguous, causality is inherently subjective. For instance, a positive correlation between duration of hospitalization and risk of death may be causal if time in the hospital increases one's risk of catching an unrelated infection, or spurious if both are caused by the severity of illness. The agent's beliefs about causality affect her behavior: she would be more reluctant to seek treatment if she believed the former rather than the latter. In this paper, we develop a theory in which an analyst can

Date: December 2022.

* 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 Arizona, BRIC, BU, DC Area Theory Workshop, Edinburgh, Israeli and Hong Kong Theory Seminars, Harvard/MIT, Michigan, LSE, RUDC 2021, St Andrews, UCLA, and UCSD for helpful comments and discussions. Thysen gratefully acknowledges financial support by the ERC Advanced Investigator grant no. 692995.

†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.

Table 1. A (Partial) Dataset

	healthy ($N_h = 36$)	diseased ($N_d = 44$)	Pr(healthy|.)
no plaque, no tangles	18	12	0.6
plaque, no tangles	7	3	0.7
no plaque, tangles	2	8	0.2
plaque, tangles	9	21	0.3

Table describes a simulated dataset with 80 total patients. Column 2 (3) shows the number of patients that are healthy (diseased) and have the characteristics in Column 1. Column 3 shows the fraction of healthy patients having the characteristics in Column 1.

use the agent's behavior to identify her subjective causal model and to test whether it is explained by misperceived causality.

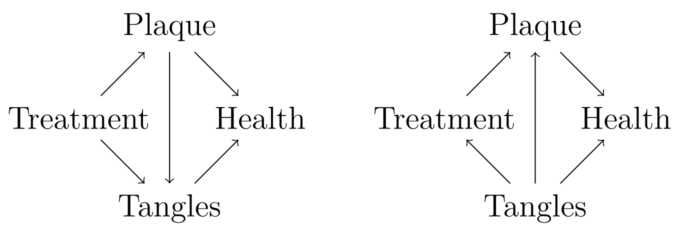
To illustrate, consider a doctor (the agent) prescribing medical treatments for Alzheimer's disease. These treatments can affect the patients' levels of two correlates, amyloid plaques and neurofibrillary tangles, in addition to Alzheimer's.[1] The doctor infers the effect of a drug using the dataset described in Table 1, as well as data demonstrating that it unambiguously decreases the chance of plaque. She prescribes the drug if she predicts that it leads to more healthy patients than doing nothing. Plaque is unconditionally correlated with disease, but the correlation reverses if one conditions on tangles. Therefore, the doctor's prediction depends crucially on how she interprets the data. If the doctor believes that tangles and health are endogenous consequences of plaque, then she predicts that the drug would be more effective than doing nothing. If she instead believes that tangles is an exogenous cause of plaque and health, then she predicts that the drug increases her patient's risk of disease. Our results show how a drug company (the analyst) can reveal the doctor's causal model from how frequently she performs each treatment (behavior) given different datasets.

Following Pearl (1995) and Spiegler (2016), we study a decision maker (DM) who makes predictions about her actions using a causal model described by a directed acyclic graph (DAG). Each node in the graph is a variable, such as the treatment and levels of plaque, tangles, and disease. 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 can represent a doctor who

[1]According to Wikipedia, Amyloid plaques are extracellular deposits of the amyloid beta protein, and neurofibrillary tangles are aggregates of hyperphosphorylated tau protein. These are the primary biomarkers of Alzheimer's disease.

believes either that tangles is an endogenous consequence of the drug (R in Figure 1) or that tangles may have affected the randomization into treatment and led to spurious correlation between the plaque and health (Q in Figure 1). A random choice rule records how often the DM takes each action after observing each dataset describing the joint distribution of variables and actions. The choice rule has a *subjective causality representation (SCR)* if she acts as if she uses a fixed DAG to predict the consequences of her actions, and then chooses the action with the highest predicted utility most frequently.

Figure 1. Two Subjective Causal Models



(a) R (b) Q

Our first main result reveals the DM's subjective causal model, the variable she cares about (the outcome), and her preferences from choices. Three features of the causal model determine her predictions. Most crucially, her predictions depend on the simplest ways in which she believes that her action can affect the outcome. These causal chains correspond to the smallest paths in the DAG from the node representing her action to that representing the outcome. For instance, the DAG R contains two shortest paths, one from treatment to plaque to health and another from treatment to tangles to health. This reflects Occam's razor: more complex explanations implied by simpler ones do not affect her predictions. Her predictions also depend on any *confounders*: exogenous variables, such as tangles for model Q, having a causal consequence in common with the action. The DM adjusts her predictions to account for spurious correlations that she thinks result from the confounders. Any two causal models with the same confounders and the same simplest causal chains from either the action or a confounder to the outcome lead to the same predictions. We show that the analyst can reveal these features by observing how the DM's choices change with the dataset used, just as the DM's choice between the drug and the status quo revealed whether her model is R or Q in the illustration. Our identification works regardless of and in complement to non-choice data about the DM's model, such as the timing of variables or pre-existing knowledge of perceptions of causality.

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 change, and thereby that policy's effectiveness, even if they are not empirically valid. For instance, a firm that appears reluctant to hire minority workers may dislike employing minorities even though they are equally or more productive (taste-based discrimination). Alternatively, its reluctance may be because minority-status 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 latter, such as affirmative action in university admissions or awarding scholarships to minority students, may do nothing for the former.[2]

We then extend our model to datasets generated by the agent's own (or others with the same model's) past behavior. Actions correspond to lotteries over the other variables, and she uses the dataset formed by combining the frequency with which she chooses each option and the resulting distribution over the other variables. The DM chooses the action with the highest predicted expected utility plus a (extreme-valued) shock. Her behavior may create a correlation between two variables that she misinterprets as a causal effect, feeding back into her predictions and thus her choices. While such a feedback effect occurs in many studies of agents with misspecified models, e.g. Esponda and Pouzo (2016), it has typically been absent in decision theoretic work on misspecification despite its potential policy implications. For example, a firm might incorrectly statistically discriminate against minority workers when it hires few of them but not when it hires more. We show in Section 4 that the feedback effect allows the model to accommodate a number of documented cognitive biases, including selection neglect, illusion of control, status-quo, and congruence biases.

To illustrate, consider a firm who thinks that worker productivity is a function of only education. If education is correlated with productivity for majority workers, but not for minorities, then the perceived return to education increases with the fraction of majority workers hired. This incentivizes the firm to hire the better-educated type more frequently, potentially reinforcing the effect. While this allows the model to accommodate the above biases, it leads to technical challenges for our analysis such as violations of (the stochastic choice analog of) independence of irrelevant alternatives, a necessary condition for a random utility model (RUM).

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

Our second main result establishes how to test whether a random choice rule over lotteries has an Endogenous SCR. We provide axioms that link the DM's subjective causal model, as inferred from our first result, to her behavior. Holding her predictions about the outcome of actions 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 her inferences about causal effects 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 effect on the outcome.

A DM with an Endogenous SCR may predict a different distribution of outcomes from taking an action than the analyst does. The result places testable restrictions on her behavior in spite of the information gap. Thus, it establishes that subjective 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 et al. (2020), Spiegler (2020), and Schumacher and Thysen (2022), which all use versions of the subjective causality representation.[3]

Related literature. Pearl (1995) argued for using and analyzing DAGs to understand causality. A large literature (e.g., Cowell et al., 1999, Pearl, 2009) develops and applies this approach for probabilistic and causal inference. Applied researchers use DAGs to estimate the causal effect of interventions from observational datasets, e.g. Tennant et al. (2020).[4]

Spiegler (2016) first modeled misspecified causal reasoning using DAGs, albeit without axiomatic foundations. He illustrates that his model has the power to capture a number of errors in reasoning, including reversed causality and omitted variables.

[3]Other models where misspecification leads to distorted beliefs include Esponda and Pouzo (2016), Bohren and Hauser (2021), Frick et al. (2020), He (2022), Heidhues et al. (2018), Samuelson and Mailath (2020), Montiel Olea et al. (2021), and Levy et al. (2022).
[4] Imbens (2020) contrasts this with the potential outcomes approach and discusses why these methods have attracted more attention outside of economics than within it.

Taken together, our results allow one to test the underlying assumptions of existing work studying the effects of causal misperception. This growing literature has been applied to monetary policy (Spiegler, 2020), political competition (Eliaz and Spiegler, 2020, Eliaz et al., 2022), communication (Eliaz et al., 2021), inference (Eliaz et al., 2020), and contracting (Schumacher and Thysen, 2022). In both these papers and the present one, the DM's behavior results from using a DAG and observational data to predict the outcome of her actions. Consequently, our results identifying the DAG from behavior increase their applicability.

We follow Spiegler in focusing on how a DM with a subjective causal model interprets data and on the feedback between her predictions and behavior. In contrast, Schenone (2020) takes a more normative approach to causality. He considers a DM who expresses preferences over act-causal-intervention pairs, and provides necessary and sufficient conditions for her beliefs to result from applying the do-operator to intervened variables for a fixed DAG and fixed prior. His analysis complements ours, showing what is possible given rich variation in exogenous interventions and mappings from variables to payoffs as opposed to rich variation in the data provided about the available actions. In contrast to both these approaches, Alexander and Gilboa (2019) model perceived causal relationships as reducing Kolmogorov complexity but do not explicitly relate their notion to choice.

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 (2022) consider an agent who categorizes alternatives based on the context, and the alternative's categorization affects her evaluation (or perception) thereof. Cerreia Viogolio et al. (2021) analyze a DM who considers several misspecified models and makes decisions that take her lack of confidence in the models into account. 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. (2022).

2. Model

2.1. **Primitives.** A DM chooses an action after observing a dataset q. A set $\mathcal{X}_0$ having at least two elements denotes the set of possible actions. Taking an action affects the distribution of a random vector $X = (X_1, \ldots, X_n)$ taking values in $\prod_{i=1}^n \mathcal{X}_i \equiv \mathcal{X}_{-0}$, where $\mathcal{X}_i$ is a finite subset of $\mathbb{R}$ with at least two distinct elements for each $i > 0$. The dataset q informing her choice describes past (joint) realizations of X and actions, with $q(a, x_1, \ldots, x_n)$ representing frequency of observations of action a and $X_i = x_i$ for every i. Formally, q is a distribution over $\mathcal{X} \equiv \mathcal{X}_0 \times \mathcal{X}_{-0}$ assigning positive probability to a finite set S_q of actions and where, to avoid issues of updating on zero probability events, $q(a, x) > 0$ for all $x \in \mathcal{X}_{-0}$ and $a \in S_q$; let $\mathcal{Q}$ be the set of such datasets. We also make the simplifying assumption that the DM's choice set equals S_q.[5] An important special case, formally studied in Section 4, is when the dataset q is derived endogenously from the DM's behavior.

The analyst observes a random choice rule $\rho : \mathcal{X}_0 \times \mathcal{Q} \to [0, 1]$ where $\sum_{a \in S_q} \rho(a, q) = 1$ and $\rho(a, q) = 0$ for every $a \notin S_q$. The choice rule describes the DM's behavior. The probability she chooses action a given dataset q is $\rho(a, q)$.

We adopt the following notational conventions. For a set B, $\Delta(B)$ denotes the set of finite-support probability measures on it. For a set $J \subset N \equiv \{0, 1, \ldots, n\}$, $\mathcal{X}_J = \prod_{j \in J} \mathcal{X}_j$, x_J denotes the event $\{y \in \mathcal{X} : y_j = x_j \text{ for all } j \in J\}$ for any $x \in \mathcal{X}_{J'}$ with $J \subseteq J'$, and $\text{marg}_J\, p$ denotes the marginal distribution of p on $\mathcal{X}_J$ when $p \in \Delta(\mathcal{X}_{J'})$ for $J \subseteq J'$. With slight abuse of notation, we sometimes write x_j instead of

[5]This avoids the DM having to form predictions about actions not in the dataset. One can easily extend the model to choice from a known subset of S_q.

$x_{\{j\}}$ and $x_\emptyset$ for an arbitrary constant random variable. For $p \in \Delta(\mathcal{X})$ and disjoint sets $A, B \subset N$, write $X_A \perp_p X_B$ if $p(x_{A\cup B}) = p(x_A)p(x_B)$ for all $x \in \mathcal{X}$. Finally, for $p_1, p_2 \in \Delta(\mathcal{X}_i)$, say that p_1 (strictly) FOSD p_2 if $p_1((-\infty, y)) \leq p_2((-\infty, y))$ for all y (without equality for some y).

2.2. **Causal Graphs.** We model perception of causality using a directed acyclic graph (DAG) R over the set N. The DAG R is an acyclic binary relation, with iRj denoting $(i, j) \in R$ and indicating that X_i is a direct cause of X_j. Visually, R describes the set of directed edges in a graph with node set N.

The following terminology will be useful. The *parents* of i, denoted $R(i)$, are the variables that directly cause X_i. The node j is an *ancestor* of i, denoted $j \in A_R(i)$, if there are $i_1, \dots, i_m \in N$ so that $jRi_1Ri_2R \dots Ri_mRi$, and j is a *descendant* of i if $i \in A_R(j)$. The tuple (i, j, k) is an *R-v-collider* if iRk, jRk, $j \not\!R i$, and $i \not\!R j$.

We follow Pearl (1995) in assuming that if one intervenes to force X_0 to equal x_0 without observing the realization of any other variables, the DAG R predicts that the random vector X equals $x \in \mathcal{X}_{-0}$ with probability

$$q_R(x \mid do(x_0)) = \prod_{j=1}^{n} q\left(x_j \mid x_{R(j)}\right) \tag{1}$$

given the dataset q. The intervention only affects the distribution of variables caused by it, which in turn only affect variables caused by them. Equation (1) makes X_i is independent of X_j conditional on $X_{R(i)}$ whenever j is not a descendant of i. In particular, the intervention does not change the distribution of its parents, even if they are correlated with it. Given the DAGs in Figure 1 and a dataset q, $q_Q(x_H|do(x_0))$ adjusts for the potential confounding effect of tangles, whereas $q_R(x_H|do(x_0))$ does not.

2.3. **Subjective Causality Representation.** We consider a DM who cares about the realization of only one of the variables, referred to as the *outcome* and labeled n^*. The other dimensions correspond to *covariates* that aid the DM in her inferences about the distribution of the outcome from taking an action. The DM has a fixed subjective causal model given by the DAG R. The DAG describes her beliefs about the causal relationships between variables. It does not impose any restrictions on the sign or the magnitude of the effects. She uses her dataset q and R to predict the outcome distribution of each action. That is, she calculates the effect of the intervention "take action a" according to Equation (1). Then, she chooses the action with the largest predicted expected utility.

Our representation restricts attention to a nontrivial and perfect DAG. A DAG R is *nontrivial* if $0 \in A_R(n^*)$ and *perfect* if there are no R-v-colliders. Neither places any restrictions on the structure of q but affects $q_R(\cdot|do(a))$. If R is trivial, then $q_R(x_{n^*}|do(a))$ is independent of a for every $x \in \mathcal{X}$ and $a \in S_q$. A perfect DAG implies that the DM does not neglect the correlation between two variables used to predict the realization of another variable. We discuss perfection further in Section 2.5.

Definition 1. Choice rule ρ has a *(perfect) subjective causality representation (SCR)* if there exists a (perfect,) nontrivial DAG R, index $n^* \in N\backslash\{0\}$, and strictly increasing u so that $U_q(a) \geq U_q(b)$ for all $b \in S_q \iff \rho(a,q) \geq \rho(b,q)$ for all $b \in S_q$, where

$$U_q(c) = \sum_{x \in \mathcal{X}_{-0}} u(x_{n^*}) q_R(x \mid do(c)).$$

Then, we say that (R, u, n^*) represents ρ and that ρ has an SCR (R, u, n^*).

The DM's most frequent choice, or choices, maximizes expected utility but with a potentially incorrect prediction about the outcome distribution. She expects to receive outcome x_{n^*} with probability $q_R(x_{n^*}|do(a))$ if she takes action a.

2.4. **Running Example.** Throughout, we illustrate using the following example. The random choice rule represents the frequency with which a doctor performs medical treatments for Alzheimer's disease on observationally identical patients. Each treatment relates to three patient characteristics (variables): plaque build-up (indexed by $P = 1$), tangles (indexed by $T = 2$), and long-term health status (indexed by $H = 3$). The doctor only cares about her patients' health. When choosing the treatment, she does not know the characteristic of her patient, but she has access to a dataset q that contains the joint distributions of other patients' treatments and characteristics. The vector (a, x_P, y_T, z_H) represents a patient who received treatment a, has plaque level x_P, tangle levels y_T and health status z_H, and $q(a, x_P, y_T, z_H)$ is the frequency with which a patient with these characteristics and that treatment appears in the dataset.

FIGURE 3. Possible DAGs in the Running Example

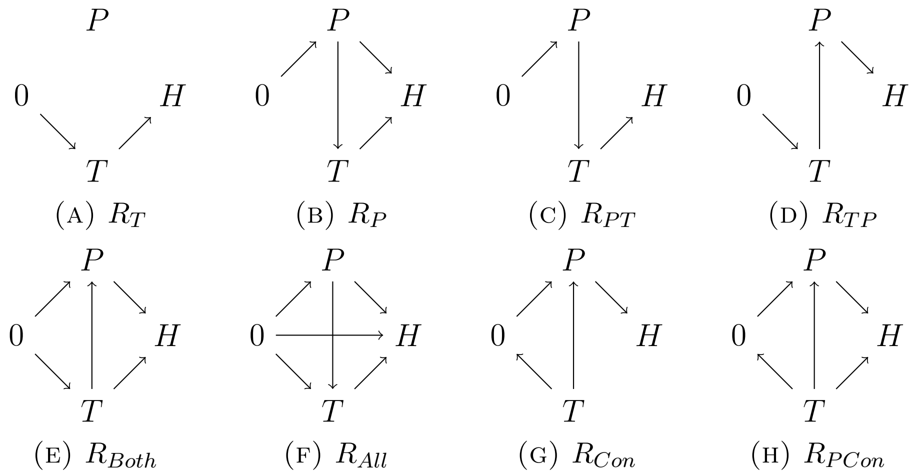


Figure 3 gives some possible DAGs, each representing a different theory of causation.[6] A doctor represented by R_T believes that the treatment directly influences tangles, and that tangles, and tangles alone, causes Alzheimer's. By contrast, one represented by R_{Both} believes that both tangles and plaque cause disease, that both the treatment and tangles cause plaque build-up, and that only the treatment causes tangles. The DAG R_{PCon} differs from R_{Both} only in that tangles affects the choice of treatment rather than vice versa. Both R_{Both} and R_{PCon} explain any dataset equally well, but they may lead to opposite predictions about which action is better. While they agree that the treatment does not directly affect health and that plaque buildup depends on both the treatment and tangles, they disagree on whether tangles influenced past treatment choices or past actions affected tangles. This may occur due to a disagreement about when tangles is measured.

Consider the dataset q described by Table 1 and where $q(\text{no plaque}|a) = q(\text{plaque}|b) \approx 1$ so that $q(\text{tangles}|a) \approx q(\text{tangles}|\text{no plaque})$, $q(\text{tangles}|b) \approx q(\text{tangles}|\text{plaque})$, and plaque is independent of tangles given treatment. Let 1 represent healthy, no plaque, and no tangles, and 0 represent diseased, plaque, and tangles. If the DM is represented

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

by R_{Both}, then she takes action a since

$$\begin{aligned} q_{R_{Both}}(\text{healthy}|do(a)) &= \sum_{z=0,1} q(z_T|a)\left[\sum_{y=0,1} q(y_P|z_T,a)q(1_H|y_P,z_T)\right] \\ &\approx \frac{3}{4} * (1 * .6 + 0 * 0.7) + \frac{1}{4} * (1 * 0.2 + 0 * 0.3) = 0.5 \\ &> 0.4 = \frac{1}{4} * 0.7 + \frac{3}{4} * 0.3 \approx q_{R_{Both}}(\text{healthy}|do(b)). \end{aligned}$$

If she is instead represented by R_{PCon}, then she takes action b since

$$\begin{aligned} q_{R_{PCon}}(\text{healthy}|do(a)) &= \sum_{z=0,1} q(z_T)\left[\sum_{y=0,1} q(y_P|z_T,a)q(1_H|y_P,z_T)\right] \\ &\approx \frac{1}{2} * (1 * 0.6 + 0 * 0.7) + \frac{1}{2} * (1 * 0.2 + 0 * 0.3) = 0.4 \\ &< 0.5 = \frac{1}{2} * 0.7 + \frac{1}{2} * 0.3 \approx q_{R_{PCon}}(\text{healthy}|do(b)). \end{aligned}$$

The expression $q_{R_{PCon}}(\text{healthy}|do(a))$ treats tangles as an exogenous confounder affecting both the treatment and plaque buildup, but $q_{R_{Both}}(\text{healthy}|do(a))$ treats it as an endogenous consequence of the action.

2.5. **Remarks on the Model.** The agent is dogmatic about her model R, in that she does not consider switching models because of her dataset. This conforms with evidence from psychology that, as summarized by Sloman (2005, p. 107), "causal explanations quickly become independent of the data from which they are derived." We view our setting as one in which the DM has already formulated her model, perhaps from past experience or earlier data, and does not deviate from it on the basis of the present dataset.

We focus on perfect DAGs for philosophical and psychological reasons. Perfection limits how wrong the DM's predictions can be, as it ensures that the DM correctly predicts the marginal distribution of individual variables. Psychology experiments (Lombrozo, 2007, Pacer and Lombrozo, 2017) indicate a preference for stories with fewer unexplained variables, and a perfect DAG has exactly one unexplained variable. Moreover, perfect DAGs are widely applied. Most of the literature following Spiegler (2016) uses perfect subjective DAGs. The first step of many algorithms using DAGs is to make the DAG perfect if it is not already (e.g. Chapter 6 of Cowell et al. (1999)). Natural procedures, such as extrapolating from multiple datasets by filling in missing correlations to maximize entropy, can be represented as coming from a perfect DAG (Spiegler, 2017).

The ideas behind our identification results apply to imperfect DAGs as well. V-colliders introduce complications similar to confounding variables (defined formally in the next section). Confounders can be treated as v-colliders with the action. However, any v-collider ancestral to the outcome affects the DM's prediction, even if it is exogenous and far removed from the variables that she thinks her action can affect. While tedious, these can be identified from the DM's behavior in a similar manner.

Our identification results only require that the DM's most frequent choices maximizes perceived expected utility. This allows application to several different choice procedures. Of particular interest are the following three. First, $\rho(\cdot, q)$ equals the uniform distribution on $\arg\max_{S_q} U_q(\cdot)$. This nests deterministic choice with $c(S_q) = \text{supp}\rho(\cdot, q)$. Second, $\rho(a, q)$ equals $\Pr(\{\varepsilon : U_q(a) + \varepsilon_a \geq U_q(b) + \varepsilon_b \ \forall b \in S_q\})$ where $\{\varepsilon_i\}_{i \in S_q}$ are independent and identically distributed. This includes Logit and Probit for particular distributions of ε_i. Our endogenous dataset results focus on Logit so that all actions are chosen with positive probability. Third, $\rho(a, q) > \frac{1}{2}$ if and only if $U_q(a) > U_q(b)$ when $X_0 = \{a, b\}$.

We assume that that the agent cares about influencing a single random variable. Our results immediately extend to multiple outcomes provided that they are all directly causally related to one another. Otherwise, the paths between the outcomes may also matter. We leave this as an open question for future work.

3. Revealing the Subjective Causal Model

We show that three features of the DAG R affect her predictions: its minimal active paths (MAPs) from the action to the outcome, confounders, and MAPs from confounders to the outcome.

A confounder is a variable that the DM thinks may affect the relationship between her action and another variable. The node i^* is an *R-confounder* if $i^* R 0$, $0 R l$ and $i^* R l$ for some $l \in A_R(n^*) \cup \{n^*\}$. Tangles is a R_{PCon}-confounder but not an R_{Both}-confounder.

A directed path in the DAG captures a chain of causal reasoning, and a MAP is a chain that cannot be made shorter by omitting variables. The finite sequence $(i_0, \ldots, i_m)$ of nodes is an *R-MAP from i_0 to i_m* if $i_j R i_k$ if and only if $k = j+1$ and i_j is an R-confounder or 0 only if $j = 0$. The path $(0, P, H)$ is both an R_{Both}-MAP and an R_{PCon}-MAP, $(0, T, H)$ is an R_{Both}-MAP but not an R_{PCon}-MAP, and $(0, T, P, H)$ is neither an R_{Both}-MAP nor an R_{PCon}-MAP.

Theorem 1. *Let ρ have a perfect SCR (R, u, n^*) and R' be a perfect DAG. Then, ρ has an SCR (R', u, n'^*) if and only if $n^* = n'^*$, R and R' have the same confounders, and R and R' have the same MAPs from a confounder or 0 to n^*.*

Theorem 1 shows that the identity of the outcome, the variables that may have affected the relationship between the action and another variable, and the simplest causal mechanisms determine her predictions. As in the running example, the DM adjusts for confounders when forming her prediction, and different confounders, or paths from confounders to the outcome, lead to different adjustments and thus different behavior. Paths from her action to the outcome represent the direct and indirect ways that she thinks her action could affect her payoff. Different paths correspond to different mechanisms, and DMs who disagree on the mechanisms make different choices.

However, only disagreements about the MAPs matter. Because a MAP cannot be made shorter, it represents one of the simplest causal mechanisms. The result thus provides a version of Occam's razor: the simplest ways in which the DM thinks her action (or a confounder) can affect the outcome determine her choices. For instance, a DM represented by R_P behaves identically to one represented by $0 \to P \to H$ even though R_P contains the additional path $(0, P, T, H)$. The latter path captures a belief that the effect of plaque on health is partly due to its effect on tangles. This does not affect her estimate of plaque's total effect on health. Since she does not think that her action directly affects tangles, she makes the same prediction with either of the two DAGs. Mathematically, tangles' effect on health integrates out through the identity $\sum_{y_T \in \mathcal{X}_T} q(x_H | y_T, x_P) q(y_T | x_P) = q(x_H | x_P)$.

The result has two immediate implications that allow us to simplify the set of DAGs considered. First, only relationships between variables that appear in at least one R-MAP affect the DM's behavior. In particular, any variables that she thinks are caused by the outcome are inconsequential for her behavior. Second, the chains of causality involving such variables determine all other causal relationships. While there may be edges not in an R-MAP, their existence and direction can either be determined from the R-MAPs or are immaterial to the DM's choices.

Remark 1. If ρ has SCRs (R, u, n^*) and (R', u', n^*), then there exists $\alpha > 0$ and β so that $u(z) = \alpha u'(z) + \beta$ for every $z \in \mathcal{X}_{n^*}$. The exact form of the converse depends on how much structure one puts on the random choice rule. For instance, the converse holds if $\rho(\cdot, q)$ is uniformly distributed on $\arg\max_{S_q} U_q(\cdot)$ and holds with $\alpha = 1$ if $\rho(\cdot)$ is Logit or Probit.

3.1. **Intuition without confounders.** We first provide an intuition for the result in the case with no confounders. Then, the result says that two DMs with SCRs behave identically if and only if they have the same tastes, care about the same outcome n^*, and their DAGs have the same MAPs from 0 to n^*.

Sufficiency is illustrated above for the case of R_P and $0 \to P \to H$. The proof extends this logic using a tool from the Bayesian networks literature known as a junction tree (e.g. Cowell et al. (1999)) to transform a perfect DAG with many paths to one with a single path but where nodes are sets of variables. We show an equivalence between the path in the junction tree to minimal active paths in the underlying DAG. Then, the logic above applies: only variables in the path between the action and the outcome matter.

We now show how one can identify the R-MAPs from behavior. Observe that when a variable is independent of its parents, the DM estimates their causal effect on it to be zero. By making certain sets of variables independent of all others, the analyst can tease out her DAG using this observation.

Definition 2. A subset $A \subseteq N$ of variables is a *ρ-separator* if $\rho(a, q) = \rho(b, q)$ for every q such that $X_A \perp_q X_{A^c}$ and all $a, b \in S_q$.

A set A of variables is a ρ-separator if the DM is indifferent between her available actions whenever X_A is independent of the other variables in her data. That is, she predicts that there is no relationship between her action and the outcome whenever there is no relationship between the variables indexed by A and the others. When using this definition or Definition 3, we omit "ρ-" when the choice rule is clear from context.

Lemma 1. *If ρ has an SCR (R, u, n^*), then $A \subseteq N$ is a ρ-separator if and only if every R-MAP from 0 to n^* intersects A.*

To illustrate, consider a doctor who is equally likely to prescribe every treatment whenever plaque build-up is independent of the other variables, i.e. $\{P\}$ separates. When one action is correlated with better health and tangles and the other is not, she is nevertheless indifferent, revealing that she does not take these correlations into account. Therefore, every causal chain in her DAG includes plaque.

In some cases, we can recover the R-MAPs from the separators alone. For example, $(0, n^*)$ is the only R-MAP if and only if every separator contains 0 or n^*. In general, one needs to know more than the separators to recover R. For instance, every non-empty subset of $\{0, H, T, P\}$ is a ρ-separator if ρ is represented by either R_{TP} or R_{PT}.

When the smallest (by set inclusion) separators are mutually disjoint, then every R-MAP from the action to outcome contains exactly one index from each of them. Intuitively, the path must contain at least one from every minimal separator by Lemma 1, and if it contained more from a given separator, then the path could be made shorter by omitting one of them. In general, the minimal separators are not disjoint. Then, the variables that make up an R-MAP correspond to a *selection from the ρ-separators*, a $\subseteq$-minimal set $J \subseteq N$ so that J intersects every ρ-separator. Similar logic applies. The path must contain at least one from every minimal separator but cannot contain too many from any of them. Note that $\{0\}$ and $\{n^*\}$ are always separators, so every selection from the separators includes both 0 and n^*.

Definition 3. For a selection J from the ρ-separators, $i \in J$ is adjacent to $j \in J$ if $X_J \perp_q X_{J^c}$ and $X_i \perp_q X_j$ implies $\rho(a, S_q) = \rho(b, S_q)$ for $a, b \in S_q$.

When $X_J \perp_q X_{J^c}$, the only variables relevant for the DM's choice given q are those indexed by J. The DM believes that every variable in the selection (except her action) is caused by exactly one of the others. If she thinks that X_i causes X_j (or vice versa) and $X_i \perp_q X_j$, then she estimates its causal effect to be zero and thus predicts that her action has no direct or indirect effect the outcome. Therefore when i is adjacent j for a selection, the DM believes that either X_i causes X_j or X_j causes X_i.

Lemma 2. *If ρ has a perfect SCR (R, u, n^*), then $(i_0, \ldots, i_m)$ is an R-MAP from* 0 *to n^* if and only if $\{i_0, \ldots, i_m\}$ is a selection from the ρ-separators, $i_0 = 0$, and i_j is ρ-adjacent to i_{j+1} for all j.*

The lemma characterizes the R-MAPs via adjacency, and also identifies n^* as the element of $\{1, \ldots, n\}$ adjacent to exactly one variable in every selection. Consider a doctor represented by R_{PT}. There is a unique selection from the separators, $J = \{0, P, T, H\}$. If tangles are independent of health, then the doctor thinks that each treatment is equally effective since she believes that tangles are the only cause of Alzheimer's. She is equally likely to prescribe any treatment, regardless of the true relationship between her action and health. Thus, we conclude that T is adjacent to H. On the contrary, the doctor may think one drug is superior even when plaque is independent of health. For instance, if tangles occur if and only if health is bad and there is no plaque build-up, then tangles are negatively correlated with both plaque and health even when they are independent of each other. Therefore, the doctor thinks that the drug leading to a lower chance of plaque build-up is best, and we conclude that P is not adjacent to H. By similar logic, P is adjacent to T, and 0 is adjacent to P.

3.2. **Intuition with confounders.** When the DM believes there are confounding variables, all representations must in addition have the same confounders and the same minimal causal chains from each confounder to the outcome. For instance, a doctor represented by R_P makes different choices than one represented by R_{PCon} even though $(0, P, H)$ is the only R_P- or R_{PCon}-MAP from 0 to H. The former has no confounders, and T is an R_{PCon}-confounder.[7] Moreover, doctors represented by R_{Con} and R_{PCon} also behave differently because (T, H) is an R_{PCon}-MAP but not an R_{Con}-MAP.

To identify the presence of confounders in the DM's causal model, we observe that in their absence only the distribution over variables involved in a causal chain from 0 to n^* affects the DM's choices. If the action is independent of all variables involved in any causal chain from 0 to n^* and the DM is not indifferent between all of her actions, then her causal model contains at least one confounder.[8] Formally, if i^* is not part of any selection from the ρ-separators, then i^* is a *revealed confounder*

[7]For the dataset q and $c = a, b$ in Section 2.4, $q_{R_P}(1_H|do(c)) = q_{R_{Both}}(1_H|do(c))$.
[8]Lemma 2 still applies, so we can identify the R-MAPs from 0 to n^* using it.

if the DM expresses a strict preference between two action for some dataset where every variable except i^* is independent of her action. We show that the revealed confounders must be confounders in the DM's DAG. We then extend the procedure above to identify the MAPs from each revealed confounder to the outcome.

3.3. **Finite observations.** The DM's DAG can be revealed using choices from a finite number of carefully constructed datasets. Given a set A, we construct a dataset so that DM's choices when facing it reveal whether or not A separates. As there are a finite number of variables, we can reveal all the separators from a finite number of datasets. Then, given a selection from the separators J and two variables $i, j \in J$, we construct another dataset that reveals whether or not i is adjacent to j.

To state the result, let $\bar{x}$ and $\underline{x}$ be such that $\bar{x}_i = \max \mathcal{X}_i$ and $\underline{x}_i = \min \mathcal{X}_i$. A dataset q has $\{i, j\}$-MLRP if $S_q = \{a, b\}$ and $\frac{q(x_i|y_j)}{q(x_i|y'_j)} \geq \frac{q(x'_i|y_j)}{q(x'_i|y'_j)}$ when $x > x'$ and $y' > y$, without equality for at least one (x, x', y, y') and where "$a > b$" is taken to be true.

Proposition 1. *Let ρ have a perfect, unconfounded SCR (R, u, n^*) and $S_q = \{a, b\}$.*

(i) If $0 \notin A$, $X_A \perp_q X_{A^c}$, and $q(\bar{x}_{A^c}|a), q(\underline{x}_{A^c}|b) > \delta$, then $\rho(a, q) = \rho(b, q)$ if and only if A ρ-separates, where $\delta = \frac{2^{-1/n}}{q(a)+(1-q(a))2^{-1/n}}$.

(ii) For a selection I from the separators and distinct $i, j \in I$, if q has $\{i', j'\}$-MLRP for all distinct $i', j' \in I$ s.t. $\{i', j'\} \neq \{i, j\}$, $X_I \perp_q X_{I^c}$, and $X_i \perp_q X_j$, then $\rho(a, q) = \rho(b, q)$ if and only if i is ρ-adjacent to j.

If a set A separates, then the DM chooses every action in her choice set with equal probability. Otherwise, there is an R-MAP between the action and outcome that does not intersect A. When the highest and lowest realizations are sufficiently strongly correlated with each other outside of A, the DM predicts that action a leads to a better outcome than does b.

If the two variables are adjacent, then $X_I \perp_q X_{I^c}$ and $X_i \perp_q X_j$ imply that the DM chooses both actions with equal probability. If not, then under the MLRP condition, every pair of variables except i and j is positively correlated with each other. As a consequence, the DM infers that all the causal effects between variables in I are positive. Therefore, she predicts that a leads to a strictly better outcome than does b whenever i does not cause j or vice versa.

A similar approach can be used when the DAG contains confounders. However, one needs additional structure on the datasets to ensure that the perceived correlation remains even after controlling for the perceived endogeneity.

4. Endogenous Datasets

The next two sections study the special case where the dataset q is derived endogenously from the DM's behavior. This section generalizes the model from Section 2 to accommodate an endogenous dataset. The next characterizes the random choice rules that have a subjective causality representation.

Consider a situation where the dataset utilized by the DM to make her predictions is generated by her own behavior as in Spiegler (2016). To model this, we take as given a function ζ that maps each action to the full-support lottery over $\mathcal{X}_{-0}$ that results from taking the action. The mapping ζ should be interpreted as known to the analyst but not the DM, who leverages her causal model to learn it. We assume that ζ and $\mathcal{X}_0$ are rich enough that there is at least one action corresponding to every such lottery, and to economize on notation, we write $a(x)$ instead of $\zeta(a)(x)$.

Let $\mathcal{S}$ be the set of finite subsets of $\mathcal{X}_0$. Then, 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 DM's dataset when facing menu S is ρ^S where

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

The dataset ρ^S combines the lottery that results from taking action a with the frequency that a is chosen, $\rho(a,S)$. Note that $\rho(\cdot,S)$ is a distribution over actions whereas ρ^S is a distribution over $\mathcal{X}$.

Definition 4. Choice rule ρ has an *Endogenous Subjective Causality Representation* if there exists a perfect, nontrivial, unconfounded DAG R, an index n^*, and a strictly increasing $u : \mathcal{X}_{n^*} \to \mathbb{R}$ so that

$$\rho(a,S) = \frac{\exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_{n^*}) \rho^S_R(x|do(a))\right)}{\sum_{a' \in S} \exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_{n^*}) \rho^S_R(x|do(a'))\right)}$$

for every $S \in \mathcal{S}$ and $a \in S$. Then, we say that ρ has an Endogenous SCR (R, u, n^*).

As before, the DM uses her model R to form predictions about the outcomes of each potential action. However, the dataset ρ^S that she uses to estimate each of the causal effects varies with how often she chooses each of the actions. Note that $\rho(a, S) = \Pr(\{\varepsilon : U_{\rho^S}(a) + \varepsilon_a \geq U_{\rho^S}(b) + \varepsilon_b \text{ for all } b \in S\})$ when $\{\varepsilon_b\}_{b \in S}$ are distributed independently and extreme value. Moreover, the representation is a personal equilibrium (Köszegi and Rabin, 2006): the DM maximizes expected utility given her beliefs, which in turn 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 equilibrium, we place no restrictions on which is selected.

Remark 2. Random choice plays two roles. First, $\rho(a, S) > 0$ for all $a \in S$, so predictions about every action are well-defined. Second, Spiegler (2016) shows that "pure" equilibria may not exist, even after defining beliefs about unchosen actions.

Remark 3. We adopt the exponential function for concreteness and applicability. Our results adapt to any other strictly increasing and positive function.

Remark 4. A confounded DAG would represent a DM who believes that a variable whose realization she does not observe affects her choices.

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, the DM may violate *regularity*, a property necessary for representation by a random utility model (RUM). Second, her behavior may be *self-confirming*: because she chooses a frequently, she thinks it is better than b, but she would reverse her ranking if she chose b more frequently. 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. We illustrate these challenges using a doctor in the running example whose behavior has an Endogenous SCR (R_P, u, H) and where $\mathcal{X}_i = \{0, 1\}$ for $i = P, T, H$.

4.1. **Regularity violation.** An Endogenous 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. By contrast, many models

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

of behavior interpreted as irrational are RUMs, such as the model of limited attention due to Manzini and Mariotti (2014).

To illustrate why regularity may be violated, consider three treatment options, ι, π, ν, that are equally likely to lead to good health. Plaque and health are independent after undergoing treatment ι, positively correlated under treatment π, and negatively correlated under treatment ν. When the doctor decides between only ι and π, 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 ν may cancel out or even reverse the perceived positive effect of plaque on health. When this effect is strong enough, she chooses the treatment with a lower probability of plaque buildup more frequently. Formally, suppose that $\iota(1_P, 1_H) = \iota(1_P, 0_H) = \frac{1}{2}$, $\pi(0_P, 0_H) = \pi(1_P, 1_H) = \frac{1}{2}$, $\nu(1_P, 0_H) = \nu(0_P, 1_H) = \frac{1}{2}$, $u(0) = 0$, and $u(1) = 6$.[10] One can verify that for $S = \{\iota, \pi\} \subset S' = \{\iota, \nu, \pi\}$, we have $\rho(\pi, S) < \frac{1}{3} = \rho(\pi, S')$, a violation of regularity.[11,12]

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 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 ι leads to better health, where none exist.

4.2. **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 predict that one action is better than another only if she chooses it sufficiently frequently. This can lead to multiple personal equilibria. Suppose that treatment b prevents the disease but often leads to plaque buildup, and that treatment a leads

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

[11]See Appendix B.1 for the derivation.

[12]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 $y \in \{0,1\}^3$.

to frequent disease but rare plaque buildup that only occurs when the patient is sick. When the doctor rarely prescribes treatment b, low plaque buildup is mistakenly believed to increase the chance of good health. Consequently, prescribing b, which raises their plaque, seems like a bad idea. Symmetrically, when she usually prescribes b, she predicts that a leads to a lower chance of good health because plaque buildup is negatively correlated with disease. Formally, let $b(1_P, 1_H) = 1$, $a(0_P, 1_H) = \frac{1}{2}q$, $a(0_P, 0_H) = \frac{1}{2}(1-q)$, and $a(1_P, 0_H) = \frac{1}{2}$ where $q \in (0,1)$. When $u(0) = 0$ and $u(1) = 30$, one can verify that $\rho(a, \{a,b\}) \approx .34$, $\rho(a, \{a,b\}) \approx 0.02$, and $\rho(a, \{a,b\}) \approx 0.99$ are all equilibria when $q = \frac{3}{4}$.[13]

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.

5. Behavioral Foundations

We now turn to the behavioral regularities that characterize the random choice rules with an Endogenous SCR. We first identify a candidate causal model for the DM. To this end, Section 5.1 adapts the techniques developed in Section 3 to the current setting. The axioms relate the predictions implied by the causal model to the DM's choices. Section 5.2 presents some familiar axioms. Section 5.3 first illustrates our approach by characterizing the choices of a doctor in the running example with model R_P. Section 5.4 presents the axiomatization for the general case.

5.1. **Identification.** The results in Section 3 generalize with minor modification. Applying the results requires constructing a menu S for which $X_A \perp_{\rho^S} X_B$. When $0 \notin A$, $X_A \perp_{\rho^S} X_B$ whenever $\mathrm{marg}_A\, a = \mathrm{marg}_A\, b$ for all $a, b \in S$ and $X_A \perp_a X_B$ for all $a \in S$. This allows easy application of Proposition 1 to reveal following from ρ.

Definition 5. The set of *minimal ρ-separators* is

$$\mathcal{A}^\rho = \{A \subseteq N : A \text{ is a minimal set that } \rho\text{-separates}\}.$$

We say that $A \in \mathcal{A}^\rho$ *is adjacent to* $B \in \mathcal{A}^\rho$, written $A \propto^\rho B$, if every $k \in A \setminus B$ is ρ-adjacent to every $l \in B \setminus A$ and $A \neq B$.

[13] See Appendix B.2 for the derivation.

Both the set $\mathcal{A}^\rho$ and the relation $\propto^\rho$ can be determined by observing ρ from a finite number of menus using Proposition 1. Together with the dataset, they determine the DM's predictions about her actions.

Proposition 2. *If ρ has an Endogenous SCR (R, u, n^*), then $\mathcal{A}^\rho$ can be ordered $\mathcal{A}^\rho = \{A_1^\rho, \ldots, A_m^\rho\}$ where $A_1^\rho = \{0\}$, $A_m^\rho = \{n^*\}$, and for every $i = 1, 2, \ldots, m$, A_i^ρ is ρ-adjacent to A_j^ρ if and only if $j \in \{i+1, i-1\}$. Moreover, for any dataset q*

$$q_R\left(x_{n^*}|do(x_0)\right) = \sum_{y \in \mathcal{X}_{\cup_{i=2}^{m-1} \mathcal{A}_i^\rho}} q\left(x_{n^*}|y_{A_{m-1}^\rho}\right) \prod_{i=2}^{m-2} q\left(y_{A_{i+1}^\rho}|y_{A_i^\rho}\right) q\left(y_{A_2^\rho}|x_0\right). \tag{2}$$

Proposition 2 reveals the DM's predictions from the minimal separators and adjacency. Consider $j \in A_{i+1}^\rho \backslash A_i^\rho$. The result shows that kRj for every $k \in A_i^\rho$ and that, by perfection, either jRk' or $k'Rj$ for every other $k' \in A_{i+1}^\rho \setminus A_i^\rho$. However, which of the two holds is immaterial for the DM's predictions. By focusing on the relationship between the variables in A_{i+1}^ρ and those in A_i^ρ, we sidestep any indeterminacy in her model.

Given Theorem 1, $\mathcal{A}^\rho$ and $\propto^\rho$ determine the R-MAPs for any R that represents ρ. This reveals the set of DAGs that could potentially represent it. Our axiomatization exploits this by identifying their implications for choice and requiring that the DM does not deviate from them.

5.2. **Basic axioms.** The first two axioms are standard, and we present them with minimal discussion. The first requires that the DM chooses every action with positive probability.

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

The second limits the perceived difference between any two options.

Axiom 2 (Bounded Misperception)**.** The quantity $\sup_{S, a, b \in S} \frac{\rho(a,S)}{\rho(b,S)}$ is finite.

The relative frequency with which the DM takes two actions, their *Luce ratio*, indicates the strength of her preference. Since the set of outcomes is finite, 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.

5.3. **Foundations - special case.** This subsection specializes the remaining axioms to a doctor represent by R_P. The first requires that R_P is a candidate for her DAG.

Axiom (CRC*)**.** The set of minimal separators $\mathcal{A}^\rho$ equals $\{\{0\}, \{P\}, \{H\}\}$. Moreover, $\{0\}$ and $\{H\}$ are both adjacent to $\{P\}$ and only to $\{P\}$.

Since $\{0\}, \{P\}, \{H\} \in \mathcal{A}^\rho$, every R-MAP of must include 0, P, and H. Further, as $\{P\}$ is adjacent to both $\{0\}$ and $\{H\}$, there is a single R-MAP $(0, P, H)$. Finally, since H is furthest from 0 in the R-MAP, this reveals that the doctor cares only about health.

Second, the doctor treats plaque as a sufficient statistic for health.

Axiom (I5*)**.** For $a, b \in S \in \mathcal{S}$, if $\operatorname{marg}_P a = \operatorname{marg}_P b$, then $\rho(a, S) = \rho(b, S)$.

According to R_P, any treatments that lead to the same chance of plaque buildup also have the same chance of good health. Therefore, if two treatments lead to the same marginal distribution over plaque, then the doctor is equally likely to choose either, regardless of any other difference between them.

Third, she chooses similarly when she makes similar predictions.

Axiom (LCI*)**.** For any $S, S_1, S_2, \dots \in \mathcal{S}$ with $a, b \in S_m \cap S$ for each m: if $\rho^{S_m}(x_H|x_P) \to \rho^S(x_H|x_P)$ for all $x \in \mathcal{X}_{-0}$, 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 makes the same predictions for a and b whether she faces S or S'. The doctor thinks that distribution of health conditional on plaque equals the causal effect of plaque on health. If this distribution is the same for ρ^{S_1} and ρ^S, then the doctor makes the same predictions about the effect of each treatment when facing either S_1 or S. Hence, she should consistently evaluate them, in the sense that Luce's choice axiom should hold for S_1 and S. Indeed, letting $S_m = S_1$ for all m, the axiom implies that if $\rho^{S_1}(x_H|x_P) = \rho^S(x_H|x_P)$ for all $x \in \mathcal{X}_{-0}$, then $\frac{\rho(a,S_1)}{\rho(b,S_1)} = \frac{\rho(a,S)}{\rho(b,S)}$. Moreover, the Luce ratio should be "almost" the same whenever these conditional probabilities are "close."

Finally, she behaves as if Logit-EU when her dataset is consistent with R_P.

Axiom (CPL*)**.** There exists a increasing $u : \mathcal{X}_H \to \mathbb{R}$ so that

$$\rho(a, S) = \frac{\exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_H) a(x)\right)}{\sum_{b \in S} \exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_H) b(x)\right)}$$

whenever $b(y_H, y_P) = b(y_P) a(y_H | y_P)$ for all $y \in \mathcal{X}_{-0}$ and $a, b \in S$.

When health is in fact independent of the treatment given plaque, the DM makes correct predictions. Hence, she should behave according to Logit-EU.

These axioms hold if and only if the doctor has an SCR with DAG R_P.

Corollary 1. *The choice rule ρ satisfies Full-support, Bounded Misperception, CRC*, I5*, LCI*, and CPL* if and only if ρ has an Endogenous SCR (R_P, u, H).*

The result is a corollary of Theorem 2, and so we defer discussion until then.

5.4. **Foundations - general case.** We now present the axioms for an arbitrary DAG. To simplify the statement in the general case, we take the outcome n^* to be known.

The third axiom ensures that we can find a candidate for the DM's DAG.

Axiom 3 (Consistent Revealed Causes, CRC)**.** The sets $\{0\}, \{n^*\} \in \mathcal{A}^\rho$. Moreover,

(i) for any distinct $A_1, \ldots, A_k \in \mathcal{A}^\rho$ with $k \geq 3$, if A_i is ρ-adjacent to A_{i+1} for all $i < k$, then A_k is not ρ-adjacent to A_1;
(ii) $|\{A' \in \mathcal{A}^\rho : A' \text{ is } \rho\text{-adjacent to } A\}| = 1$ for each $A \in \{\{0\}, \{n^*\}\}$; and
(iii) $|\{A' \in \mathcal{A}^\rho : A' \text{ is } \rho\text{-adjacent to } A\}| = 2$ for each $A \in \mathcal{A}^\rho \setminus \{\{0\}, \{n^*\}\}$.

First, both the action and the outcome separate. Second, there are no cycles of causes. Since we infer causality from adjacency, this requires that the adjacency relation has no cycles except those caused by symmetry. Finally, each separator is adjacent to the correct number of others according to Proposition 2. When $|\mathcal{A}^\rho| \leq 4$, condition (i) is redundant given (ii) and (iii).

Lemma 3. *Axiom 3 holds if and only if $\mathcal{A}^\rho$ can be ordered $\mathcal{A}^\rho = \{A_1^\rho, \ldots, A_m^\rho\}$ where $A_1^\rho = \{0\}$, $A_m^\rho = \{n^*\}$, and for every $i = 1, 2, \ldots, m$, A_i^ρ is ρ-adjacent to A_j^ρ if and only if $j \in \{i+1, i-1\}$.*

Given Proposition 2, Lemma 3 shows that we can infer what a DM with an Endogenoenous SCR would predict whenever ρ satisfies Axiom 3. This will be useful in that the remaining axioms relate the DM's inferences about causal effects to her choices.

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 $A \in \mathcal{A}^\rho$ is ρ-adjacent to $\{0\}$ and $\operatorname{marg}_A a = \operatorname{marg}_A b$, then $\rho(a, S) = \rho(b, S)$.

The covariates directly caused by the action are a sufficient statistic for the DM's prediction of the outcome distribution. Whenever two actions lead to the same distribution over these covariates, she predicts that they lead to the same outcome distribution. She is therefore indifferent between any two actions with identical immediate implications. The axiom only considers the marginal distributions on X_A of a and b, remaining agnostic about their distribution on other variables and any other available actions.

The fifth axiom ensures that if she makes similar predictions in two contexts, then she makes similar choices when facing them.

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

$$\rho^{S_m}\left(y_{A^\rho_{i+1}} | y_{A^\rho_i}\right) \to \rho^S\left(y_{A^\rho_{i+1}} | y_{A^\rho_i}\right)$$

for every $y \in \mathcal{X}_{-0}$ and $i = 2, \dots, |\mathcal{A}^\rho| - 1$, then $\frac{\rho(a,S_m)}{\rho(b,S_m)} \to \frac{\rho(a,S)}{\rho(b,S)}$.

LCI relates the DM's inferences about the causal effects to her choices. As per Equation (2), the axiom's hypothesis ensures that her predictions about the outcome of a and b from ρ^{S_m} are close to those from ρ^S for large m. If that is the case, then LCI requires that their Luce ratio from S_m is close to their Luce ratio from S. By taking $S_1 = S_m$ for all m, we see that LCI implies that if her inferences are the same, then their Luce ratio is the same.

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

Definition 6. A menu $S \in \mathcal{S}$ *leads to correct predictions* if

$$b\left(y_{\bigcup_{i=2}^{|\mathcal{A}|} A_i^\rho}\right) = b\left(y_{A_2^\rho}\right) \prod_{i=2}^{|\mathcal{A}^\rho|-1} a\left(y_{A_{i+1}^\rho} | y_{A_i^\rho}\right),$$

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

From Equation (2), we see that only the set of variables indexed by $\bigcup_{i=1}^{|\mathcal{A}^\rho|} A_i^\rho$ are relevant for the DM's predictions, and that the DM thinks that any variable in A_{i+1}^ρ is independent of any in $\bigcup_{k=1}^{i-1} A_k^\rho$ conditional on those indexed by A_i^ρ. A correctly perceived menu satisfies these conditional independences, so her prediction of the distribution of all relevant variables, including the outcome, is correct.

The DM behaves in a standard fashion when her predictions are correct.

Axiom 6 (Correct Predictions Logit-EU, CPL)**.** There exists a strictly increasing u so that $\rho(a, S) = \frac{\exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_{n^*}) a(x)\right)}{\sum_{b \in S} \exp\left(\sum_{x \in \mathcal{X}_{-0}} u(x_{n^*}) b(x)\right)}$ for any $a \in S$ whenever S leads to correct predictions.

If S leads to correct predictions, then every conditional independence assumption implied by her causal model holds in ρ^S. Hence, her predictions about each of the actions are correct. She should therefore behave according to Logit-EU. This axiom can be replaced by assuming that analogues of the independence, monotonicity, and continuity axioms hold for correctly perceived menus; see the online appendix for details.

Theorem 2. *A random choice rule ρ has an Endogenous SCR if and only if ρ satisfies Full-support, Bounded Misperception, CRC, I5, LCI, and CPL.*

The result highlights the connection between SCR and the Logit-EU model. Notice that if $\mathcal{A}^\rho = \{\{0\}, \{n^*\}\}$ and ρ satisfies Axioms 1-6, then the choice rule has a Logit-EU representation. The axioms relate deviations from Logit-EU to inconsistent predictions about causal effects. I5 says that two alternatives are chosen with same probability whenever they coincide on the distribution of variables the action is revealed to cause, whereas Logit-EU requires coincidence on the outcome distribution. LCI requires that the Luce ratio is constant only when predictions are constant, whereas Logit-EU requires it to be constant across all menus. CPL limits deviations from Logit-EU to situations where the DM's predictions do not match reality.

We outline the proof for sufficiency here, and defer a formal proof to the appendix. By Axiom 3, we can define a candidate DAG R^ρ and n^* to represent ρ. Given $S \in \mathcal{S}$, we construct a correctly predicted S'_1 so that for every $a \in S$, there is an $a' \in S'_1$ with the same distribution that she would predict for a given the candidate DAG, i.e., $a'(\cdot) = \rho^S_{R^\rho}(\cdot|a)$. We then show that for any $a, b \in S$, the DM chooses a and b from S with the same frequency as she chooses a' and b' from S'_1. To do so, we add distinct alternatives to S'_1 to form a nested sequence $(S'_m)_{m=1}^\infty$ where each S'_m is correctly predicted. Bounded Misperception implies that the probability of choosing anything in S from $S'_m \cup S$ goes to zero as $|S'_m|$ goes to infinity. Therefore, the inferences that the DM makes when facing $S'_m \cup S$ approach those she makes from S'_1, which in turn equal 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.

6. Discussion

Theorem 1 shows that one can use choice of actions to reveal the relevant parts of a subjective causal model. It suggests the types of questions, and the regularities in data, that could be used in surveys or experiments to infer or to test a subjective model. The approach provides both an alternative approach to asking subjects directly to describe their model, as in Andre et al. (2022), and a complementary way to test whether subjects actually use the model that they described.

We conclude by discussing comparative behavior and alternative interpretations of the model.

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.[14] 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 7. Say that ρ_2 *has a coarser model than* ρ_1 if $\rho_1(a,q) \geq \rho_1(b,q)$ for all $b \in S_q \iff \rho_2(a,q) \geq \rho_2(b,q)$ for all $b \in S_q$ whenever $X_i \perp_q X_{N\setminus\{i\}}$ for all $i \in N$ that are not in any R_2-MAP from 0 or an R_2-confounder to n^*.

Consider DM1 represented by ρ_1 and DM2 represented by ρ_2. As revealed by Theorem 1, the variables that do not belong to an R_2-MAP from 0 or an R_2-confounder to n^* are irrelevant for DM2's predictions. The condition says that whenever those variables are independent of all other variables, i.e. they are actually irrelevant when forming predictions using any DAG, 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 ρ_i have a perfect SCR (R_i, u_i, n^*) for $i = 1, 2$. Then, ρ_2 has a coarser model than ρ_1 if and only if there exists $N' \subset N$ so that ρ_2 has an SCR $(R_1 \cap N' \times N', u_1, n^*)$.*

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, more variables may be relevant for DM1's predictions than for DM2's.

6.2. **Interpretations of the model.** Our main interpretation of an SCR is the one discussed above: it describes a DM who predicts the outcome of her action using a causal model and adjusts her predictions for the spurious correlation caused by any confounders.

Alternatively, the model may 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 from the partial datasets using the distribution that maximizes entropy subject to matching the marginal distributions over the datasets. Identifying her DAG corresponds to identifying the

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

considered subsets, and the DAG will necessarily be perfect. In the 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 contains the correlations between plaque, tangles and 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.[15]

Another interpretation is a particular kind of correlation neglect caused by limited memory. Without confounders, the DAG reduces the number of parameters needed to reconstruct the joint distribution over the variables. In the running example, a DM with DAG R_T can store all the information, she deems relevant for such reconstruction using only 6 parameters when all variables are binary, but it would require $2^4 - 1 = 15$ parameters to record the probability of each possible realization without this assumption.

Another interpretation is that the DM estimates a structural equation model (SEM). She estimates each variable X_i based on the variables indexed by $R(i)$. Her estimates have a causal interpretation if she includes all the variables (and only those) that she believes have a direct causal impact on it. See Chapter 5 of Pearl (2009) for a more in-depth discussion of this.

For a final interpretation, we note that when ρ has an SCR (R, u, n^*), q_R minimizes Kullback-Liebler divergence from q among all the probability distributions on $\mathcal{X}$ that are consistent with ρ^S.[16] A ρ with an Endogenous SCR 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 steady state of a learning process with a set of parameters (probability distributions) that does not include the "true" one.

[15] 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.pdf for the evidence submitted and the FDA's evaluation thereof.

[16] See Section 5.5 of Hajek et al. (1992).

Appendix A. Proofs omitted from the main text

A.1. **Notation.** For a DAG Q, let $\tilde{Q}$ be the skeleton or undirected version of Q, i.e. $j\tilde{Q}i$ if and only if either iQj or jQi, (i,j,k) be a Q-v-collider if and only if iQj, jQk, $i\not{Q}j$ and $j\not{Q}i$, and Q^* be the DAG that drops all edges into 0 so $p_Q(x|do(a)) = p_{Q^*}(x|a)$ for all full support p.

Definition 8. 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 $C \subseteq \{0, \ldots, n\}$ 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 perfect DAG R can be linked to form a junction tree. Call this the *maximal clique junction tree (MCJT)* for R.

A.2. **Proof of Sufficiency for Theorem 1.** Suppose that ρ has a perfect SCR (R, u, n^*) and R' is perfect and non-trivial DAG so that the set of R-confounders equals the set of R'-confounders, the set of R'-MAPs from an R-confounder or 0 to n^* equals the set of R-MAPs from an R-confounder or 0 to n^*. We show that $q_{R^*}(x_{n^*}|a) = q_{R'^*}(x_{n^*}|a)$ for all q and $a \in S_q$. There is no loss in considering only an R for which $R \cap A_{R^*}(n^*)^2 = R$.

Let C^*_R (or just C^* when R is clear) be the set consisting of 0 and all R-confounders. Notice C^* is a clique: for $j, k \in C^*$, $j\tilde{R}k$ if $j = 0$ or $k = 0$ and otherwise, $jR0$ and $kR0$ so $j\tilde{R}k$ by perfection.

Set $R^0 = R$, $N^0(R) = A_R(n^*)$. Inductively define $N^{i+1}(R)$ and R^{i+1} as follows. There is a unique maximal R^i-clique $C^{i\dagger}$ containing n^*, since jR^in^* and kR^in^* implies $j\tilde{R}^ik$ by perfection and $n^*\not{R}^ij$ for all j. For any MCJT for R^i, there is a unique path from any maximal clique containing C^* to $C^{i\dagger}$. Take a maximal clique containing C^* and the MCJT with one of the shortest such paths, and let $\mathcal{C}^i = \{C^i_1, \ldots, C^i_m\}$ be the maximal cliques in this path ordered so that so that there is an edge from C^i_j to C^i_{j+1}, and $C^* \subseteq C^i_1$. If $\cup\mathcal{C}^i \neq N^i(R)$, then set $N^{i+1}(R) = \cup\mathcal{C}^i$. If $\cup\mathcal{C}^i = N^i(R)$ and there is $j_{i+1} \in N^i(R) \setminus [C^* \cup \{n^*\}]$ that is not contained in at least two maximal R^i-cliques, then set $N^{i+1}(R) = N^i(R) \setminus j_{i+1}$. Otherwise, $N^{i+1}(R) = N^i(R)$. Set $R^{i+1} = R^i \cap [N^{i+1}(R) \times N^{i+1}(R)]$.

Since there are finite variables and cliques, there is $\bar{i}$ so that $N^{i+1}(R) = N^i(R)$ for all $i > \bar{i}$. Let $N^*(R) = N^{\bar{i}+1}(R)$.

Lemma 4. *If ρ has a perfect SCR (R, u, n^*), then it also has an SCR $(R \cap N^*(R)^2, u, n^*)$.*

Proof. Clearly, $p_{R^{0*}}(x_{n^*}|x_{C^*}) = p_{R^*}(x_{n^*}|x_{C^*})$. Assume (IH) that $p_{R^*}(x_{n^*}|x_{C^*}) = p_{R^{i*}}(x_{n^*}|x_{C^*})$. We show that (IH) implies $p_{R^*}(x_{n^*}|x_{C^*}) = p_{R^{i+1*}}(x_{n^*}|x_{C^*})$. This establishes the result.

If $\cup\mathcal{C}^i \neq N^i(R)$, then by properties of the MCJT, we can treat C_1^i as ancestral and all the cliques outside of $\mathcal{C}^i$ as being non-ancestors of n^*. Hence

$$p_{R^*}(x_{n^*}|x_{C^*}) = p_{R^{i*}}(x_{n^*}|x_{C^*}) = p_{R^{i+1*}}(x_{n^*}|x_{C^*}).$$

That is, we can drop all variables that do not appear in at least one clique in $\mathcal{C}^i$.

If $\cup\mathcal{C}^i = N^i(R)$, let $(C_1^i, \cdots, C_m^i)$ be the MCJT for R^i, WLOG a single path ordered naturally and having $C^* \subset C_1^i$ and $n^* \in C_m^i$. Let k be so that $j_{i+1} \in C_k^i$ and $j_{i+1} \notin C_{k'}$ for all $k' \neq k$.

Let $l(k') < k'$ be the highest index for each $k' > 1$ so that $C_{k'} \cap C_{k'-1} \subset C_{l(k')}$. Define Q^i so that $jQ^ij' \Leftrightarrow jR^ij'$ when $j, j' \in C^*$, jQ^ij' whenever $j \in C^*$ and $j' \in C_1^i \setminus C^*$, jQ^ij' whenever $j \in C_{k'} \cap C_{l(k')}$ and $j' \in C_{k'} \setminus C_{l(k')}$, jQ^ij_{i+1} whenever $j \in C_k$, and $j\tilde{Q}^ij'$ for each distinct $j, j' \in C_{k'}$ for all k'.

Notice Q^{i*} has the same skeleton and the same v-colliders as R^i and that $j_{i+1} \notin A_{Q^i}(n^*)$. Hence

$$p_{[Q^i \cap N^{i+1}(R)^2]^*}(x_{n^*}|x_{C^*}) = p_{Q^{i*}}(x_{n^*}|x_{C^*}) = p_{R^{i*}}(x_{n^*}|x_{C^*}) = p_{R^*}(x_{n^*}|x_{C^*})$$

where the last equality holds by hypothesis and the second by Theorem 1 of Verma and Pearl (1991). Moreover, $[Q^i \cap N^i(R)^2]^*$ has the same skeleton and same v-colliders as R^{i+1*}, so $p_{R^*}(x_{n^*}|x_{C^*}) = p_{R^{i+1*}}(x_{n^*}|x_{C^*})$. □

Lemma 5. *If $(B_1, \ldots, B_m)$ is any ordering of the maximal cliques of a perfect and non-trivial DAG R that satisfies the running intersection property with $B_1 \supset C^*$, then*

$$p_{R^*}(x_{\bar{B}_m}) = p(x_{C^* \setminus \{0\}})p(x_0)p(x_{B_1}|x_{C^*}) \prod_{j=1}^{m} p(x_{B_j}|x_{B_j \cap \bar{B}_{j-1}}) \tag{3}$$

for any p and where $\bar{B}_k = \cup_{i=1}^k B_i$.

Proof. Let p^* be defined by the RHS of Equation (3) where $\{B_1, \ldots, B_m\}$ are the maximal R-cliques ordered to satisfy the running intersection property, have $B_1 = C_1$; let $l(k) < k$ be the highest index for each $k > 1$ so that $B_k \cap \bar{B}_{k-1} \subset B_{l(k)}$, and define Q from R and $(B_1, \ldots, B_m)$ as we did Q^i from R^i and $(C_1^i, \ldots, C_m^i)$ (ignoring j_{i+1}). Then, Q^* has the same skeleton as R^* since they have the same maximal cliques, and Q^* has the same v-colliders as R^* since any Q^*-v-colliders are between variables in C^*. By Theorem 1 of Verma and Pearl (1991), $p_{R^*} = p_{Q^*}$.

Proceed by induction. Since $C^* \setminus \{0\}$ and $\{0\}$ are ancestral cliques in Q^* that are Q^*-independent, $p_{Q^*}(x_{C^*}) = p(x_{C^* \setminus \{0\}})p(x_0)$. That $p^*(x_{\bar{B}_1}) = p_{Q^*}(x_{\bar{B}_1})$ follows from the formula for $p_{Q^*}(x_{\bar{B}_1})$.

Suppose (IH) that $p^*(x_{\bar{B}_{M-1}}) = p_{Q^*}(x_{\bar{B}_{M-1}})$. For any $j \in B_M \setminus B_{l(M)}$, observe that by construction of Q, kQj only if $k \in B_M \cup B_{l(M)}$.

Pick $j_1 \in B_M \setminus B_{l(M)}$ so that $j' \not{Q} j_1$ for all $j' \in B_M \setminus B_{l(M)}$; j_1 exists since Q is acyclic. Then,

$$p_{Q^*}(x_{\bar{B}_{M-1}}, x_{j_1}) = p^*(x_{\bar{B}_{M-1}})p(x_{j_1}|x_{Q(j_1)}) = p^*(x_{\bar{B}_{M-1}})p^*(x_{j_1}|x_{B_M \cap B_{l(M)}})$$

since $Q(j_1) = B_M \cap B_{l(M)}$. Recursively define $j_k \in B_M \setminus [B_{l(M)} \cup \{j_1, \ldots, j_{k-1}\}]$ so that $j' \not{Q}^* j_k$ for all $j' \in B_M \setminus [B_{l(M)} \cup \{j_1, \ldots, j_{k-1}\}]$ and assume that

$$p_{Q^*}(x_{\bar{B}_{M-1}}, x_{j_1}, \ldots, j_{k-1}) = p^*(x_{\bar{B}_M - 1})p(x_{j_1}, \ldots, x_{j_{k-1}}|x_{B_M \cap B_{l(M)}}).$$

Then,

$$\begin{aligned}
&p_{Q^*}(x_{\bar{B}_{M-1}}, x_{j_1}, \ldots, j_k) \\
=&p^*(x_{\bar{B}_{M-1}})p(x_{j_1}, \ldots, x_{j_{k-1}}|x_{B_M \cap B_{l(M)}})p(x_{j_k}|x_{Q(j_k)}) \\
=&p^*(x_{\bar{B}_{M-1}})p(x_{j_1}, \ldots, x_{j_{k-1}}|x_{B_M \cap B_{l(M)}})p(x_{j_k}|x_{j_1}, \ldots, x_{j_{k-1}}, x_{B_M \cap B_{l(N)}}) \\
=&p^*(x_{\bar{B}_{M-1}})p(x_{j_1}, \ldots, x_{j_k}|x_{B_M \cap B_{l(M)}})
\end{aligned}$$

since $Q(j_k) = \{j_1, \ldots, j_{k-1}\} \cup [B_M \cap B_{l(M)}]$ by construction and that B_M is a Q-clique. Hence $p^*(x_{\bar{B}_M}) = p_{Q^*}(x_{\bar{B}_{M-1}})p(x_{B_M \setminus B_{M-1}}|x_{B_M \cap B_{l(M)}}) = p_{Q^*}(x_{\bar{B}_M})$. Inductively, $p^* = p_{Q^*} = p_{R^*}$. □

By Lemma 4, the MCJT for a perfect R such that $R = R \cap N^*(R)^2$ consists of single path. We say that $(C_1, \ldots, C_m)$ is the MCJT for such an R if each C_i is a maximal clique for each i, there is an edge between C_i and C_{i+1} for each i, and $C_R^* \subseteq C_1$.

Lemma 6. *If $(C_1, \ldots, C_m)$ is the MCJT for a perfect and non-trivial DAG $R \subset N^*(R)^2$, $A_0 = C^*$, $A_m = \{n^*\}$, 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 $j \in A_{i-1}$. □

Definition 9. For a DAG $Q \subseteq N^*(Q)^2$, $k\hat{Q}j$ if and only if kQj and either:

(a) $k \in C^*$ and either $j \notin C^*$ or $j = 0$,
(b) k precedes j on a Q-MAP from $l \in C^*$ to n^*, or
(c) there exists a node $l \in N$ such that $l\hat{Q}k$ and $l \not{Q} j$.

Lemma 7. *If $(C_1, \ldots, C_m)$ is the MCJT for a perfect and non-trivial DAG $R \subset N^*(R)^2$ and $C_0 = C_R^* \subset C_1$ and $\{k, j\} \nsubseteq C_0$, then $k\hat{R}j$ if and only if there exists i so that $k \in C_i \cap C_{i-1}$ and $j \in C_i \setminus C_{i-1}$.*

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 $k \in C^*$. If $j \in C_1 \setminus C_0$, then kRj. To see this, suppose not, i.e. $k \not{R} j$. As $j, k \in C_1$ this implies that jRk. If $jR0$, then $j \notin N^*(R)$ since $j \notin C^*$ implies all paths from j to n^* go through C^*. Hence $0Rj$ and $k \neq 0$, but jRk implies $0RjRkR0$, a cycle. Conclude kRj.

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 $j \in C_i \setminus C_{i-1}$. If $k \in C_0$ and jRk, then there exists $l \in C_{i-1} \setminus C_{i-2}$ and $l' \in C_{i-1} \setminus C_i$. Now, $j\tilde{R}l$ by construction and $l'Rl$ by IH. We must have lRj since jRl would imply $j\tilde{R}l'$, contradicting that $j \notin C_{i-1}$ and $l' \notin C_i$. If $l' \notin C_0$, then by IH, there is an R-path from k to l, leading to a cycle when combined with $lRjRk$. If $l' \in C_0$ and either $l' = k$ or kRl', then there is a cycle. If $l' \in C_0$ and $l'Rk$, then $l'\tilde{R}j$ by perfection, contradicting that $j \notin C_{i-1}$ and $l' \notin C_i$. Hence kRj and, since $k \in C^*$, $k\hat{R}j$.

So suppose that $k \notin C_0$. Then 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 $C_{i'} \nsubseteq C_{i'+1}$. By Lemma 6, $l \in C_{i'} \cap C_{i'-1}$, and by IH,

$l\hat{R}k$. Since $j \notin C_{i''}$ for any $i'' < i$ and $l \notin C_{i''}$ for any $i'' > i'$, $l \not\tilde{R} j$, so by perfection, $j \not R k$. Then, kRj since $k, j \in C_i$, and $k\hat{R}j$ by Definition 9.c.

To complete the proof, we show that "if $j \in C_i \backslash C_{i-1}$ and $k\hat{R}j$, then $k \in C_{i-1} \cap C_i$." Let $j \in C_i \setminus C_{i-1}$ and $k\hat{R}j$. If $i = 1$ and $k \in C^*$, then $k \in C_0 \cap C_1$. If $i = 1$ and $k \notin C^*$, then $j, k \in C_1 \setminus C_0$. Then $k'Rj$ and $k'Rk$ for all $k \in C^*$ by necessity. But this rules out Definitions 1.(a)-(c) holding for k and j, since no R-MAP can contain both. For $i > 1$, take any $j' \in C_i \setminus C_{i+1}$, noting $j' \in C_{i-1} \cap C_i$ by Lemma 6. 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 $k \notin C_{i-1}$. Hence $k \notin C^*$. Since any path from a node in C^* to $n+1$ contains some $j' \in C_{i-1} \cap C_i$, and $j'\hat{R}j$ and $j'\hat{R}k$ by necessity, no R-MAP includes both j and k. Therefore, by Definition 9 there is l so that $l\hat{R}k$ and $l \not\hat{R} j$. Then, $l \in C_i$, since otherwise $k\hat{R}l$ by necessity. Thus, jRl. But then $jRlRkRj$, a contradiction. Hence $k \in C_{i-1} \cap C_i$. □

Lemma 8. *For a perfect and non-trivial DAG $R \subseteq N^*(R)^2$, 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 $j \not\hat{R} k$. If $\{j, k\} \subset C^*$, then $k \neq 0$ by Definition 9(a), and $j \neq 0$ as $k \in C^*$. Thus, j and k are R-confounders, and so $j\hat{R}0$ and $k\hat{R}0$ by Definition 9. Otherwise, let $(C_1, \ldots, C_m)$ be a MCJT for R, noting that $j, k \in C_i$ for some i and that if $k \notin C_{i-1}$, then $j \notin C_{i-1}$ by Lemma 7 and $j \not\hat{R} k$. This combined with Lemma 6 implies that either $j, k \in C_i \cap C_{i-1}$ or $j, k \in C_i \cap C_{i+1}$. In the former case pick $l \in C_i \setminus C_{i-1}$ and in the latter pick $l \in C_{i+1} \setminus C_i$. In either case, $j\hat{R}l$ and $k\hat{R}l$ by Lemma 7. □

By Lemmas 7 and 8, any two DAGs that have the same confounders have the same C^* and the same MAPs from C^* to n^* correspond to the same MCJT after removing unnecessary variables. Therefore, if R and R' have the same confounders and the same MAPs from a confounder or 0 to n^*, then $q_R(x_{n^*}|do(a)) = q_{R'}(x_{n^*}|do(a))$ for any q and $a \in S_q$.

A.3. **Necessity.** It will be useful to introduce the following class of datasets. Partition every $\mathcal{X}_{j'}$ into $\{\bar{E}_{j'}, \underline{E}_{j'}\}$ where $x_{j'} \in \bar{E}_{j'}$ and $x'_{j'} \in \underline{E}_{j'}$ implies $x_{j'} > x'_{j'}$ and let $\bar{E}_0 = \{a\}$ and $\underline{E}_0 = \{b\}$. A dataset q is *pseudo-binary* if $S_q = \{a, b\}$ and the distribution within each $\bar{E}_{j'}$ and $\underline{E}_{j'}$ is independent of all events outside of j'.

Proof of Lemma 1. Let $(i_0, \ldots, i_m)$ be an R-MAP from 0 to n^* that does not intersect A. Pick a pseudo-binary dataset q that factorizes as

$$q(x) = q(x_0) \prod_{j=1}^{m} q(x_{i_j}|x_{i_{j-1}}) \prod_{k' \notin \{i_0,\ldots,i_m\}} q(x_{k'})$$

and where $q(\bar{E}_{i_j}|\bar{E}_{i_{j-1}}) > q(\bar{E}_{i_j}|\underline{E}_{i_{j-1}})$ for all j. This implies that $X_A \perp_q X_{A^c}$ since $A \subseteq N \setminus \{i_0, \ldots, i_m\}$. For such a dataset q, $\rho(a, q) > \rho(b, q)$ since $q_R(\cdot|do(a))$ FOSD $q_R(\cdot|do(b))$, so A does not separate.

Suppose that A intersects all paths from 0 to n^* and let R' remove all links from A to A^c and vice versa. Observe that when $X_A \perp_q X_{A^c}$,

$$q(x_i|x_{R(i)}) = q(x_i|x_{R(i)\cap A}, x_{R(i)\setminus A}) = q(x_i|x_{R(i)\cap A}) = q(x_i|x_{R'(i)})$$

when $i \in A$, and for $i \notin A$,

$$q(x_i|x_{R(i)}) = q(x_i|x_{R(i)\cap A}, x_{R(i)\setminus A}) = q(x_i|x_{R(i)\setminus A}) = q(x_i|x_{R'(i)}).$$

Hence $q_R(x) = q_{R'}(x)$ for all x. Since $0 \notin A_{R'^*}(n^*)$, 0 and n^* are q_{R^*} independent and $\rho(a, q) = \rho(b, q)$ for all $a, b \in S_q$. $\square$

Lemma 9. *Let ρ have a perfect SCR $(R \subseteq N^*(R)^2, u, n^*)$ and $R' = R \cap [N^*(R \setminus [R(0) \times N])]^2$. Then, $A \in \mathcal{A}$ if and only if $A = C_i \cap C_{i+1}$ for some i where $(C_1, \ldots, C_m)$ is a MCJT for R', $C_0 = \{0\}$, and $C_{m+1} = \{n^*\}$.*

Proof. Let ρ have a perfect SCR $(R \subseteq N^*(R)^2, u, n^*)$, $R' = R \cap [N^*(R \setminus [R(0) \times N])]^2$, $(C_1, \ldots, C_m)$ be a MCJT for R', $C_0 = \{0\}$, $C_{m+1} = \{n^*\}$, $A_i = C_i \cap C_{i+1}$ for each i, and $B_i = \cup_{j=1}^{i} C_j$ for each i. Moreover, observe that every R-MAP from 0 to n^* is an R'-MAP, and every R'-AP (R'-MAP) from 0 to n^* is a R-AP (R-MAP). Clearly, $A_0, A_m \in \mathcal{A}$.

Pick any $i \geq 0$. We first show that every R-AP from 0 to n^* intersects A_{i+1}. Take any such 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, $i_{k-1} \in C_{j-1} \cap C_j \subset B_{j-1}$ by Lemma 7, so $j - 1 = i$ and $i_k \in A_{i+1}$. By Lemma 1, A_{i+1} separates.

Let $A \in \mathcal{A} \setminus \{\{0\}, \{n^*\}\}$. By Lemma 1, A intersects every R-AP from 0 to n^*. By Theorem 4.4 of Cowell et al. (1999), A is a clique, so $A \subset C_i$ for some i. We show that either $A = A_i$ or $A = A_{i-1}$. Since both A_i and A_{i-1} separate and A is minimal, it suffices to show that either $A_{i-1} \subseteq A$ or $A_i \subseteq A$.

For contradiction, suppose that $A_i \cap A_{i-1} \not\subseteq A$. There exists $j \in A_i \cap A_{i-1} \setminus A$. 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 7, $j' \hat{R}' j$ and $j \hat{R}' l$, so there exists a R-AP that does not intersect A. Therefore, $A_i \cap A_{i-1} \subseteq A$.

If $A_{i-1} \not\subseteq A$, then there is $j \in A_{i-1} \setminus A$. For any $k \in A_i \setminus A_{i-1}$, $j \hat{R} k$ by Lemma 7 and since A blocks all R-APs, $k \in A$. Since k was arbitrary, $A_i \subseteq A$. If $A_i \not\subseteq A$, then there is $k \in A_i \setminus A \subset A_i \setminus A_{i-1}$. For any $j \in A_{i-1}$, $j \hat{R} k$ by Lemma 7 and since A blocks all R-APs, $j \in A$. Since j was arbitrary, $A_{i-1} \subseteq A$.

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 6, $x \in A_i \cap A_j$ implies that $x \in [C_{i+1} \cap C_{j+1}] \setminus R(0)$. Since $(C_1, \dots, C_m)$ is a MCJT, $x \in C_{i+2}$. But then every $x \in A_i$ is also in A_{i+1}, and by Lemma 6, $C_{i+1} = A_i \cup A_{i+1} = A_{i+1} \subseteq C_{i+2}$, contradicting that C_{i+1} is maximal. $\square$

Proof of Lemma 2. Let $(C_1, \dots, C_m)$ by a MCJT for $R' = R \cap [N^* (R \setminus [R(0) \times N])]^2$, and take $A_i = C_i \cap C_{i+1}$ with $A_0 = \{0\}$ and $A_m = \{n^*\}$. By Lemma 9, $\mathcal{A} = \{A_0, \dots, A_m\}$.

Consider any selection J from $\mathcal{A}$. For all i, let $F(i) = \min\{k : i \in A_k\}$ and $L(i) = \max\{k : i \in A_k\}$. Observe $0 \in J$. Take $i_0 = 0$, noting $F(0) = L(0) = 0$. Let $J_0 = J \setminus \{0\}$. While $J_{l-1} \neq \emptyset$, let k_l be the lowest index for which $J_{l-1} \cap A_{k_l} \neq \emptyset$. Note that k_l exists since $J \setminus J_{l-1}$ does not intersect some $A \in \mathcal{A}$ by minimality; $k_l > F(i_{l-1})$ and, when $l > 1$, $k_l > L(i_{l-2}) + 1$ since otherwise J can be made smaller; and $k_l \leq L(i_{l-1}) + 1$ since otherwise $J \cap A_{L(i_{l-1})+1} = \emptyset$. Moreover, $|J_{l-1} \cap A_{k_l}| = 1$. To see this, suppose not, and there are $j_1, j_2 \in J_{l-1} \cap A_{k_l}$. Then there is κ_i so that $j_i \in A_k$ for $k \in [k_l, \kappa_i]$ by Lemma 6 and $(C_1, \dots, C_m)$ is a MCJT. If $\kappa_1 \geq \kappa_2$ ($\kappa_2 \geq \kappa_1$), then $J \setminus \{j_2\}$ ($J \setminus \{j_1\}$) still intersects every A, contradicting minimality of J.

Define $i_l \in J_{l-1} \cap A_{k_l}$ and $J_l = J_{l-1} \setminus \{i_l\}$. By above, $i_{l-1} \in A_{k_l - 1}$, so $i_{l-1} R i_l$ and $F(i_l) > L(i_{l-2}) + 1$ implies that $i_{l'} \not{R} i_l$ for all $l' < l - 1$ by Lemma 7. Moreover, minimality of J requires that $L(i_l) > L(i_{l-1})$. Repeating until $l = |J| - 1$, conclude that $(i_0, \dots, i_{|J|-1})$ is a R-MAP from 0 to n^*.

Conversely, pick any R-MAP $(i_0 = 0, \dots, i_m = n^*)$ and let $J = \{i_0, \dots, i_m\}$. By Lemma 1, $J \cap A \neq \emptyset$ for any separator A. Clearly, $\{i_0\} = J \cap A_0$ and $\{i_m\} = J \cap A_m$, so $J \setminus \{i_0\} \cap A_0 = \emptyset$ and $J \setminus \{i_m\} \cap A_m = \emptyset$. By Lemmas 6 and 7, for each $l \in (0, m)$, there exists k_l so that $i_l \in A_{k_l}$ and $i_{l+1} \in A_{k_l+1} \setminus A_{k_l}$. Since $i_{l'} \not R i_{l+1}$ for all $l' \neq l$ by definition, $i_{l'} \notin A_{k_l}$ for all $l' \neq l$ by Lemma 6. Therefore, $J \setminus \{i_l\} \cap A_{k_l} = \emptyset$ for all $l \in (0, m)$, and so no proper subset of J intersects every $A \in \mathcal{A}$.

Now, let I be any selection from $\mathcal{A}$. By above, $I = \{i_0, \dots, i_m\}$ for some R-MAP $(i_0, \dots, i_m)$ from 0 to n^*. We show that i_j is adjacent to i_{j+1} for each j. Fix $j < m$ and consider q so that $X_{i_j} \perp_q X_{i_{j+1}}$ and $X_I \perp_q X_{I^c}$. Note that

$$q(x_{i_k}|x_{R(i_k)}) = q(x_{i_k}|x_{R(i_k)\cap I}, x_{R(i_k)\cap I^c}) = q(x_{i_k}|x_{R(i_k)\cap I}) = q(x_{i_k}|x_{i_{k-1}})$$

for $k > 0$. Therefore,

$$q_R(x_{n^*}|a) = \sum_y q(y_{i_1}|a) \prod_{k=1}^{m-2} q(y_{i_{k+1}}|y_{i_k}) q(x_{n^*}|y_{i_{m-1}}),$$

which does not vary with a if $q(y_{i_{j+1}}|y_{i_j}) = q(y_{i_{j+1}})$ for all y.

Now, pick any $j < k - 1$; we show that i_j is not adjacent to i_k. Pick a pseudo-binary dataset q that factorizes as

$$q(x) = q(x_0) q(x_{i_k}) \prod_{l \neq k-1, k-2} q(x_{i_{l+1}}|x_{i_l}) q(x_{i_{k-1}}|x_{i_{k-2}}, x_{i_k}) q(x_{I^c})$$

and where

$$q(\bar{E}_{i_{k-1}}|x_{\{i_k, i_{k-2}\}}) = \begin{cases} \frac{3}{4} & \text{if } x_{\{i_k, i_{k-2}\}} \in \bar{E}_{i_k} \times \bar{E}_{i_{k-2}} \\ \frac{1}{4} & \text{otherwise} \end{cases}$$

and $q(\bar{E}_{i_l}|\bar{E}_{i_{l-1}}) > q(\bar{E}_{i_l}|\underline{E}_{i_{l-1}})$ for $l \neq k, k-1$. By construction, $X_I \perp_q X_{I^c}$ and $X_{i_j} \perp_q X_{i_k}$. Nevertheless, $q(\bar{E}_{i_l}|\bar{E}_{l-1}) > q(\bar{E}_{i_l}|\underline{E}_{i_{l-1}})$ for $l = k - 1$ and $l = k$, so $q_{R^*}(\bar{E}_{n^*}|a) > q_{R^*}(\bar{E}_{n^*}|b)$, so $\rho(a, q) > \rho(b, q)$. Conclude that i_j and i_k are not adjacent. $\square$

Let N^* be the variables in some R-MAP for 0 to n^*, i.e. $N^* = \{0\} \bigcup_{i=1}^{|\mathcal{A}|} \mathcal{A}_i^*$ and $\hat{N}$ the remaining variables.

Lemma 10. *Let $i^* \in C^* \setminus \{0\}$. If $(i_0 = 0, \ldots, i_M = n^*)$ and $(j_0 = i^*, \ldots, j_{k'} = i_k)$ are R-MAPs with $\{j_1, \ldots, j_{k'-1}\} \subseteq \hat{N} \setminus C^*$ and $k' > 1$, then $i_{k-1} R j_l$ for all $l \in \{1, \ldots, k'-1\}$ and $i^* R i_{l'}$ for all $l' \in \{1, \ldots, k-1\}$.*

Proof. Note that $i_{k-1} \tilde{R} j_{k'-1}$ as R is perfect, so suppose for contradiction that $j_{k'-1} R i_{k-1}$. If $k = 1$, then $i_0 = 0$ and then $j_{k'-1} \in C^*$, a contradiction. If $k > 1$, then there exists some $l < k-1$ such that $i_l R j_{k'-1}$ as otherwise $j_{k'-1} \in C^*$. But then $(i_0, \ldots, i_l, j_{k'-1}, i_{k-1}, \ldots, i_M)$ is an R-MAP, which contradicts that $j_{k'-1} \notin N^*$.

Suppose that (*) $i_{k-1} R j_{l'}$ for $l' \in \{l, \ldots, k\}$ and, for contradiction, that $i_{k-1} \not R j_{l-1}$. By perfection, $j_{l-1} R i_{k-1}$. Again, $k-1 = 0$ implies that $j_{l-1} \in C^*$, a contradiction, so there exists some $l' < k-1$ such that $i_{l'} R j_{l-1}$. But then $(i_0, \ldots, i_{l'}, j_{l-1}, \ldots, j_{k'-1}, i_k, \ldots, i_M)$ is an R-MAP, contradicting that $j_{l-1} \notin N^*$. Induction establishes that $i_{k-1} R j_l$ for $l = 1, \ldots, k'-1$. Note that $i^* \tilde{R} i_{k-1}$ by perfection and $i^* R i_{k-1}$ as R is acyclic. The same arguments inductively establish that $i^* R i_{l'}$ for all $l' \in \{1, \ldots, k-1\}$. □

Definition 10. The covariate $i^* \notin N^*$ is a *revealed confounder* if

$$X_{N \setminus \{i^*, 0\}} \perp_q X_0$$

does not imply that $\rho(a, q) = \rho(b, q)$ for all $a, b \in S_q$ and all q. Let $C^*(\rho)$ be the set of all revealed confounders.

Lemma 11. *i^* is a R-confounder if and only if $i^* \in C^*(\rho)$.*

Proof. Suppose that i^* is an R-confounder. This implies that there is an R-MAP $(i_0 = i^*, \ldots, i_m = n^*)$ that does not go through 0 or any other j^* such that $j^* R 0$. Let $I = \{i_0, \ldots, i_m, 0\}$ and construct a pseudo-binary dataset q that factorizes according to

$$q(x) = q(x_{i_0} \mid x_{\{i_1, 0\}}) q(x_0) q(x_{i_1}) \prod_{k' > 1} q(x_{i_{k'}} \mid x_{i_{k'-1}}) \prod_{j \notin I} q(x_j)$$

where $q(\bar{E}_{i_1}) \geq \frac{5}{6}$, $q(\bar{E}_{i_{k'}} \mid \bar{E}_{i_{k'-1}}) > q(\bar{E}_{i_{k'}} \mid \underline{E}_{i_{k'-1}})$ for $k' \geq 2$, and

$$q(\bar{E}_{i_0} \mid x_{\{i_1, 0\}}) = \begin{cases} \frac{3}{4} & \text{if } x_{\{i_1, 0\}} \in \bar{E}_{i_1} \times \bar{E}_0, \\ \frac{1}{4} & \text{otherwise.} \end{cases}$$

By construction, $X_0 \perp_q X_{N \setminus \{0, i^*\}}$, $q(\bar{E}_{i_0}) = \frac{1}{4} + \frac{1}{2} q(\bar{E}_{i_1}) q(a) < \frac{3}{4}$,

$$q(\bar{E}_{i_1} \mid \bar{E}_{i_0}, a) = \frac{3}{q(\bar{E}_{i_1})^{-1} + 2} > \frac{1}{q(\bar{E}_{i_1})^{-1} + 2} = q(\bar{E}_{i_1} \mid \underline{E}_{i_0}, a),$$

and $q(\bar{E}_{i_1} \mid E_{i_0}, b) = q(\bar{E}_{i_1})$. Hence,

$$q_{R^*}(\bar{E}_{i_1} \mid a) < \frac{3}{4}1 + \frac{1}{4}\frac{1}{3} = \frac{5}{6} \leq q(\bar{E}_{i_1}) = q(\bar{E}_{i_1} \mid b).$$

Combined with the above, we have $q_{R^*}(\bar{E}_{n^*} \mid b) > q_{R^*}(\bar{E}_{n^*} \mid a)$. Thus, $\rho(b, q) > \rho(a, q)$ and $i^* \in C^*(\rho)$.

Now suppose that $j^* \in \hat{N}$ is not an R-confounder. WLOG, assume $R = R \cap N^*(R)^2$ so there exists a MCJT for R, $(C_1, \ldots, C_m)$ such that $C^* \subset C_1$ and $(C_1, \ldots, C_m)$ satisfies the running intersection property with $C_k \cap \bar{C}_{k-1} \subset C_{k-1}$ when $\bar{C}_k = \cup_{k'=1}^{k} C_{k'}$ for every k.

Consider any q that satisfies $X_{N \setminus \{0, j^*\}} \perp_q X_0$. By Lemma 5 we have

$$\begin{aligned} q_{R^*}(x_{\bar{C}_k}) &= q(x_0) q(x_{C^* \setminus \{0\}}) q(x_{C_1 \setminus C^*} \mid x_{C^*}) \prod_{k'=2}^{k} q(x_{C_{k'} \setminus C_{k'-1}} \mid x_{C_{k'} \cap C_{k'-1}}) \\ &= q(x_{C_k \setminus C_{k-1}} \mid x_{C_k \cap C_{k-1}}) q_{R^*}(x_{\bar{C}_{k-1}}). \end{aligned} \tag{4}$$

We use this to show that $q_{R^*}(C_k) = q(C_k)$ for every k. For $k = 1$, this follows directly from the assumptions on q and $j^* \notin C^*$: $q_{R^*}(C_1) = q(x_0) q(x_{C^* \setminus \{0\}}) q(x_{C_1 \setminus C^*} | x_{C^*}) = q(x_0) q(x_{C^* \setminus \{0\}} \mid x_0) q(x_{C_1 \setminus C^*} | x_{C^*}) = q(x_{C_1})$. Assume (IH) that $q_{R^*}(x_{C_{k-1}}) = q_{R^*}(x_{C_{k-1}})$. Then,

$$\begin{aligned} q_{R^*}(x_{C_k}) &= \sum_{y \in \mathcal{X}_{C_{k-1} \setminus C_k}} q(x_{C_k \setminus C_{k-1}} \mid x_{C_k \cap C_{k-1}}) q_{R^*}(x_{C_{k-1} \cap C_k}, y_{C_{k-1} \setminus C_k}) \\ &= \sum_{y \in \mathcal{X}_{C_{k-1} \setminus C_k}} q(x_{C_k \setminus C_{k-1}} \mid x_{C_k \cap C_{k-1}}) q(x_{C_{k-1} \cap C_k}, y_{C_{k-1} \setminus C_k}) = q(x_{C_k}), \end{aligned}$$

where the second equality follows from (IH).

Let k^* be the last index for which $0 \in C_{k^*}$. If $n^* \in C_{k^*}$, then $q_{R^*}(x_{n^*}|a) = q_{R^*}(x_{n^*}|b)$ for all $a, b \in S_q$. If $j^* \notin C_{k^*} \cap C_{k^*+1}$, then $q(x_{C_{k^*} \cap C_{k^*+1}}|a) = q(x_{C_{k^*} \cap C_{k^*+1}}|b)$ since $X_{C_{k^*} \cap C_{k^*+1}} \perp_q X_0$. Then, $q_{R^*}(x_{n^*}|a) = q_{R^*}(x_{n^*}|b)$ for all $a, b \in S_q$ follows from Eq (4). Therefore, $j^* \notin C^*(\rho)$ follows if we show $j^* \in C_{k^*} \cap C_{k^*+1}$ or $n^* \in C_{k^*}$. Suppose not, so $n^* \notin C_{k^*}$ and $j^* \in C_{k^*} \cap C_{k^*+1}$. By Lemma 7, $0 R j^*$. For any $l \in C_{k^*+2} \setminus C_{k^*+1}$, $0 R j^* R l$, $0 \not{R} l$, and there is an R-AP from l to n^*. But this implies that there exists an R-MAP $(0, j^*, \ldots, n^*)$, contradicting that $j^* \notin N^*$. $\square$

Definition 11. There is a *revealed confounding path* from $i^* \in C^*(\rho)$ to i_k in the R-MAP $(i_0 = 0, \ldots, i_M = n^*)$ if

$$\begin{aligned} X_j \perp_q X_{N \setminus j} \quad \forall j \in N^* \cup C^*(\rho) \setminus \{i^*, i_0, \ldots, i_M\} \\ \& \; X_{i_k} \perp_q X_{i_{k-1}} \end{aligned}$$

do not imply that $\rho(a,q) = \rho(b,q)$.

First condition implies that the only R-confounder that matters is i^*, and the only R-MAP from 0 to n^* that matters is $(i_0, \ldots, i_M)$. The remaining condition ensures that in the absence of a path from i^* to i_k, the DM will express indifference.

Lemma 12. *For R-MAP $(i_0 = 0, \ldots, i_M = n^*)$, there is a revealed confounding path from $i^* \in \hat{N}$ to i_k $\iff$ there exists an R-MAP from i^* to i_k that does not intersect $N^* \cup C^*$.*

Proof. Suppose that i^* is an R-confounder and there exists an R-MAP from i^* to i_k, $(j_0 = i^*, j_1, \ldots, j_m = i_k)$. that doesn't intersect N^*.

Let $C = \{i_0, \ldots, i_M, j_0, \ldots, j_m\}$. Pick a pseudo-binary dataset q that factorizes according to

$$q(x) = q(x_{\{i_0, j_0\}}) \prod_{k' \neq k} q(x_{i_{k'}} | x_{i_{k'-1}}) \prod_{k'=1}^{m-1} q(x_{j_{k'}} | x_{j_{k'-1}}) q(x_{i_k} | x_{i_{k-1}, j_{m-1}}) \prod_{j' \notin C} q(x_{j'})$$

where

(i) $q(\bar{E}_{i_{k'}} | \bar{E}_{i_{k'-1}}) > q(\bar{E}_{i_{k'}} | \underline{E}_{i_{k'-1}})$ for $k' \neq k$;
(ii) $q(\bar{E}_{j_{k'}} | \bar{E}_{j_{k'-1}}) > q(\bar{E}_{j_{k'}} | \underline{E}_{j_{k'-1}})$ for $k' < m$; and
(iii) for appropriate $z, z', \epsilon \in (0,1)$,

$X_{\{i_0, j_0\}}$	$q(.)$	$X_{\{i_{k-1}, j_{m-1}\}}$	$q(\bar{E}_{i_k} \vert \cdot)$
$\bar{E}_{i_0} \times \bar{E}_{j_0}$	$z\frac{1}{4}$	$\bar{E}_{i_{k-1}} \times \bar{E}_{j_{m-1}}$	$\frac{q\left(\bar{E}_{j_{m-1}} \vert \underline{E}_{i_{k-1}}\right)}{q\left(\bar{E}_{j_{m-1}} \vert \bar{E}_{i_{k-1}}\right)} z' + \epsilon$
$\underline{E}_{i_0} \times \bar{E}_{j_0}$	$(1-z)\frac{3}{4}$	$\underline{E}_{i_{k-1}} \times \bar{E}_{j_{m-1}}$	$z' + \epsilon$
$\bar{E}_{i_0} \times \underline{E}_{j_0}$	$z\frac{3}{4}$	$\bar{E}_{i_{k-1}} \times \underline{E}_{j_{m-1}}$	ϵ
$\underline{E}_{i_0} \times \underline{E}_{j_0}$	$(1-z)\frac{1}{4}$	$\underline{E}_{i_{k-1}} \times \underline{E}_{j_{m-1}}$	ϵ

so $X_{i_{k-1}} \perp_q X_{i_k}$ and $q(\bar{E}_{i_k}|x_{j_{m-1}}, \bar{E}_{i_{k-1}}) > q(\bar{E}_{i_k}|x_{j_{m-1}}, \underline{E}_{i_{k-1}})$ for all x.[17]

Then, $q_{R^*}(x_C)$ equals

$$\begin{aligned}
&q(x_{i_0})q(x_{j_0}) \prod_{k'=1}^{m-1} q(x_{j_{k'}}|x_{R(j_{k'})}) \prod_{k'=1}^{M} q(x_{i_{k'}}|x_{R(i_{k'})}) \\
=&q(x_{i_0})q(x_{j_0}) \prod_{k'=1}^{m-1} q(x_{j_{k'}}|x_{j_{k'-1}}, x_{i_{k-1}}) \prod_{k'=1}^{k-1} q(x_{i_{k'}}|x_{i_{k'-1}}, x_{j_0}) \prod_{k'=k+1}^{M} q(x_{i_{k'}}|x_{i_{k'-1}}) q(x_{i_k}|x_{i_{k-1}}, x_{j_{m-1}}) \\
=&q(x_{i_0})q(x_{j_0}) \prod_{k'=1}^{m-1} q(x_{j_{k'}}|x_{j_{k'-1}}) \prod_{k'=1}^{k-1} q(x_{i_{k'}}|x_{i_{k'-1}}) \prod_{k'=k+1}^{M} q(x_{i_{k'}}|x_{i_{k'-1}}) q(x_{i_k}|x_{i_{k-1}}, x_{j_{m-1}})
\end{aligned}$$

where the second equality follows from the construction of q.

A simple induction argument establishes that $q(\bar{E}_{i_{k'}}|\bar{E}_{i_{k''}}) > q(\bar{E}_{i_{k'}}|\underline{E}_{i_{k''}})$ whenever $k > k' > k''$ or $k' > k'' \geq k$, and $q(\bar{E}_{j_{m-1}}|\bar{E}_{j_1}) > q(\bar{E}_{j_{m-1}}|\underline{E}_{j_1})$. Then, $q_{R^*}(\bar{E}_{i_k}|c)$ equals

$$\sum_x q_{R^*}(x_{j_{m-1}}) \Big\{ q(\bar{E}_{i_{k-1}}|c) \Big[q(\bar{E}_{i_k}|\bar{E}_{i_{k-1}}, x_{j_{m-1}}) - q(\bar{E}_{i_k}|\underline{E}_{i_{k-1}}, x_{j_{m-1}}) \Big] + q(\bar{E}_{i_k}|\underline{E}_{i_{k-1}}, x_{j_{m-1}}) \Big\}$$

and

$$q_{R^*}(\bar{E}_{n^*}|c) = q_{R^*}(\bar{E}_{i_k}|c)[q(\bar{E}_{n^*}|\bar{E}_{i_k}) - q(\bar{E}_{n^*}|\underline{E}_{i_k})] + q(\bar{E}_{n^*}|\underline{E}_{i_k}).$$

Combining implies $q_{R^*}(\bar{E}_{n^*}|a) > q_{R^*}(\bar{E}_{n^*}|b)$.

Now, suppose that there is no path in $\hat{N}$ from i^* to i_k. Observe that if there is a path in $\hat{N}$ from i^* to i_{k+1}, there is also a path to i_k by Lemma 10. Consider any q satisfying the conditions. Define R' that removes all edges involving a node in $N^* \cup C^*(\rho) \setminus \{i^*, i_0, \ldots, i_m\}$ but otherwise agree with R. Take $R'' = R' \cap N^*(R')^2$, and arguments as in Lemma 1 show that $q_{R^*}(x_{n^*}|c) = q_{R'^*}(x_{n^*}|c) = q_{R''^*}(x_{n^*}|c)$. Lemma 7 shows that $N^*(R')$ equals the variables that are part of some R'-MAP from i^* or 0 to n^*. Let $l \in R''(i_k)$. Either l is in an R''-MAP from 0 to n^*, in which case $l = i_{k-1}$, or l is in an R''-MAP from i^* to n^*, in which case $l = i_{k-1}$ or l is part of a path in $\hat{N}$ from i^* to i_k, a contadiction. Therefore, $R''(i_k) = \{i_{k-1}\}$ and $q(x_{i_k}|x_{R''(i_k)}) = q(x_{i_k}|x_{i_{k-1}}) = q(x_{i_k})$ since $X_{i_k} \perp_q X_{i_{k-1}}$. Then, the DM is indifferent, and there is no revealed confounding path from i^* to i_k. $\square$

[17]The quantity $\frac{q(\bar{E}_{j_{m-1}}|\underline{E}_{i_{k-1}})}{q(\bar{E}_{j_{m-1}}|\underline{E}_{i_{k-1}})}$ is pinned down by (1), (2), and the first part of (3), so $z', \epsilon \in (0,1)$ exist satisfying the conditions so that q is a well-defined conditional probability.

Definition 12. If $(i_0 = 0, \ldots, i_m = n^*)$ is an R-MAP and there is a revealed confounding path from $i^* \in C^*(\rho)$ to i_k, then $B \subset \{i^*\} \cup [\hat{N} \setminus C^*(\rho)]$ is a (i^*, i_k)-separator if $\rho(a,q) = \rho(b,q)$ whenever

$$\begin{aligned} &X_j \perp_q X_{\{j\}^c} \quad \forall j \in N^* \cup C^*(\rho) \setminus \{i^*, i_0, \ldots, i_m\} \\ &X_{i_k} \perp_q X_{i_{k-1}} \\ \& \ &X_B \perp_q X_{B^c}. \end{aligned}$$

The first two conditions are the same as before, and the third condition is analogous to separation.

Remark 5. $\{i^*\}$ is a (i^*, i_k)-separator but $\{i_k\}$ is not.

Lemma 13. *B is a (i^*, i_k)-separator $\iff$ B intersects all R-paths from i^* to i_k in $\hat{N}$.*

Proof. If B does not intersect a path from i^* to i_k contained in $[(N^* \cup C^*) \setminus \{i^*, i_k\}]^c$, then we can find a minimal R-path $(j_0 = i^*, j_1, \ldots, j_m = i_k)$ with $j_l \notin B \cup N^* \cup C^*$ for all $l \neq 0, m$. Then, the dataset q from Lemma 12 constructed for this path leads to $\rho(a,q) \neq \rho(b,q)$.

If B intersects all paths from i^* to i_k, then consider R' that drops all edges involving a node in B and a node in B^c. Since $X_B \perp_q X_{B^c}$, $q_{R'^*}(x_{n^*}|x_0) = q_{R^*}(x_{n^*}|x_0)$ for all $x \in S_q \times \mathcal{X}_{-0}$. Applying Lemma 12 to R' and q establishes that $\rho(a,q) = \rho(b,q)$. □

Definition 13. Let $\mathcal{A}^{(i^*,i_k)}$ be the set of minimal (i^*, i_k)-seperators, and say that J is a selection from $\mathcal{A}^{(i^*,i_k)}$ if $J \cap A \neq \emptyset$ for all $A \in \mathcal{A}^{(i^*,i_k)}$ and no proper subset of J has a non-empty intersection with every $A \in \mathcal{A}^{(i^*,i_k)}$. For any selection J from $\mathcal{A}^{(i^*,i_k)}$, $i \in J$ is *adjacent* to $j \in J$ if $\rho(a,q) = \rho(b,q)$ for all $a, b \in S_q$ whenever

$$\begin{aligned} &X_l \perp_q X_{N \setminus l} \quad \forall l \in N \setminus [\{i_0, \ldots, i_m\} \cup J], \\ &X_i \perp_q X_j | X_{i_{k-1}}, \text{ and} \\ \& \ &X_{i_k} \perp_q X_{i_{k-1}}. \end{aligned}$$

The first two conditions are analogous to the conditions in Definition 3, and the third condition ensure that the DM expresses indifference whenever all variables outside the R-MAP between 0 and n^* are independent.

Lemma 14. *The sequence* $(j_0 = i^*, \ldots, j_{M-1}, i_k)$ *is an* R*-MAP from* $i^* \in C^*(\rho)$ *to* i_k *and* $J = \{j_0, \ldots, j_{M-1}\}$ *does not intersect* N^* *if and only if* J *is a selection from* $\mathcal{A}^{(i^*,i_k)}$ *and* j_l *is adjacent to* j_{l+1} *for all* $l < M-2$.

Proof. As in Lemma 2, J is a selection from $\mathcal{A}^{(i^*,i_k)}$ if and only if there is an R-MAP $(j_0 = i^*, \ldots, j_M = i_k)$, $\{j_0, \ldots, j_{M-1}\} = J$, and $J \cap C^*(\rho) = \{i^*\}$.

Let $N^\dagger = \{i_0, \ldots, i_m\} \cup J$. We show that $j_{k'}$ is adjacent to $j_{k'+1}$. Pick any q satisfying the conditions for $j_{k'}$ and $j_{k'+1}$ to be adjacent. Then by Lemma 10,

$$q(x_{i_l}|x_{R(i_l)}) = q(x_{i_l}|x_{R(i_l)\cap N^\dagger}, x_{R(i_l)\setminus N^\dagger}) = q(x_{i_l}|x_{R(i_l)\cap N^\dagger}) = q(x_{i_l}|x_{i_{l-1}}, x_{j_0})$$

for $l < k$. Similarly, $q(x_{i_l}|x_{R(i_l)}) = q(x_{i_l}|x_{i_{l-1}})$ for $l > k$. Moreover,

$$q(x_{j_l}|x_{R(j_l)}) = q(x_{j_l}|x_{R(j_l)\cap N^\dagger}, x_{R(j_l)\setminus N^\dagger}) = q(x_{j_l}|x_{R(j_l)\cap N^\dagger}) = q(x_{j_l}|x_{j_{l-1}}, x_{i_{k-1}})$$

for $l > 0$. In particular, $q(x_{j_{k'+1}}|x_{R(j_{k'+1})}) = q(x_{j_{k'+1}}|x_{i_{k-1}})$ by the second condition in Definition 13.

Note that for $s = k'+2, ..., M$,

$$\sum_{y_{j_{s-1}}} q(x_{j_s}|y_{j_{s-1}}, x_{i_{k-1}})q(y_{j_{s-1}}|x_{i_{k-1}}) = \sum_{y_{j_{s-1}}} q(y_{j_{s-1}}, x_{j_s}|x_{i_{k-1}})$$
$$= \sum_{y_{j_{s-1}}} q(y_{j_{s-1}}|x_{j_s}, x_{i_{k-1}})q(x_{j_s}|x_{i_{k-1}}) = q(x_{j_s}|x_{i_{k-1}}).$$

Since j_s is the only variable dependent on j_{s-1} in $q_{R^*}(\cdot|x_0)$ and $q(x_{j_{k'+1}}|x_{R(j_{k'+1})}) = q(x_{j_{k'+1}}|x_{i_{k-1}})$, we can successively apply the above to drop $x_{j_{s-1}}$ for $s = k'+2, ..., M$ from the expression for $q_{R^*}(x_{n^*}|x_0)$. Since $X_{i_k} \perp_q X_{i_{k-1}}$,

$$q(x_{j_M}|x_{i_{k-1}}) = q(x_{i_k}|x_{i_{k-1}}) = q(x_{i_k})$$

so the DM is indifferent, and $j_{k'}$ and $j_{k'+1}$ are adjacent.

We now show that j_l is not adjacent to j_{k^*+1} for any $l < k^*$. Pick a pseudo-binary dataset q that factorizes as

$$q(x) = q(x_{\{i_0,j_0\}})q(j_{k^*+1}) \prod_{k'\neq k} q(x_{i_{k'}}|x_{i_{k'-1}}) \prod_{k'\notin\{k^*,k^*+1\}} q(x_{j_{k'}}|x_{j_{k'-1}}) \prod_{j'\notin N^\dagger} q(x_{j'})\times$$
$$\times\, q(j_{k^*}|j_{k^*+1}, j_{k^*-1})q(i_k|i_{k-1}, j_{M-1})$$

and where

$$q(\bar{E}_{i_k}|x_{\{i_{k-1},j_{M-1}\}}) = \begin{cases} \frac{3}{4} & \text{if } x_{\{i_{k-1},j_{m-1}\}} \in \bar{E}_{i_{k-1}} \times \bar{E}_{j_{M-1}} \\ \frac{1}{4} & \text{otherwise} \end{cases},$$

$$q(\bar{E}_{j_{k^*}}|x_{\{j_{k^*-1},j_{k^*+1}\}}) = \begin{cases} \frac{3}{4} & \text{if } x_{\{j_{k^*-1},j_{k^*+1}\}} \in \bar{E}_{j_{k^*-1}} \times \bar{E}_{j_{k^*+1}} \\ \frac{1}{4} & \text{otherwise} \end{cases},$$

$q(\bar{E}_{j_{k'}}|\bar{E}_{j_{k'-1}}) > q(\bar{E}_{j_{k'}}|\underline{E}_{j_{k'-1}})$ for $k' \neq k^*, k^*+1$, and $q(\bar{E}_{i_{k'}}|\bar{E}_{i_{k'-1}}) > q(\bar{E}_{i_{k'}}|\underline{E}_{i_{k'-1}})$ for $k' \neq k$. By construction, $X_l \perp_q X_{k^*+1}|X_{i_{k-1}}$ for all $l < k^*$. We can also calculate that $q(\bar{E}_{j_{k'}}|\bar{E}_{k'-1}) > q(\bar{E}_{j_{k'}}|\underline{E}_{j_{k'-1}})$ for $k' = k^*, k^*+1$, so $q_{R^*}(\bar{E}_{j_m}|\bar{E}_{j_0}) > q_{R^*}(\bar{E}_{j_m}|\underline{E}_{j_0})$. As in Lemma 2, this leads to $\rho(a,q) > \rho(b,q)$. Conclude that j_{k^*+1} and j_l are not adjacent for all $l < k^*$. □

A.4. **Proof of Proposition 1.** For part (1), if A separates then we have $\rho(a,q) = \rho(b,q)$ by definition. So suppose that A does not separate.

Let $(C_1, \ldots, C_m)$ be a MCJT representation for $R \cap N^*(R)^2$ with $0 \in C_1$ and $n^* \in C_m$, and $B_k = C_k \setminus A$. By the properties of the junction tree, it follows that C_k is adjacent to C_{k+1} for every k. Furthermore, $B_k \cap B_{k-1} \neq \emptyset$ for every k. If $B_k \cap B_{k-1} = \emptyset$ for some k, then A intersect every R-MAP from 0 to n^*, a contradiction given Lemma 1. By Lemma 5 and $A \perp_q A^c$, we note that $q_{R^*}(\bar{x}_{n^*} \mid a) > q(\bar{x}_{B_1\setminus\{0\}} \mid a) \prod_{k=2}^m q(\bar{x}_{B_k\setminus B_{k-1}} \mid \bar{x}_{B_k\cap B_{k-1}})$. As A does not separate, $B_k \cap B_{k-1} \neq \emptyset$ for every $k \in \{2, \ldots, m\}$ and so

$$\begin{aligned} q(\bar{x}_{B_k\setminus B_{k-1}}|\bar{x}_{B_k\cap B_{k-1}}) &= \frac{q(\bar{x}_{B_k}|a)q(a) + q(\bar{x}_{B_k}|b)q(b)}{q(\bar{x}_{B_k\cap B_{k-1}}|a)q(a) + q(\bar{x}_{B_k\cap B_{k-1}}|b)q(b)} \\ &> \frac{f(n,q(a))q(a)}{1-(1-q(a))f(n,q(a))} = 2^{-1/n}, \end{aligned}$$

since $f(n,q(a)) \leq q(\bar{x}_{B_k}|a) < 1$, $0 < q(\bar{x}_{B_k}|b) \leq 1 - f(n,q(a))$, and $q(b) = 1 - q(a)$.

As there are no more than n minimal separator sets, $q_{R^*}(\bar{x}_{n^*}|a) > \left(2^{-1/n}\right)^n > \frac{1}{2}$. Symmetrically, $q_{R^*}(\underline{x}_{n^*}|b) > \frac{1}{2}$. Therefore, $U_q(a) > \frac{1}{2}u(\bar{x}) + \frac{1}{2}u(\underline{x}) > U_q(b)$ so $\rho(a,q) > \rho(b,q)$.

For part (2), if i is adjacent to j, then $X_i \perp_q X_j$ and $X_I \perp_q X_{I^c}$ is sufficient to imply $\rho(a,q) = \rho(b,q)$ by definition. We show that if i is not adjacent to j, then the conditions on q imply that $\rho(a,q) > \rho(b,q)$. Since I is a selection from separators, Lemma 2 implies that $I = \{i_0, i_1, \ldots, i_m\}$ and $(i_0, i_1, \ldots, i_m)$ is an R-MAP. We show that $\mathrm{marg}_{i_l} q_R(\cdot|a)$ FOSD $\mathrm{marg}_{i_l} q_R(\cdot|b)$ for all l, provided that $\{i_l, i_{l+1}\} \neq \{i, j\}$ for all l. By Lemma 2, $\{i_l, i_{l+1}\} \neq \{i, j\}$ for all l if and only if i is not adjacent to j. For $l = 1$, this follows per hypothesis. Assume $\mathrm{marg}_{i_l} q_R(\cdot|a)$ FOSD $\mathrm{marg}_{i_l} q_R(\cdot|b)$. By (i, j)-MLRP, $\mathrm{marg}_{i_{l+1}} q(\cdot|x_{i_l})$ FOSD $\mathrm{marg}_{i_{l+1}} q(\cdot|x'_{i_l})$ when $x > x'$, so for any event $E = (-\infty, k) \cap X_{i_{l+1}}$, the function $f = x \mapsto q(E|x_{i_l})$ is decreasing. Then,

$$q_R(E_{i_{l+1}}|a) = \sum_{x \in \mathcal{X}_{i_l}} q_R(x_{i_l}|a) f(x) \leq \sum_{x \in \mathcal{X}_{i_l}} q_R(x_{i_l}|b) f(x) = q_R(E_{i_{l+1}}|b),$$

establishing that $\mathrm{marg}_{i_{l+1}} q_R(\cdot|a)$ FOSD $\mathrm{marg}_{i_{l+1}} q_R(\cdot|b)$. Inductively extend to i_m, and conclude $\rho(a, q) > \rho(b, q)$. The proof of Lemma 2 constructs a q satisfying these properties. $\square$

A.5. **Proof of Proposition 2.** Equation (2) follows from Lemma 5, Lemma 6, and Lemma 9. That $\mathcal{A}^\rho$ can be ordered uniquely follows from Lemma 2 and Lemma 7. $\square$

A.6. **Proof of Lemma 3.** Necessity is obvious, so suppose Axiom 3 holds. The result is clearly true when $|\mathcal{A}^\rho| \leq 4$, so suppose $|\mathcal{A}^\rho| = m > 4$.

We show that $\{0\} \not\propto^\rho \{n^*\}$. If not, then $\{0\} \propto A$ iff $A = \{n^*\}$ and $\{n^*\} \propto A$ iff $A = \{0\}$. Moreover, for any $A_2 \in \mathcal{A}^\rho \setminus \{\{0\}, \{n^*\}\}$, there exist distinct $A_1, A_3 \in \mathcal{A}^\rho \setminus \{\{0\}, \{n^*\}\}$ so that $A_1 \alpha^\rho A_2 \propto^\rho A_3$ and $A_1 \not\propto^\rho A_3$ by Axiom 3. So there are distinct $A_1, \ldots, A_k$ so that $A_1 \propto^\rho \cdots \propto^\rho A_k$ for $k = 3$. Suppose there are distinct $A_1, \ldots, A_k$ so that $A_1 \propto^\rho \cdots \propto^\rho A_k$ for $k \geq 3$. Since $|\{A' \in \mathcal{A}^\rho : A' \propto^\rho A_k\}| = 2$, there exists $A_{k+1} \neq A_{k-1}, \{0\}, \{n^*\}$ so that $A_{k+1} \propto^\rho A_k$. By Axiom 3, $A_{k+1} \neq A_l$ for $l < k - 1$. Therefore, there are distinct $A_1, \ldots, A_K$ so that $A_1 \propto^\rho \cdots \propto^\rho A_K$ for $K = k + 1$. Inductively, this is true for $K = m + 1$. But $|\mathcal{A}^\rho| = m$ by hypothesis, a contradiction.

Now $\{0\} = A_1^\rho$, and $\{0\} \propto^\rho A$ for some unique $A \neq \{n^*\}$. Label this A as A_2^ρ. Suppose that $A_1^\rho \propto^\rho \cdots \propto^\rho A_k^\rho$. As above, there exists $A_{k+1} \neq A_{k-1}^\rho, \{0\}$ so that $A_{k+1} \propto^\rho A_k^\rho$. By Axiom 3, $A_{k+1} \neq A_l^\rho$ for $l < k - 1$. Let $A_{k+1}^\rho = A_{k+1}$. If $A_{k+1}^\rho = \{n^*\}$ and $k + 1 \neq m$, then the same arguments as to why $\{0\} \not\propto^\rho \{n^*\}$ imply a contradiction. Therefore, if $k + 1 = m$, then the proof is complete. Otherwise, we can inductively find A_{k+2}^ρ. $\square$

A.7. **Necessity for Theorem 2.** Suppose that ρ has an Endogenous SCR (R, u, n^*). Proposition 2 implies that ρ is represented by R^ρ where $j\tilde{R}^\rho k$ for every distinct $j, k \in A_i^\rho \cup A_{i+1}^\rho$ for every i, $jR^\rho k$ whenever there exists $i \geq 0$ such that $j \in A_i^\rho$ and $k \in A_{i+1}^\rho \setminus A_i^\rho$, $R^\rho(j) \subset A_{i+1}^\rho \cup A_i^\rho$ when $j \in A_{i+1}^\rho \setminus A_i^\rho$, and $R^\rho(j) = \emptyset$ whenever $j \notin A_i^\rho$ for all i. Since

$$\int u(c)d\rho_{R^\rho}^S(c_{n^*}|a') \in \left[\min_{x\in\mathcal{X}_{n^*}} u(x), \max_{x\in\mathcal{X}_{n^*}} u(x)\right],$$

Axioms 1 and 2 hold. Axioms 4, 5, and 6 follow from Proposition 2 and continuity of the expected utility. □

A.8. **Sufficiency for Theorem 2.** Suppose that ρ satisfies the axioms. Lemma 3 implies $A_1^*, \ldots, A_{|\mathcal{A}|}^*$ exist, and so R^ρ as defined in necessity is a perfect, unconfounded, nontrivial DAG. We show that

$$\frac{\rho(a, S)}{\rho(b, S)} = \frac{\exp[\int_{\mathcal{X}_{n^*}} u(c)d\rho_{R^\rho}^S(c_{n^*}|a)]}{\exp[\int_{\mathcal{X}_{n^*}} u(c)d\rho_{R^\rho}^S(c_{n^*}|b)]} \tag{5}$$

for any $a, b \in S$ and any $S \in \mathcal{S}$. If so, then ρ has an Endogenous SCR (R^ρ, u, n^*) since $\sum_{a\in S} \rho(a, S) = 1$.

Pick any $S \in \mathcal{S}$ and any $a, b \in S$. Let $a'(y) = \rho_{R^\rho}^S(y|a)$ and $b'(y) = \rho_{R^\rho}^S(y|b)$ 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 Axiom 6. If $a' = b'$, then $\text{marg}_{A_1^*} a = \text{marg}_{A_1^*} b$, so $\rho(a, S) = \rho(b, S)$ by Axiom 4, and Equation (5) 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 $a'', b'' \in S'' \in \mathcal{S}$, $\frac{\rho(a'',S'')}{\rho(b'',S'')} \leq K$. For $S_m \cup S = \{s_1, \ldots, s_M\}$ and any $i, j \in \{1, \ldots, M\}$ with $i \neq j$, we have $\rho(s_i, S_m \cup S) \geq K^{-1}\rho(s_j, S_m \cup S)$. Then,

$$1 = \sum_{i\neq j} \rho(s_i, S_m \cup S) + \rho(s_j, S_m \cup S) \geq [(M-1)K^{-1} + 1]\rho(s_j, S_m \cup S)$$

so $\rho(s_j, S_m \cup S) \leq \frac{K}{M+K-1}$. In particular $\rho(c, S_m \cup S) \to 0$ for $c \in S$ as $m \to \infty$.

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

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

for every $x \in \mathcal{X}_{-0}$ since $\hat{a}(x_E|x_{A_i^\rho}) = a'(x_E|x_{A_i^\rho})$ for all $\hat{a} \in S_m$. This converges to $\rho^{S_1}(x_E|x_{A_i^\rho}) = a'(x_E|x_{A_i^*})$ 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 (5) holds for a, b. Since a, b, and S were arbitrary, ρ has an Endogenous SCR (R^ρ, u). □

A.9. **Proof of Proposition 3.** Suppose that ρ_i has a perfect SCR (R_i, u_i, n^*) for $i = 1, 2$ and that ρ_2 has a coarser model than ρ_1. Let N^* be the set of all i that are part of an R_2-MAP from 0 or an R_2-confounder to n^*. Define $R_3 = R_1 \cap N^* \times N^*$ and ρ_3 to have an SCR (R_3, u, n^*). Pick any q, and set $\hat{q} = x \mapsto q(x_{N^*}) \prod_{i \notin N^*} q(x_i)$. By construction for $i = 2, 3$, $q_{R_i}(x_{n^*}|do(x_0)) = \hat{q}_{R_i}(x_{n^*}|do(x_0))$ so $\arg\max_{S_q} \rho_i(\cdot, q) = \arg\max_{S_q} \rho_i(\cdot, \hat{q})$. Also, $\hat{q}_{R_1}(x_{n^*}|do(x_0)) = \hat{q}_{R_3}(x_{n^*}|do(x_0))$ so $\arg\max_{S_q} \rho_1(\cdot, \hat{q}) = \arg\max_{S_q} \rho_3(\cdot, \hat{q})$. By hypothesis, $\arg\max_{S_q} \rho_1(\cdot, \hat{q}) = \arg\max_{S_q} \rho_2(\cdot, \hat{q})$. Combining, $\arg\max_{S_q} \rho_2(\cdot, q) = \arg\max_{S_q} \rho_3(\cdot, q)$. Since q was arbitrary, R_3 represents ρ_2.

Conversely, let ρ_i have a perfect SCR (R_i, u_i, n^*) for $i = 1, 2$, $u_2 = u_1$ and $R_2 = R_1 \cap [N' \times N']$ for some $N' \subset N$. By Lemmas 4 and 7, $N^*(R_2)$ equals the indexes that appear an R_2-MAP from 0 or an R_2-confounder to n^*. Pick any $q \in \mathcal{Q}$ so that $X_i \perp_q X_{N \setminus \{i\}}$ for all $i \notin N^*(R_2)$. Since $N' \supset N^*(R_2)$, we also have $X_i \perp_q X_{N \setminus \{i\}}$ for all $i \notin N'$, so for every i and $x \in \mathcal{X}$,

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

Hence, $q_{R_1} = q_{R_2}$, and so $\rho_1(a, q) \geq \rho_1(b, q)$ for all $b \in S_q$ if and only if $\rho_2(a, q) \geq \rho_2(b, q)$ for all $b \in S_q$. □

Appendix B. Examples

B.1. **Violating Regularity.** If $\rho(\iota, S) = z$, then

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

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.

B.2. **Self-confirming choices.** 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 $u(1) = \lambda > u(0) = 0$,

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

For large enough λ, $\rho(a, S) \approx 0$ and $\rho(a, S) \approx 1$ are both solutions.

References

Alexander, Yotam and Itzhak Gilboa (2019), "Subjective causality."

Andre, Peter, Carlo Pizzinelli, Christopher Roth, and Johannes Wohlfart (2022), "Subjective models of the macroeconomy: Evidence from experts and a representative sample." *Review of Economic Studies*, Forthcoming.

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

Bohren, J. A. and D. Hauser (2021), "Learning with heterogenous misspecified models: Charaterization and robustness." *Econometrica*, 89, 3025–3077.

Brady, Richard L and John Rehbeck (2016), "Menu-dependent stochastic feasibility." *Econometrica*, 84, 1203–1223.

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

Cattaneo, Matias, Xinwei Ma, Yusufcan Masatlioglu, and Elchin Suleymanov (2020), "A random attention model." *Journal of Political Economy*, 128.

Cerreia Viogolio, Simone, Lars Peter Hansen, Fabio Maccheroni, and Massimo Marinacci (2021), "Making decisions under model misspecification." *working paper*.

Chambers, Christopher P, Tugce Cuhadaroglu, and Yusufcan Masatlioglu (2022), "Behavioral Influence." *Journal of the European Economic Association*, URL `https://doi.org/10.1093/jeea/jvac028`. Jvac028.

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

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

Eliaz, Kfir, Simone Galperti, and Ran Spiegler (2022), "False narratives and political mobilization." *arXiv preprint arXiv:2206.12621*.

Eliaz, Kfir and Ran Spiegler (2020), "A model of competing narratives." *American Economic Review*, 110, 3786–3816.

Eliaz, Kfir, Ran Spiegler, and Heidi C Thysen (2021), "Persuasion with endogenous misspecified beliefs." *European Economic Review*, 134, 103712.

Eliaz, Kfir, Ran Spiegler, and Yair Weiss (2020), "Cheating with models." *American Economic Review: Insights*.

Ellis, Andrew and Yusufcan Masatlioglu (2022), "Choice with endogenous categorization." *The Review of Economic Studies*, 89, 240–278.

Ellis, Andrew and Michele Piccione (2017), "Correlation misperception in choice." *American Economic Review*, 107, 1264–92.

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

Esponda, Ignacio and Demian Pouzo (2016), "Berk–nash equilibrium: A framework for modeling agents with misspecified models." *Econometrica*, 84, 1093–1130, URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA12609`.

Esponda, Ignacio and Emanuel Vespa (2018), "Endogenous sample selection: A laboratory study." *Quantitative Economics*, 9, 183–216.

Eyster, Erik and Matthew Rabin (2005), "Cursed equilibrium." *Econometrica*, 73, 1623–1672.

Frick, Mira, Ryota Iijima, and Yuhta Ishii (2020), "Misinterpreting others and the fragility of social learning." *Econometrica*, 88, 2281–2328.

Gul, Faruk and Wolfgang Pesendorfer (2006), "Random expected utility." *Econometrica*, 74, 121–146.

Hajek, Petr, Tomas Havranek, and Radim Jirousek (1992), *Uncertain Information Processing in Expert Systems*. CRC Press.

He, Kevin (2022), "Mislearning from censored data: The gambler's fallacy and other correlational mistakes in optimal-stopping problems." *Theoretical Economics*, 17, 1269–1312.

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

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

Jehiel, Philippe and Frédéric Koessler (2008), "Revisiting games of incomplete information with analogy-based expectations." *Games and Economic Behavior*, 62, 533–557.

Ke, Shaowei, Chen Zhao, Zhaoran Wang, and Sung-Lin Hsieh (2020), "Behavioral neural networks." *Working paper*.

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

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

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

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

Levy, Gilat, Ronny Razin, and Alwyn Young (2022), "Misspecified politics and the recurrence of populism." *American Economic Review*, 112, 928–62.

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

Lombrozo, Tania (2007), "Simplicity and probability in causal explanation." *Cognitive psychology*, 55, 232–257.

Lu, Jay (2016), "Random choice and private information." *Econometrica*, 84, 1983–2027.

Luce, R Duncan (1959), "Individual choice behavior."

Manzini, Paola and Marco Mariotti (2014), "Stochastic choice and consideration sets." *Econometrica*, 82, 1153–1176.

Montiel Olea, Jose Luis, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat (2021), "Competing models." *working paper*.

Pacer, M and Tania Lombrozo (2017), "Ockham's razor cuts to the root: Simplicity in causal explanation." *Journal of Experimental Psychology: General*, 146, 1761.

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

Pearl, Judea (1995), "Causal diagrams for empirical research." *Biometrika*, 82, 669–688.

Samuelson, L. and G. Mailath (2020), "Learning under diverse world views: Model based inference." *American Economic Review*, 110, 1464 – 1501.

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

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

Schumacher, Heiner and Heidi Christina Thysen (2022), "Equilibrium contracts and boundedly rational expectations." *Theoretical Economics*, 17, 371–414.

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

Sloman, Steven (2005), *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.

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

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

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

Tennant, Peter W G, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, Georgia D Tomova, Mark S Gilthorpe, and George T H Ellison (2020), "Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations." *International Journal of Epidemiology*, 50, 620–632, URL `https://doi.org/10.1093/ije/dyaa213`.

Verma, T. S. and Judea Pearl (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, 129–140.