SUBJECTIVE CAUSALITY IN CHOICE

ANDREW ELLIS AND HEIDI CHRISTINA THYSEN

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

1. Introduction

As every economics undergraduate knows, correlation does not imply causation. However in many economic environments, causal relationships are hard to determine, and an agent's perception of them is not observable. While most accept that ice cream consumption does not cause car theft, reasonable actors may disagree about whether the Phillips curve reflects causation or correlation. In this paper, we develop a theory in which an analyst can use the agent's behavior, in the form of a random choice rule, to identify her subjective causal model and to test whether misperceived causality explains her choices.

We motivate our results with a pair of examples. Consider a firm (the agent) faced with a series of hiring decisions between workers who differ according to a publicly-observable characteristic or type, e.g. their CV. The worker's type correlates with education, ability, and productivity, which are revealed to the firm only after it hires him or her. The firm uses a subjective causal model to infer the productivity of each

Date: May 2021.

Department of Economics, London School of Economics and Political Science. Email: a.ellis@lse.ac.uk and h.c.thysen@lse.ac.uk.

from the characteristics of its past hires, and employs each type in proportion to its expected production.[1] A university (the analyst) may want to know whether or not the firm thinks that education *causes* productivity, conditional on ability. If the firm thinks so, then relaxing admission standards does not lower the value of a degree, keeping the curriculum fixed; otherwise, it might because of a decrease in signaling value. Because the university cannot observe the firm's causal model, she must infer it from the frequently each type of worker is hired. Our result allows the university to test whether its hiring decisions (behavior) are consistent with a belief that education causes productivity.

For a second example, consider a group of firms (the agents), each a local monopolist in distinct but observationally-identical markets, choosing what price to set for spirits. They base their pricing decision on what they perceive the effect of an increase in price on quantity demanded to be, which they infer from a common causal model and the history of joint realizations of prices, quantities demanded, and other covariates with demand. A policy maker (the analyst) sets the level of an excise tax on spirits. Our results show how the policy maker can identify the firms' perception of the causal effect of an increase in price and so calculate the incidence of each level of tax from the fraction of firms who set each price (behavior) in different market conditions.[2]

This paper provides a theoretical methodology for identifying an agent's subjective causal model from her behavior. We study a decision maker (DM) who chooses from a set of actions, each determining a probability distribution over a vector of variables. Her random choice rule has a *subjective causality representation* if she learns the consequences of her actions from the data generated by her past choices using a subjective causal model described by a directed acyclic graph (DAG), and then chooses each action with a frequency proportional to its expected utility. The choices form a personal equilibrium: how frequently she chooses each action affects her inferences, which in turn affects the likelihood each is chosen. We show how to identify the DM's subjective causal model (her DAG) and preferences from her behavior. Then, we turn to the question of how to test whether a random choice rule has a subjective causality representation and provide necessary and sufficient conditions for one to exist.

[1] Perhaps she also (privately) observes a match-specific component, such as interview performance, relevant to productivity but independent of the other variables.

[2] While we focus on the individual interpretation of random choice for exposition, our results apply equally well with a group interpretation as in this second example.

The DM mistakes correlation for causation when her causal model is misspecified. This creates a particular challenge for the analyst: the DM's own behavior may create a correlation between two variables that she misinterprets as a causal effect, the magnitude of which affects how likely she is to choose each action. As an example, suppose that a firm thinks that education alone causes productivity when it is actually caused by ability. If some high-ability workers are more likely to have high-education than all low-ability ones, then the perceived return to education increases with the fraction of high-education workers hired by the firm. This incentivizes the firm to hire even more highly-educated workers, reinforcing the effect; see Section 2.3 for a formal treatment.

Our first main result identifies the causal model that explains the DM's behavior. This entails revealing her perceived chains of causality; for example, the firm may think that who it hires affects the education of the workforce, which in turn affects its productivity. We show that the DM's perceived causal chains identify all the relevant variables and the causal relationships between them. We reveal the set of variables that makes up each by observing that if *every* chain passes through a set of variables, then independence between those variables and the others implies indifference between all actions. For instance, the above firm is equally likely to choose either of a pair of workers whose productivity and ability are independent of their level of education. Hence if the DM is *not* indifferent when the variables *outside* a given set are independent, then she thinks that a chain passes through that set. We then determine the direction of causality by varying the correlation between the variables in the same causal chain.

While a large literature in economics focuses on empirically determining causality (e.g. Card (1999)), an agent's perception of causal relationships, regardless of its validity, can affect the result of a policy intervention. For instance, a firm that appears to offer workers with a particular trait a lower wage may do so because it dislikes employing workers with the trait even when the trait has no impact on productivity (taste-based discrimination). Alternatively, it may offer a lower wage because the trait is correlated with another attribute, such as education, that the firm thinks affects productivity (statistical discrimination). Policies that attempt to remedy the former, such as affirmative action for or awarding scholarships to students with the trait, may do nothing for the latter.[3]

[3] 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 misspecified causal model can explain the DM's behavior by providing necessary and sufficient conditions for a random choice rule to have a subjective causality representation. The axioms link her perceptions of alternatives, as inferred from our first result, to her behavior. Holding her perception constant, her behavior conforms closely to Logit with an expected utility Luce index (henceforth, Logit-EU). Put another way, her choices are inconsistent with Logit-EU only when her inferences change. For example, the axioms require that the DM chooses two actions with the same relative frequency from two menus in which she infers that each has the same distribution, and that she is equally likely to choose any actions with identical distributions over a subset of variables that she thinks is a sufficient statistic for the outcome. However, her evaluation of alternatives varies across menus as her perception of them changes. As a consequence, she may violate a number of standard axioms, including a necessary condition for a random utility representation known as regularity.

An agent with a subjective causality representation perceives her options differently than the analyst does. The result places testable restrictions on her behavior in spite of the information gap. Thus, it establishes that misspecified causality provides enough discipline on how her beliefs are distorted to be testable; without any discipline on belief distortion, testing would be impossible. More broadly, this paper adapts decision theory 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 towards providing testable implications for the growing literature studying agents with misspecified models, especially Spiegler (2016), Spiegler (2020), Eliaz and Spiegler (2018), Eliaz et al. (2019), Eliaz et al. (2020), and Schumacher and Thysen (2020) that all use versions of the subjective causality representation.[4]

The paper proceeds as follows. The next subsection reviews the related literature. Section 2 introduces our setup and model, formalizes our running example, and then discusses alternative interpretations of the model. Section 3 reveals the DM's subjective causal model from her choices. Theorem 1 shows that the DM's behavior reveals the

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

minimal causal chains in her model and that these identify the relevant portions of the DAG. Section 4 describes the axioms that characterize the model. Theorem 2 proves that they are necessary and sufficient for a subjective causality representation.

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

Pearl (1995) argued for using and analyzing DAGs to understand causality. A large literature (e.g., Cowell et al., 1999, Koller and Friedman, 2009, Pearl, 2009) develops and applies this approach for probabilistic and causal inference. The typical exercise either uses a DAG to estimate the causal effect of a particular intervention or to infer which DAG(s), if any, are consistent with a given dataset.[5] Schenone (2020) introduces the DAG approach to causality into a decision theory framework. In the model, an agent expresses preference over act-causal-intervention pairs. For instance, the DM expresses a preference over which of two workers to hire, identical except one of whom has been forced to obtain exactly 11 years of education. He provides necessary and sufficient conditions for the agent's beliefs to result from applying the "do-operator" to intervened variables for a fixed DAG and prior. The DAG is identified from the behavior. This approaches is complementary with the one taken by this paper. It is mainly concerned with a normative definition of causality as a manifestation of rationality and consequently takes interventions as observable. In contrast, this paper uses DAGs to capture flaws in reasoning and focuses on identification from only choices without interventions.

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

More generally, our paper is related to the decision theory literature studying DMs who misperceive the world. Lipman (1999) studies a DM who may not understand all the logical implications of information provided to her. Ellis and Piccione (2017) develop a model where agents misperceive the correlation between actions. Kochov (2018) models an agent who does not accurately foresee future consequences of her action. In all three, the misperception is fixed and unaffected by the agent's 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 indices, 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 Manzini and Mariotti (2014), Brady and Rehbeck (2016), and Cattaneo et al. (2020) models of limted attention. This paper uses random choice identify features of explicit boundedly rational behavior.

2. Model

2.1. **Setting.** Each action a determines a distribution over a payoff-relevant consequence and n covariates. The ith covariate takes a value in $\mathcal{X}_i$ and the consequence belongs to the set $\mathcal{X}_{n+1}$. For a non-empty-set S, let $\Delta(S)$ is the set of finite support probability distributions over S. Each action is a member the set $\mathcal{X}_0 = \Delta(\prod_{i=1}^{n+1} \mathcal{X}_i)$, and it will be convenient to denote $\mathcal{X}_{-0} = \prod_{i=1}^{n+1} \mathcal{X}_i$ and $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_{-0}$. We require that $\mathcal{X}_{n+1}$ is a compact subset of a topological space with $|\mathcal{X}_{n+1}| \geq 2$, and take $\mathcal{X}_i = \mathbb{R}$ for simplicity.[6]

The DM's (stochastic) choice of action determines the distribution of a random vector $X = (X_0, X_1, \ldots, X_{n+1})$. If the DM chooses $a \in \mathcal{X}_0$, then $a(x_1, \ldots, x_{n+1})$ is the probability that $X_i = x_i$ for every $i \in \{1, \ldots, n+1\} \equiv N^*$. We identify the distribution over actions with the 0th random variable. The last index $n+1$ denotes consequence. The set $N = \{1, \ldots, n\}$ indexes the set of covariates. By convention,

[6] We can take each $\mathcal{X}_i$ to be an arbitrary set with $|\mathcal{X}_i| \geq \min\{|\mathcal{X}_{n+1}|, |\mathbb{N}|\}$. This would increase the notational complexity, particularly for Definition 3, Proposition **??**, and Lemma 10, but not substantively change the arguments.

capital letters refer to variables and lower case letters to realizations. We denote by $\operatorname{marg}_J p$ the marginal distribution of p on the variables indexed by J. With slight abuse of notation, we sometimes identify the action $a \in \mathcal{X}_0$ with the element of $\Delta\mathcal{X}$ that has a marginal on $\mathcal{X}_{-0}$ equal to a and attaches probability 1 to $X_0 = a$.

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

$$\mathcal{S} = \left\{ S \subset \mathcal{X}_0 : \prod_{j \in N} \operatorname{supp}(\operatorname{marg}_j a) = \operatorname{supp}(\operatorname{marg}_N b) \text{ for all } a, b \in S \text{ and } S \text{ is finite} \right\}.$$

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

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

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

Note $\rho(\cdot, S)$ is a distribution over actions whereas ρ^S is a distribution over $\mathcal{X}$.[7]

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

2.2. **Subjective Causality Representation.**

A directed acyclic graph (DAG) over a set M is an asymmetric, acyclic binary relation $R \subset M \times M$, where iRj denotes $(i,j) \in R$. A DAG R over $\{0, 1, \ldots, n+1\}$ describes the DM's perception of causality. Here, iRj indicates that the DM thinks that X_i causes X_j and corresponds to a directed edge in a graph. We often write $R(i)$ for the indexes of the variables that cause X_i

[7] Cerreia-Vioglio et al. (2019) denote ρ^S by $\overline{\rho(S)}$ but interpret the DM as preferring $\overline{\rho(S)}$ to any lottery in $co(S) \subset \Delta\mathcal{X}$.

according to R, termed the *parents* of i. For a DAG R and $p \in \Delta\mathcal{X}$, p_R is the probability distribution so that $p_R(x) = \prod_{j=0}^{n+1} p(x_j|x_{R(j)})$. See Spiegler (2016) for a discussion of how the DAG maps into various behavioral biases and for further discussion.

A DAG R has free-will if 0 is ancestral and $n+1$ is a descendant of 0: there is no $i \in N^*$ with $iR0$ and $0Ri_1Ri_2R\dots R\ (n+1)$ for some $i_1, i_2, \dots \in N$. A free-will DAG indicates two things. First, the action is not conditioned on any of the other variables. Second, there is a channel through which the choice of action can influence the distribution over consequences. Say that (i,j,k) is an R-v-collider if iRk, jRk, $j\not R i$, and $i\not R j$. A DAG R is perfect if there are no R-v-colliders. We focus on perfect DAGs because otherwise the perceived marginal distribution of some variable may diverge from its true distribution (Spiegler, 2017).

Definition 1. The random choice rule ρ has a *subjective causality representation (SCR)* if there exists a free-will DAG R and a continuous, non-constant u so that

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

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

The representation corresponds to the following "as if" procedure. The DM maximizes expected utility but with a potentially incorrect perceptions of the distribution of consequences resulting from her actions. She derives this perception from her choices from S, which determine a "dataset" ρ^S of the frequency of each realization of the random vector X. She applies her causal model to infer the conditional distribution of consequences and covariates for each action, and updates the probability of consequence c if she takes the action a to $\rho_R^S(c_{n+1}|a)$. She then chooses each action with a probability proportional to the exponential of her perceived expected utility, or equivalently, picks an action if it has the highest utility after adding iid extreme-value shocks to each action's expectation. For comparison, a ρ has a *Logit-EU* representation if there is a continuous, non-constant u so that for every $a \in S$ and $S \in \mathcal{S}$,

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

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

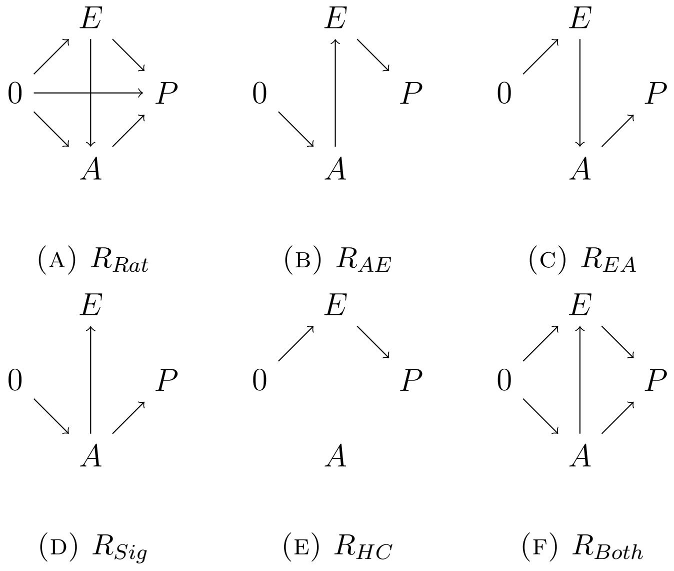
An SCR is a personal equilibrium (Köszegi and Rabin, 2006): the DM maximizes expected utility given her beliefs that depend on her choices. It is easy to show that an equilibrium exists for any $S \in \mathcal{S}$ using Brouwer's fixed point theorem. For menus with more than one, we place no restrictions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

2.3. **Running Example.** Return to the firm example to illustrate our framework. Hiring a given type of worker corresponds to an action. The first variable represents ability ($A = 1$), the second education ($E = 2$), and the third productivity ($P = 3$). Throughout, each takes one of two values $H > L$. For example, the vector (H, L, H) represents a worker who has ability H, education L, and productivity H, and $a(H, L, H)$ indicates the probability that a type-a worker has those characteristics. The firm prefers to hire high productivity workers.

Figure 1 gives some possible DAGs for the firm. Each represents a different theory of causation. A firm represented by R_{HC} thinks that education, and education alone, causes productivity, one represented by R_{Sig} thinks that ability causes both education and productivity, and one represented by R_{EA} thinks that education causes ability and that ability causes productivity. In contrast, one represented by R_{Rat} is rational and always correctly infers the distribution. With any DAG other than R_{Rat}, the firm potentially misperceives the joint distribution. For instance, the R_{Sig}-firm necessarily believes ability to be independent of productivity.

The behavior of an agent may endogenously create correlations that she misinterprets as causation. For instance, consider a firm whose behavior has a SCR (R_{HC}, u). There are three equally productive types of workers, ι, π, ν, but P is independent of E for type-ι workers, positively correlated for type-π, and negatively correlated for type-ν. As it hires type-π workers more often, the correlation between education and productivity increases. The firm mistakes this correlation for causation, so it perceives an increased return to education and hires the more-educated types more often.

This can lead to a violation of *regularity*, the requirement that $\rho(a, S) \geq \rho(a, S')$ whenever $a \in S \subset S'$. Every random choice rule represented by a random utility



(a) R_{Rat} (b) R_{AE} (c) R_{EA}

(d) R_{Sig} (e) R_{HC} (f) R_{Both}

Figure 1. Possible DAGs in Running Example

model satisfies regularity, so the class of SCRs and RUMs do not coincide. To see why the violation occurs, observe that when the firm decides between only ι- and π-type workers, education is positively correlated with productivity. Since it mistakes the correlation for causation, it hires the type with highest expected education more often than the other. However, when it chooses between all three, the type-ν workers may cancel out or even reverse the perceived positive relationship between education and productivity. When this effect is strong enough, it can lead to an increase in the probability of choosing the less-educated type.

Formally, assume that $\iota(H_E H_P) = \iota(H_E L_P) = \frac{1}{2}$, $\nu(H_E L_P) = \nu(L_E H_P) = \frac{1}{2}$, and $\pi(L_E L_P) = \pi(H_E H_P) = \frac{1}{2}$.[8] Suppose that $u(L) = 0$ and $u(H) = 6$. Consider menus $S = \{\iota, \pi\}$ and $T = \{\iota, \nu, \pi\}$.[9] If $\rho(\iota, S) = z$, then

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

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

[9] We note that $S, T \notin \mathcal{S}$, so this example is 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 a with $a' = (1-\epsilon)a + \epsilon b$ where $\epsilon > 0$ is small enough and $b(y) = \frac{1}{8}$ for each $y \in \{H, L\}^3$.

and since $\iota(H_E) > \pi(H_E) > 0$, we have $1 > z > \frac{1}{2}$. Then, $\frac{1}{2} < \rho^S(H_P|H_E) < \frac{2}{3}$, so $\rho^S_R(H_P|\pi) < \frac{1}{3}$, while $\rho^S_R(H_P|\iota) > \frac{1}{2}$. 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 productivity is independent of the firm's action conditional on education according to R, the firm is indifferent between two worker types whose distribution over education is identical. Therefore, $\rho(\nu,T) = \rho(\pi,T) = \gamma$ since the two have the same distribution over education. Then,

$$\rho^T(H_P|H_E) = \frac{(1-2\gamma)\frac{1}{2}+\gamma\frac{1}{2}+\gamma(0)}{(1-2\gamma)+2\gamma\frac{1}{2}} = \frac{1}{2} = \frac{(1-2\gamma)(0)+\gamma(0)+\gamma\frac{1}{2}}{(1-2\gamma)(0)+2\gamma\frac{1}{2}} = \rho^T(H_P|L_E),$$

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

2.4. **Interpretations of model.** Our main interpretation of an SCR is that it describes a DM endowed with a fixed causal model that she utilizes to infer the consequences of her choices. Her model maps the distribution ρ^S of the random vector X into a perceived distribution ρ^S_R of each action. We are agnostic as to why the DM infers the overall distribution rather than just the relationship between action and consequence. She may recognize that the data set is endogenous and attempt to adjust for any endogeneity by applying her causal model. Or perhaps she (incorrectly) anticipates the arrival of additional information or the possibility of taking other actions. She may think it easier or quicker to learn the stronger correlations between the covariates that make up a causal chain than the weaker correlation between her action and the outcome.

Alternatively, the DM may have limited data access (Spiegler, 2017). In this interpretation, she only considers or observes the distributions of several overlapping subsets of variables and then extrapolates to form a distribution over all variables using the principle of insufficient reason. Formally, she uses the distribution that maximizes entropy subject to matching the marginal distribution over each subset of variables in her database. Identifying her subjective causal model corresponds to identifying the subsets of whose distribution she matches. In the running example, consider a firm with access to two datasets, $\{\{0,E\},\{E,P\}\}$. That is, she does not directly observe the correlation between the type she hires and productivity, instead only observing the relationships between type and education and betweeen education and productivity. This would occur, for instance, when the firm relies on reports from the HR director screening potential new hires and the factory foreman overseeing the workers. In this case, the extrapolation procedure leads to the same behavior a firm who has a SCR (R_{HC}, u).

The SCR can also describe a DM economizing on the information she stores. The DAG R embeds conditional independence assumptions that reduce the number of moments needed to reconstruct the distribution. In the running example, a firm with two worker-types available can store all the relevant information according to the DAG R_{HC} using only 6 parameters.[10] In contrast, it would require $2^4 - 1 = 15$ parameters to record the probability of each possible realization of X without these assumption.

For a final interpretation, we note that when ρ has an SCR, ρ^S_R minimizes Kullback-Lieber divergence from ρ^S among all the probability distributions on $\mathcal{X}$ that are consistent with R. Then, ρ represents a single agent Berk-Nash equilibrium (Esponda and Pouzo, 2016) with extreme-value errors. As in that model, we can intrepret the behavior as the steady state of a learning process with a set of parameters (probability distributions) that do not include the "true" one.[11]

3. Identifying a Subjective Causal Model

In this section, we identify the DM's subjective causal model from her choice behavior. In a perfect, free-will DAG R, information flows along the chains of causal relations from actions to consequences. Formally, a causal chain is an *R-Active Path from* 0 *to* $n+1$ *(or, a R-AP)*: a finite sequence of variables $(i_1, i_2, \dots, i_m)$ with $i_1 = 0$, $i_m = n+1$, and $i_j R i_{j+1}$ for every $j < m$. This represents a chain of causal reasoning: according to R, variable 0 causes i_1, which in turn causes i_2, and so on, ending with a cause of the outcome variable. A *minimal R-AP, or R-MAP,* is a R-AP that cannot be made shorter. That is, $(i_1, \dots, i_m)$ is an R-MAP if it is a R-AP and $i_j \not R i_{j'}$ whenever

[10] For $S = \{a, b\}$, the numbers $p(a)$, $p(H_A)$, $p(H_E|a)$, $p(H_E|b)$, $p(H_P|H_E)$, and $p(H_E|L_E)$ fully determine p_R.

[11] Whether there exists a learning process that necessarily converges to the steady state captured by SCR remains an open question.

$j' \neq j+1$. The main result shows that these minimal causal chains suffice to identify a perfect, free-will DAG.

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

Only variables that appear in at least one R-MAP affect the DM's behavior. For a firm represented by R_{HC}, A does not belong to an R-MAP, so another DAG R' that represents ρ must contain R_{HC} but may add links to and from A, provided that doing so does not create a cycle, a R'-MAP, or a v-collider. The relationships in causal chains determine all other key causal relationships. While there may be others, their direction is immaterial for the DM's choices. For instance, a firm represented by R_{Both} may also be represented by a DAG R' that reverses the link between A and E.

We argue next that if R and R' both represent ρ, then the R-MAPs and R'-MAPs must coincide. Then, we construct a candidate DAG that represents ρ whenever it has a perfect SCR. First, we show that a set of variables contains a causal chain only if the DM expresses a preference between actions in a menu for which all other variables are statistically independent. To determine the order of variables in a causal chain, we establish that, ceteris paribus, she perceives a stronger relationship between her action and her payoff when choosing from a menu of actions consistent with her model than one that is not.

We begin by revealing the sets containing the covariates in an active path. For $K \subseteq N$, we say that X_K is *independent within* $S \in \mathcal{S}$, written $X_K \perp_S X_{N^* \setminus K}$, if $\operatorname{marg}_K a = \operatorname{marg}_K b$ for any $a, b \in S$ and $a(x_K, x_{N^*\setminus K}) = a(x_K)a(x_{N^*\setminus K})$ for every $x \in \mathcal{X}_{-0}$ and every $a \in S$. If X_K is independent within S, then regardless of how the DM chooses from S, X_K is independent of the other random variables in the resulting joint distribution. An experimenter can create this independence by intervening to set their values without changing the others in a randomized controlled trial. Alternatively, we explore allowing for an exogenous dataset in Section 5 where independence can be induced using unavailable actions.

The following definitions are key to our identification.

Definition 2. The set $K \subseteq N$ *separates* if $\rho(a, \{a, b\}) = \frac{1}{2}$ whenever $X_K \perp_{\{a,b\}} X_{N^* \setminus K}$, and $I \subseteq N$ *is a ρ-MAP* if it is a minimal set so that $N \setminus I$ does not separate.

Whether a set of covariates separates, and so whether it is a ρ-MAP, depends solely on ρ. In the randomized controlled trial interpretation, the intevention leads to separation when it controls for all the variables through which the DM thinks the treatment can affect the outcome. Then, the perceived average treatment effect equals zero.

When R represents ρ, each ρ-MAP equals the set of covariates in a R-MAP.

Lemma 1. *If ρ has an SCR (R, u), then $I \subseteq N$ contains the covariates in an R-MAP if and only if $N \setminus I$ does not separate.*

That is, $N \setminus I$ does not separate only if there exists an R-MAP $(i_1, \ldots, i_m)$ so that $\{i_2, \ldots, i_{m-1}\} \subset I$. Therefore, a ρ-MAP I is exactly the covariates in an R-MAP. Since a ρ-MAP is defined by ρ and not R, every R' that represents ρ must have an R'-MAP with covariates I. When $\emptyset$ is a ρ-MAP, the DM thinks that actions directly cause outcomes, so $0R(n + 1)$ and there are no other ρ-MAPs or R-MAPs. Similarly when N is a ρ-MAP, she thinks that there is exactly one causal chain from actions to outcomes, and that chain includes every covariate. In this case, there are also no other ρ-MAPs.

To illustrate, consider a firm that is equally likely to choose each type of worker whenever education is unrelated to the other variables, i.e. $\{E\}$ separates. One can construct such menus where the type of worker hired is positively correlated with ability and productivity. Since the firm nevertheless hires every type equally frequently, it must think that the correlations between type, ability, and productivity are all spurious. Thus, its choices reveal that the firm thinks that every causal chain includes E. By contrapositive, if it is not equally likely to hire every type for some menu where E is independent, i.e. $\{E\}$ does not separate, then its causal model has *some* causal chain that does not pass through $\{E\}$. In other words, $\{A\}$ contains the covariates in an R-MAP, and the firm's subjective DAG R includes either the chain $0RARP$ or $0RP$.

In the running example, the ρ-MAPs suffice to distinguish between firms with any of the DAGs in Figure 1 except R_{EA} and R_{AE}.[12] We establish that the DM's behavior reveals the order of causation, allowing us to distinguish the other two DAGs as well.

Lemma 2. *If ρ has SCRs (R, u) and (R', u'), then R agrees with R' on $I \cup \{0, n+1\}$ for every ρ-MAP I.*

The two Lemmas establish that the set of R-MAPs coincides with the set of R'-MAPs when R and R' both represent ρ. To illustrate, consider a firm for which $\{E, A\}$ is a ρ-MAP. By Lemma 1, we know that its DAG is either R_{EA} or R_{AE}.[13] We compare two menus of workers, one consistent with R_{EA} and the other consistent with R_{AE} but that are otherwise equivalent. The firm identifies a stronger correlation between the worker's type and productivity, and hires the more productive type more often when the menu is consistent with its subjective DAG.

More specifically, we compare $\rho(a, \{a, b\})$ and $\rho(a', \{a', b'\})$, with types defined as follows. Type a (a') workers are more likely to have high education (ability) than type b (b'), and type a' (b') workers are as likely to have high ability as type a (b) workers are to have high education:

$$a'(H_A) = a(H_E) > b(H_E) = b'(H_A).$$

Education (ability) alone determines ability (education):

$$\begin{aligned} &a'(H_E|H_A) = a(H_A|H_E) = b(H_A|H_E) = b'(H_E|H_A) \\ >&a'(H_E|L_A) = a(H_A|L_E) = b(H_A|L_E) = b'(H_E|L_A). \end{aligned}$$

Ability (education) alone determines productivity:

$$\begin{aligned} &a(H_P|H_A, x_E) = b(H_P|H_A, x'_E) = a'(H_P|H_E, x_A) = b'(H_P|H_E, x'_A) \\ >&a(H_P|L_A, x_E) = b(H_P|L_A, x'_E) = a'(H_P|L_E, x_A) = b'(H_P|L_E, x'_A) \end{aligned}$$

[12] The ρ-MAPs are $\emptyset$ for R_{Rat}, $\{E\}$ for R_{HC}, $\{A\}$ for R_{Sig}, both $\{E\}$ and $\{A\}$ for R_{Both}, and $\{E, A\}$ for either R_{AE} or R_{EA}.

[13] While neither of the DAGs are the most natural causal models, we stick with the running example for consistency of exposition. For a more reasonable economic example, a DM who believes that the Phillips curve is a causal relationship may either believe that inflation causes low unemployment or that low unemployment causes inflation.

for any $x, x' \in \{H, L\}$. Moreover, both correlations are positive. Note that the relationships resulting from choice in $\{a, b\}$ are consistent with R_{EA}, and that those from choice in $\{a', b'\}$ are consistent with R_{AE}.

While a and a' (as well as b and b') workers are equally productive, the firm does not realize this. If the firm is more likely to employ type a than a', then its choices reveal that it thinks that education causes ability; otherwise, it thinks that ability causes education. Suppose that the firm's model is R_{EA}, so it decomposes the causal effect of its hiring on productivity into the effects of hiring on education, of education on ability, and of ability on productivity. The relationships between hiring and education as well as between ability and productivity are weaker when facing $\{a', b'\}$ than when facing $\{a, b\}$, while that between ability and education is the same. In fact when facing $\{a', b'\}$, education is a sufficient statistic for productivity, so adding ability in the causal chain garbles the relationship between hiring and productivity. As a consequence, the firm perceives a smaller causal effect of her action in $\{a', b'\}$ than in $\{a, b\}$, and so is more likely to choose the worse type in the former than the latter.[14]

3.1. Constructing the revealed DAG.

Theorem 1 shows that the minimal causal chains characterize the subjective DAG. In this subsection, we construct a revealed DAG R^ρ directly from ρ that represents it whenever it has a perfect SCR.[15] We use this revealed DAG in our axiomatization to determine when the DM's perception of her actions remains constant, but any other perfect, free-will DAG R with the same R-MAPs would work equally well.

The key tool to construct R^ρ is a *Markovian family of menus for* $I \subseteq N$. Such a family of menus is indexed by their potential orderings, bijections $\pi : \{0, \ldots, |I|+1\} \to I \cup \{0, n+1\}$ with $\pi(0) = 0$ and $\pi(|I|+1) = n+1$; call the set of such bijections the indexes for I. For any index π for I, let $R_\pi = \{(\pi(k), \pi(k+1)) : k = 0, ..., |I|\}$.

Definition 3. The family of menus $\{\{a^\pi, b^\pi\}\}_\pi$ is *Markovian* for $I \subseteq N$ if π ranges over the indexes for I and

(1) $a^\pi = a^\pi_{R_\pi}$, $b^\pi = b^\pi_{R_\pi}$, and $a^\pi \neq b^\pi$;
(2) for every $i = 1, ..., |I|$, $a^\pi(x_{\pi(i+1)}|x_{\pi(i)}) = b^\pi(x_{\pi(i+1)}|x_{\pi(i)}) = a^{\pi'}(x_{\pi'(i+1)}|x_{\pi'(i)})$;

[14] This is a consequence of the logit structure: the difference in utilities maps to the difference in choice probabilities.

[15] R^ρ is defined regardless of whether ρ has a perfect SCR.

(3) for every pair of indexes π, π' and $d \in \{a, b\}$, $\operatorname{marg}_{\pi(1)} d^{\pi} = \operatorname{marg}_{\pi'(1)} d^{\pi'}$;
(4) $|\operatorname{supp}\operatorname{marg}_i \left(\frac{1}{2}a^{\pi} + \frac{1}{2}b^{\pi}\right)| = 2$ for each $i \in I \cup \{n+1\}$; and
(5) for each $i < |I|$, $\operatorname{marg}_{\pi(i)} a^{\pi}$-a.e. x, x', and $\operatorname{marg}_{\pi(i+1)} a^{\pi}$-a.e. y, y' with $x > x'$ and $y > y'$, $a^{\pi}(y_{\pi(i+1)}|x_{\pi(i)}) > a^{\pi}(y_{\pi(i+1)}|x'_{\pi(i)})$.

Each of the actions defines a first-order Markov chain on the variables indexed by I. The indexes vary the order of the chain but not the transition probabilities. The initial distribution of the first variable, $X_{\pi(1)}$, is the same for every a^{π} and the same for every b^{π}, but is not identical for a^{π} and b^{π}. For $j > 1$, the distribution of $X_{\pi(j)}$ conditional on the realization of $X_{\pi(j-1)}$ is the same for a^{π} and b^{π}. Each variable has a binary support. A high value of $X_{\pi(i)}$ increases the likelihood of a high value for the $X_{\pi(i+1)}$ for each $i \leq |I|$. Note that $\{\{a, b\}, \{a', b'\}\}$ above is a Markovian family of menus for $\{E, A\}$.

We have defined a ρ-MAP in terms of whether its complement separates. This requires observing choices from every menu where those variables are independent. One can instead test whether a set is a ρ-MAP using choices from a small number of carefully selected menus. Specifically, for each $I \subseteq N$ let $S_I = \{a^{\pi}, b^{\pi}\}$ be a member of a Markovian family of menus for I with

$$\int_{\mathcal{X}_{n+1}} \left[a^{\pi}\left(c_{n+1}|x_{\pi(|I|)}\right) - a^{\pi}\left(c_{n+1}|x'_{\pi(|I|)}\right)\right] u(c)dc \neq 0 \tag{1}$$

for some x, x' so that $a^{\pi}\left(x_{\pi(|I|)}\right), a^{\pi}\left(x'_{\pi(|I|)}\right) > 0$.[16] If ρ has an SCR (R, u), then $I \subseteq N$ is a ρ-MAP if and only if $\rho(a, S_I) \neq \frac{1}{2}$ for any $a \in S_I$ and $\rho(a, S_{I\setminus\{i\}}) = \frac{1}{2}$ for each $i \in I$ and $a \in S_{I\setminus\{i\}}$.

The *subjective ordering of a ρ-MAP I* is the index π^* that maximizes the probability of choosing the better action in some Markovian family of menus for I for which Equation (1) holds for all π.[17] For any indexes π, π', the actions a^{π} and $a^{\pi'}$ have the same distribution over consequences, as do b^{π} and $b^{\pi'}$. As in the illustration, the DM perceives a smaller effect of her choice when her causal model is inconsistent with the menu.

[16] The utility index u is easily identified from ρ.
[17] This is equivalent to maximizing the difference in choice frequencies between actions.

We use these observations to identify the causal relationships that must belong to the DM's DAG. First, *i is a directly revealed cause of j, or $i\bar{R}^{\rho}j$,* if $iR_{\pi^*}j$ where π^* is the subjective ordering of a ρ-MAP. There are additional causal relationships that can be derived from the directly revealed causes. We illustrate this for a ρ represented by R in Figure 2. Observe that the R-MAPs are $(0,1,3,4)$ and $(0,2,4)$. If ρ also has a SCR (R',u'), then R' includes $2R'3$ in addition to the directly revealed causes. If $3\not{R}'2$ and $2\not{R}'3$, then $(2,3,4)$ is a R'-v-collider. If $3R'2$, then either $(0,3,2)$ is a R'-v-collider, or $0R'3$ and so $(0,1,3,4)$ is not an R'-MAP. In this case, we say that 2 is an indirectly revealed cause of 3. More generally, *i is an indirectly revealed cause of j* if there are variables h and k so that i and j both must cause k, h must cause i, and h must not cause nor be caused by j. This occurs whenever there are variables $i_0,...,i_m,j_0,...,j_M$ so that $i_0=0$, $i_m=i$, $j_0=j$, $j_M=n+1$, $i_k\bar{R}^{\rho}i_{k+1}$ for all $k<m$, $j_k\bar{R}^{\rho}j_{k+1}$ for all $k<M$, $i\bar{R}^{\rho}j_k$ for some k, and $i_k\not{\bar{R}}^{\rho}j_{k'}$ for all k' and all $k<m$. Finally, *i is a revealed cause of j, written $i\hat{R}^{\rho}j$,* if i is a direct or indirect revealed cause of j.

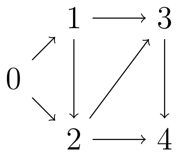


FIGURE 2. X_2 must cause X_3 but $\{2,3\}\not\subseteq I$ for any ρ-MAP I

Definition 4. The *revealed DAG R^{ρ}* satisfies $iR^{\rho}j$ if and only if $i\hat{R}^{\rho}j$ or there is $k\in N$ so that (i,j,k) is a $\hat{R}^{\rho}$-v-collider and $i<j$.

The revealed DAG links common parents by making the one with a lower index the cause.[18] Note that R^{ρ} is defined regardless of whether ρ actually has an SCR. Its interpretation as the DM's causal model relies on ρ having a perfect SCR.

Corollary 1. *If ρ has a perfect SCR, then ρ has an SCR (R^{ρ},u) for some u.*

Hence, R^{ρ} is natural candidate for the DM's subjective causal model. In particular, R^{ρ} reveals the contexts in which the DM makes the same inferences about the

[18] Resolving ambiguity according to any linear order would work equally well.

consequence of an action. This observation is key for the foundations we provide in the subsequent section.

4. Behavioral Foundations for Subjective Causality

Based on our identification in the previous section, this section characterizes the random choice rules that have a subjective causality representation. Simple axioms that relate the DM's behavior in different menus are shown to be equivalent to her choices having a perfect SCR. Throughout, the properties are illustrated in the context of the running example. For these purposes, we assume that the firm's revealed DAG R^ρ equals R_{AE}, i.e. it thinks that its hiring decisions cause ability, ability causes education, and education causes productivity.

The first axiom is standard.

Axiom 1 (Basic)**.** For any $S \in \mathcal{S}$ and $a \in S$, $\rho(a, S) > 0$, and there exists $S \in \mathcal{S}$ and $a, b \in S$ with $\rho(a, S) \neq \rho(b, S)$.

The DM chooses every available alternative with positive probability and does not always choose every alternative with equal frequency.

The first new axiom says that if the DM perceives two actions to be the same, then she chooses each with the same probability.

Axiom 2 (Indifferent If Identical Immediate Implications, I5)**.**
For $F = \{k \in N^* : 0R^\rho k\}$, if $\text{marg}_F\, a = \text{marg}_F\, a'$ and $a, a' \in S$, then $\rho(a, S) = \rho(a', S)$.

The covariates directly caused by the DM's action are a sufficient statistic for her perception of it. That is, if two actions have the same distribution over these covariates, then she perceives them to have the same consequence distribution. She is therefore indifferent between any two actions with identical immediate implications according to her subjective causal model. In the running example, it implies that the firm is equally likely to hire any two types of workers having the same distribution over ability, regardless of how they differ in their distributions over other variables. In the firm's mind, ability suffices to determine the other variables, including productivity.

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

Axiom 3 (Bounded Misperception)**.** The Luce ratio is bounded: $\sup_{S,a,b\in S} \frac{\rho(a,S)}{\rho(b,S)} < \infty$.

The Luce ratio indicates the strength of the DM's preference between two actions. Since the set of consequences is compact, there is a best and worst outcome. These provide a natural limit to how much she can prefer one action to another. The axiom thus bounds the size of the mistakes that the DM can make.[19] In the running example, the relative frequency of hiring any pair of workers is bounded above by the relative frequency with which it would hire a (known) high-productivity worker relative to a (known) low-productivity worker.

The following definitions help state the remaining axioms.

Definition 5. A lottery $p \in \Delta\mathcal{X}$ (sequence of lotteries $p_1, p_2, \cdots \in \Delta\mathcal{X}$) *leads to (approximately) the same R-inferences* as $q \in \Delta\mathcal{X}$, written $p|R = q|R$ ($p_m|R \to q|R$), if $p(y_k|y_{R(k)}) = q(y_k|y_{R(k)})$ ($p_m(y_k|y_{R(k)}) \to q(y_k|y_{R(k)})$) for q-a.e. $y \in \mathcal{X}_{-0}$ and $k \in N^*$ such that $0 \not R k$.
A menu $S \in \mathcal{S}$ is *R-Markov* if $a|R = b|R$ and $a_R = a$ for all $a, b \in S$.

To interpret the definitions, consider a DM with causal model $R = R_{AE}$. For menus $S, T \in \mathcal{S}$, ρ^S leads to the same R-inferences as ρ^T if the relationship between ability and education in addition to the relationship between education and productivity is the same for both the datasets ρ^S and ρ^T. Consequently, the causal effects of ability on education and of education on productivity are perceived to be the same. She infers the same effect on productivity of hiring type $a \in S \cap T$ in either of the two menus. Similarly for menus $S_m, T \in \mathcal{S}$, $(\rho^{S_m})_m$ leads to approximately the same R-inferences as ρ^T if the two above relationships are arbitrarily close for large m. When m is big enough, she infers almost the same effect on productivity of hiring type $a \in S_m \cap T$ in either of the two menus. When facing an R-Markov menu, the DM's model is correct, and she perceives every alternative correctly.

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

Axiom 4 (Luce's Choice Axiom Given Inferences, LCI)**.** For any $S, S_1, S_2, \cdots \in \mathcal{S}$ with $a, b \in S_m \cap S$ for each m: if $\rho^{S_m}|R^\rho \to \rho^S|R^\rho$, then $\frac{\rho(a,S_m)}{\rho(b,S_m)} \to \frac{\rho(a,S)}{\rho(b,S)}$.

[19] The axiom is implied by Independence and Luce's Choice Axiom for the Logit-EU model.

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 is "close" to holding whenever the DM's inferences about the relationships captured by R^ρ are also close. In particular, when her inferences are the same, the choice axiom holds: if $S, S' \in \mathcal{S}$, $a, b \in S \cap S'$, and $\rho^S|R^\rho = \rho^{S'}|R^\rho$, then $\frac{\rho(a,S')}{\rho(b,S')} = \frac{\rho(a,S)}{\rho(b,S)}$. Put another way, the statistical relationships captured by R^ρ suffice for determining the DM's choices.

In the running example, suppose that given the firm's choices when facing $\{a, b\}$ and $\{a, b, c\}$, the statistical relationship between ability and education is the same as is that between education and productivity. Then, its perception of the productivity of type-a and type-b workers is the same given either the dataset $\rho^{\{a,b\}}$ or the dataset $\rho^{\{a,b,c\}}$. LCI requires that relative probability of hiring a to hiring b is the same when facing either $\{a, b\}$ or $\{a, b, c\}$.

The next two axioms ensure that ρ has a Logit-EU representation for menus where the DM's inferences are correct. Axiom 4 guarantees that Luce's Choice Rule holds for all such menus.

Axiom 5 (R^ρ-Markov Independence)**.** If $\{\alpha p + (1-\alpha)r, r\}, \{\beta p + (1-\beta)r, r\} \in \mathcal{S}$ are R^ρ-Markov, then

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

For R^ρ-Markov menus, the DM's perception of each alternative is correct. For any such menu, the independence axiom holds. Moreover, the relative probability of choosing $\alpha p + (1-\alpha)r$ to r is log linear in α because of the Logit functional form.

When the DM perceives her actions correctly, her behavior is suitably continuous in the consequence distribution.

Axiom 6 (R^ρ-Markov Continuity)**.** If $\{p, q\}, \{p_1, q_1\}, \{p_2, q_2\}, \cdots \in \mathcal{S}$ are all R^ρ-Markov, $\operatorname{marg}_{n+1} p_m \to \operatorname{marg}_{n+1} p$, and $\operatorname{marg}_{n+1} q_m \to \operatorname{marg}_{n+1} q$, then

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

Only the consequence matters for payoffs. So when the DM's inferences are correct, what determines her preferences over lotteries are their distributions over that variable. If the distributions over consequences for two pairs of lotteries are close, then their relative choice frequencies are also close. In particular, if two correctly-perceived pairs of lotteries have the same consequence distributions, then they have the same Luce ratio. That is, if $\text{marg}_{n+1} p' = \text{marg}_{n+1} p$, $\text{marg}_{n+1} q' = \text{marg}_{n+1} q$, and both $\{p, q\}$ and $\{p', q'\}$ are R^ρ-Markov, then $\frac{\rho(p',\{p',q'\})}{\rho(q',\{p',q'\})} = \frac{\rho(p,\{p,q\})}{\rho(q,\{p,q\})}$.

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

Theorem 2. *A random choice rule ρ has a perfect subjective causality representation if and only if R^ρ is a perfect, free-will DAG and ρ satisfies Basic, I5, Bounded Misperception, LCI, R^ρ-Markov Independence, and R^ρ-Markov Continuity.*

The result highlights the connection between SCR and the Logit-EU model. Notice that if Axioms 1, 4, 5, and 6 hold when the part of their hypotheses involving R^ρ are dropped, the choice rule has a Logit-EU representation. The axioms thus indicate the circumstances under which the choice rule does not diverge from Logit. I5 says two alternatives are chosen with same probability whenever they coincide on the distribution of the R^ρ-immediate implications of the action, whereas Logit-EU requires coincidence on the consequence distribution. Bounded Misperception gives a maximum deviation in the relative choice frequencies. LCI restricts violations of Luce's Choice Axiom to when inferences change. R^ρ-Markov Independence and Continuity show independence and continuity hold whenever alternatives are perceived correctly.

We outline the proof for sufficiency here, and defer a formal proof to the appendix. We first show that the choice rule has a Logit-EU representation when restricted to R^ρ-Markov menus. Then, we relate the DM's choices from S to those from a R^ρ-Markov copy of S, S'_1. That is, for every $a \in S$, there is an $a' \in S'_1$ so that $a'(\cdot) = \rho^S_{R^\rho}(\cdot|a)$ and S'_1 is R^ρ-Markov. Our goal is to show that for any $a, b \in S$, a and b are chosen with the same relative frequency in S as a' and b' are in S'_1. To do so, we add distinct alternatives to S'_1 to form a nested sequence of menus $(S'_m)_{m=1}^\infty$ while maintaining that each S'_m is R^ρ-Markov. Bounded Misperception implies that the probability of choosing anything in S from $S'_m \cup S$ goes to zero as the number of alternatives in S'_m goes to infinity. In particular, the inferences that the DM makes from $S'_m \cup S$ approach those she makes from S_1', which are in turn equal to those she makes from S. LCI implies that the relative frequency with which a' and b' are chosen from $S_m' \cup S$ converges to that for a' and b' in S_1'. Moreover, a and a' (as well as b and b') are chosen from $S_m' \cup S$ with the same probability. Applying LCI another time, we see that a and b are chosen with the same relative frequency in S as a' and b' are in S_1', completing the proof.

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

Corollary 2. *A random choice rule ρ has a Logit-EU representation if and only if $0 \; R^\rho \; (n+1)$ and ρ satisfies Basic, I5, Bounded Misperception, LCI, R^ρ-Markov Independence, and R^ρ-Markov Continuity.*

Note that $0 \; R^\rho \; (n+1)$ if and only if $\emptyset$ is a ρ-MAP, and if so, $\emptyset$ is also the only ρ-MAP as noted earlier. When $0 \; R^\rho \; (n+1)$, those are the only nodes related by R^ρ, so $p|R^\rho = q|R^\rho$ for any $p, q \in \Delta\mathcal{X}$. Consequently, LCI becomes the Luce Choice Axiom. It is easy to verify that the remainder of the axioms ensure that the Luce Index has the desired form.

5. Discussion and Extensions

This section concludes the paper by looking at some implications of the model and considering how our modeling decisions affect our results. We discuss some of the biases in reasoning that arise with a subjective causal model. Then, the behavior of two DMs with nested causal models are compared. We then examine how to extend our analysis to eliminate stochasticity and the endogeneity of the dataset.

5.1. Biases.

A DM with SCR may be subject to illusion of control: she may overestimate her ability to control events. As Langer (1975, p. 311) writes, "In skill situations there is a causal link between behavior and outcome.... Success in luck or chance activities is apparently uncontrollable. The issue of present concern is whether or not this distinction is generally recognized. The position taken here is that it is not." The firm in Section 2.3 is subject to illusion of control when choosing from $\{\pi, \iota\}$. Its action does not affect productivity, yet it would be willing to pay a premium to choose one worker over another.

While our model is formally static, it can be viewed as the steady state of a learning process. With such an interpretation, her initial behavior can be self-confirming, leading to status quo bias (Samuelson and Zeckhauser, 1988), omission bias (Ritov and Baron, 1992), and other biases. Status quo bias is a tendency toward "maintaining one's current or previous decision." The related omission bias indicates a preference for inaction. Formally, such biases occur in SCR when there exist multiple personal equilibria. Then, a DM who begins choosing according to one particular personal equilibrium tends to stay there. Congruence bias (Wason, 1960) refers to a tendency not to test alternative hypotheses. For similar reasons to the above, the DM's behavior may conform to a sub-optimal personal equilibrium. She fails to experiment with other actions sufficiently frequently to push her beliefs towards the better equilibrium.

A misspecified causal model can lead to an agent treating Berkson's bias as an actual causal relationship. Berkson's bias, also known as collider bias, is a statistical artifact that leads to an increase in correlation between two covariates when conditioning on a common consequence. The following example illustrates the criticism by Griffith et al. (2020) of Miyara et al. (2020)'s finding that smoking may prevent symptomatic Covid-19. Suppose that Serious Covid (C) and Smoking (S, the agent's action) both decrease Lung Functionality (L), and that C and poor L both increase the odds of Hospitalization (H). However, the agent thinks that S affects L, L affects C, and C alone affects H. This agent can become convinced that Smoking decreases risk of Serious Covid, even if in reality Smoking and Serious Covid are independent or even moderately positively correlated. See also Spiegler (2016)'s dieter's dilemma.

5.2. **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 (2005), 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.[20] 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?

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

Definition 6. Say that ρ_2 *has a coarser model than* ρ_1 if $\rho_1(\cdot, S) = \rho_2(\cdot, S)$ whenever $X_{N \setminus N^*(\rho_2)} \perp_S X_{N^*(\rho_2)}$.

Consider DM1 represented by ρ_1 and DM2 represented by ρ_2. As revealed by Theorem 1, DM2 considers the variables outside $N^*(\rho_2)$ irrelevant for determining the consequence of her action. The condition says that if those variables are actually irrelevant when choosing from S, i.e. they are independent of the other variables, then the two DMs choose identically from S. This ensures that whenever DM2 thinks a variable is relevant, so does DM1.

Proposition 1. *Let ρ_1 and ρ_2 have perfect SCRs. If ρ_2 has a coarser model than ρ_1, then $R^{\rho_2} = R^{\rho_1} \bigcap [N^*(\rho_2)]^2$ and the utility indices are equal up to addition of a constant. The converse holds up to the selection of a personal equilibrium.*

The result shows that the comparison reveals when the models of two DMs are nested. Specifically, they agree on the causal relationship between any two variables that both consider relevant and on the desirability of outcomes. However, they may disagree on which variables are relevant, with DM1 considering any variable relevant that DM2 does. In this sense, DM2 has a coarser model than DM1.

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

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

$$\prod_{j \in N} \operatorname{supp}(\operatorname{marg}_j q) = \operatorname{supp}(\operatorname{marg}_N q).$$

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

Definition 7. The augmented random choice rule ρ^* has an *Exogenous SCR (ESCR)* if there exists a free-will DAG R and a continuous, non-constant u so that

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

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

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

5.4. **Deterministic choice.**

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

Formally, we suppose that the DM's behavior is described by a choice correspondence $c : \mathcal{S} \rightrightarrows \Delta(\mathcal{X}_0)$ where $p(S) = 1$ for all $p \in c(S)$ and $c(S) \neq \emptyset$ for each $S \in \mathcal{S}$.[21] For any $p \in \Delta(\mathcal{X}_0)$, write $p^X \in \Delta(\mathcal{X})$ for the resulting dataset, i.e. p^X is the lottery so that $p^X(a, y) = p(a)a(y)$ for every $(a, y) \in \mathcal{X}$.

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

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

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

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

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

Again, Theorem 1 continues to hold. The arguments that establish necessity require some changes. As above, Lemma 1 requires only minor alterations: we replace "$\rho(\{a, b\})(a) = \frac{1}{2}$" with "$c(\{a, b\}) = \Delta\{a, b\}$" in Definition 2. Similarly, we replace "ρ-MAP" with "c-MAP" by making the natural substitutions.

Adapting Lemma 2 requires more work. Consider a Markovian family of menus for a c-MAP J. The construction of the family of menus ensures that any misperception preserves the ordinal preference between a^{π} and b^{π}, though it may affect the perceived magnitude of their difference. If a^{π} would be better than b^{π} if correctly perceived, then the personal equilibrium for each menu in the family is a deterministic choice of a^{π}. To reveal her subjective ordering, we augment the Markovian family with an outside option that she necessarily perceives "almost" correctly but that is worse than a correctly-perceived a^{π}. Whenever π disagrees with the DM's DAG and the outside option is good enough, there is a personal equilibrium where she chooses the outside option. In Appendix B.1, we formalize these the arguments.

Appendix A. Proofs from Main Text

A.1. **Proof of Lemma 1.**

Lemma 3. *For any $p \in \Delta\mathcal{X}$ and free-will DAG R, if there is no from 0 to j for any $j \in J$, then $p_R(x_J|a) = p_R(x_J)$ for p-a.e. $a \in \mathcal{X}_0$. Moreover, if there is no from 0 to j for every $j \in J$ within the set L and $X_i \perp_{\{p\}} X_{-i}$ for all $i \in N \setminus L$, then $p_R(x_J|a) = p_R(x_J)$ for p-a.e. $a \in \mathcal{X}_0$.*

Proof. Let R^l be a linear order that completes R, i.e. iRj implies iR^lj and R^l is complete and transitive (R^l exists since R is acyclic). Relabel so that $0R^l1R^l2\dots$, and note that $R(i) \subset \{1, \dots, i\}$. We prove the result by induction. Let (IHm) be "If there is no R-AP from 0 to j for all $j \in J$ where $J \subset \{1, \dots, m\}$, then $p_R(x_J|a) = p_R(x_J)$." Consider $m = 1$. Then, $J = \{1\}$. If there is no R-AP from 0 to 1, then clearly $p_R(x_1|a) = p(x_1)$. So (IH1) is true.

Assume (IHm) is true. Consider any $J' \subset \{1, \dots, m+1\}$ for which there is no R-AP from 0 to j for all $j \in J'$. If $m+1 \notin J'$, the claim follows from IHm, so consider $J' = J \bigcup \{m+1\}$ for $J \subset \{1, \dots, m\}$ and no R-AP from 0 to j for all $j \in J'$. Let $\tilde{J} = R(m+1) \setminus J$ and $\hat{J} = R(m+1) \bigcap J$. Let

$$\mathcal{X}(\bar{x}_{J'}, a) = \{x \in \mathcal{X} : x_0 = a,\ x_{J'} = \bar{x}_{J'}\}.$$

Then,

$$\begin{aligned} p_R(\bar{x}_{J'}|a) &= \sum_{x \in \mathcal{X}(\bar{x}_{J'}, a)} \frac{\prod_{k=1}^m p(x_k|x_{R(k)})}{p(x_0 = a)} p(\bar{x}_{m+1}|x_{R(m+1)}) \\ &= \sum_{x \in \mathcal{X}(\bar{x}_{J'}, a)} p_R(\bar{x}_J, x_{(J')^c}|a) p(\bar{x}_{m+1}|x_{R(m+1)}) \\ &= \sum_{y_{\tilde{J}} \in \mathcal{X}_{\tilde{J}}} p_R(\bar{x}_J, y_{\tilde{J}}|a) p(\bar{x}_{m+1}|\bar{x}_{\hat{J}}, y_{\tilde{J}}). \end{aligned}$$

Since there is no R-AP from 0 to $m+1$, there is no R-AP from 0 to any $j \in J \bigcup \tilde{J}$, and clearly $J \bigcup \tilde{J} \subset \{1, \dots, m\}$. By (IHm), $p_R(\bar{x}_J, y_{\tilde{J}}|a) = p_R(\bar{x}_J, y_{\tilde{J}})$, so $p_R(\bar{x}_{J'}|a) = p_R(\bar{x}_{J'})$. This completes the induction step, (IHm) so is true for all m, which in turn establishes the result.

For the moreover, assume that $X_{N \setminus L} \perp_{\{p\}} X_L$. Let Q be R where every link between a node in $N \setminus L$ and a node in L is dropped. Then, $p_Q(x) = p_R(x)$ for every

x since when $k \in L$ we can write

$$\begin{aligned} p(x_k|x_{R(k)}) &= p(x_k|x_{R(k)\cap L}, x_{R(k)\setminus L}) \\ &= p(x_k|x_{R(k)\cap L}) = p(x_k|x_{Q(k)}). \end{aligned}$$

and if $k \notin L$ we can write

$$\begin{aligned} p(x_k|x_{R(k)}) &= p(x_k|x_{R(k)\cap L}, x_{R(k)\setminus L}) \\ &= p(x_k|x_{R(k)\setminus L}) = p(x_k|x_{Q(k)}). \end{aligned}$$

Apply the preceding argument to p_Q to get the result. □

Proof of Lemma 1. For necessity, consider a set $J = N \setminus I$ that does not separate. For contradiction, assume that I does not contain an R-AP. Then, all R-APs from 0 to $n+1$ go through J. Since J does not separate, there exists $B = \{a, b\}$ where $X_j \perp_B X_{-j}$ for all $j \in J$ and $\rho^B(a) > \frac{1}{2}$. Let Q be R where all links involving a node in J are dropped and $q = \rho^B$. By construction, $q(x_j x_{-j}) = q(x_j)q(x_{-j})$ for all $x \in \mathcal{X}$ and $j \in J$. Then, $q_Q(x) = q_R(x)$ for every $x \in \mathcal{X}$ since we can write

$$\begin{aligned} q(x_i|x_{R(i)}) &= q(x_i|x_{R(i)\cap J}, x_{R(i)\setminus J}) \\ &= q(x_i|x_{R(i)\setminus J}) = q(x_i|x_{Q(i)}). \end{aligned}$$

Since

$$q_R(\bar{x}_{n+1}|a) = \sum_{x \in \mathcal{X}(\bar{x}_{n+1}, a)} q_R(x_{R(n+1)}|a) q(\bar{x}_{n+1}|x_{R(n+1)}),$$

where $\mathcal{X}(\bar{x}_J, a) = \{x \in \mathcal{X} : x_0 = a,\ x_J = \bar{x}_J\}$ and $R(n+1)$ satisfies the hypothesis of Lemma 3 for Q, $\rho^B_R(x_{n+1}|a) = \rho^B_R(x_{n+1}|b)$ for any a, b in B. But this contradicts $\rho(a, B) > \frac{1}{2}$.

For sufficiency, consider any $I = \{i_1, \ldots, i_m\} \subset J$ so that $0 R i_1 R i_2 R \ldots R i_m R (n+1)$. We want to show that $N \setminus J \equiv J^c$ does not separate. The claim is true if there exists a, b with $\rho(a, \{a, b\}) \neq \frac{1}{2}$ and $X_{j'} \perp_{\{a,b\}} X_{-j'}$ for every $j' \in N \setminus J$. If $0\ R\ (n+1)$, then this is trivial since $c(x_{n+1}) = [\alpha a + (1-\alpha)b]_R(x_{n+1}|c)$ for every $x \in \mathcal{X}_{n+1}$, $\alpha \in (0,1)$, and $c \in \{a, b\}$. Otherwise, pick a Markovian family of menus $\{\{a^\pi, b^\pi\}\}_\pi$ for I (Definition 3) where $\int a^\pi(x_{n+1})u(x)dx \neq \int b^\pi(x_{n+1})u(x)dx$ for each index π, noting that $X_{j'} \perp_{\{a^\pi, b^\pi\}} X_{-j'}$ holds for every $j' \in N \setminus J$ and each index π. Let π^* be the index such that $\pi^*(j) = i_j$ for $j \leq m$, and label $a = a^{\pi^*}$ and $b = b^{\pi^*}$. For any $\alpha \in (0,1)$, let

$q \in \Delta(\mathcal{X})$ equal $\alpha x + (1-\alpha)b$. Note that $q = q_R$, so $q_R(x_{n+1}|c) = c(x_{n+1})$ for $c = a, b$. This establishes that $\rho(a, \{a, b\}) \neq \frac{1}{2}$, completing the proof. □

A.2. **Proof of Proposition ??.** When a probability measure $p \in \Delta\mathcal{X}$ has binary support on each of its components, let i denote the event "variable i equals the larger outcome" and $-i$ denote the event "variable i equals the smaller outcome."

Lemma 4. *Fix a DAG R and a set $J = \{j_1, \ldots, j_m\}$ with $0Rj_1Rj_2R\ldots j_mR(n+1)$ and no other links in J, for any Markovian family of menus $\{\{a^\pi, b^\pi\}\}_\pi$ for J and π^* be the bijection so that $\pi^*(i) = j_i$ for each i. For any index π and any $\alpha > 0$, if $a(j_1) > b(j_1)$, $\rho = \alpha a + (1-\alpha)b$, and $p^\pi = \alpha a^\pi + (1-\alpha)$, then*

$$p^\pi_R(n+1 \mid -i) < p^\pi_R(n+1 \mid i) \text{ and } p^\pi_R(n+1 \mid i) \leq \rho_R(n+1 \mid i)$$

when i is the first index such that $\pi(i) \neq \pi^*(i)$.

Proof of Lemma 4. We prove by induction on $|J|$. The basic case is when $|J| = 1$. Pick any index π, and note that $\pi = \pi^*$, so the Lemma is trivially holds given the construction.

Induction Step: (IH) Suppose that the Lemma is true for any R', J' and $\pi^{*\prime}$ as in the statement with $|J'| \leq m$. Consider a R, J and π^* as in the statement with $|J| = m+1$. Relabel the variables in J such that π^* is the identity map on $1, \ldots, |J|$. Let $a = a^{\pi^*}$ and $b = b^{\pi^*}$, and $\rho = \alpha a + (1-\alpha)b$. This construction implies strict MLRP, and so $a(i) > b(i)$ for each $i \in J \cup \{n+1\}$. Also by construction,

$$p^\pi(\pi(i)) = \rho(i) \;\&\; p^\pi(x_{\pi(i+1)} \mid x_{\pi(i)}) = \rho(x_{\pi(i+1)} \mid x_{\pi(i)})$$

for every index π and every $i \in J$.

Pick any index π; if $\pi = \pi^*$ the conclusion holds, so assume $\pi \neq \pi^*$. Let $i^* = \min_{i \in J} i \neq \pi(i)$. Define a new DAG $\hat{R} = R \cap \{i^*+1, ..., |J|, n+1\}^2$, an index π' on $J' = J \setminus \{1, \ldots, i^*\}$ so that $\pi'(i) = |\{j \in J' : \pi^{-1}(j) \leq i\}|$, and actions $\hat{a}^{\pi'}, \hat{b}^{\pi'}$ so that

$$\operatorname{marg}_{\{n+1\} \cup J'} \hat{c}^{\pi'} = \operatorname{marg}_{\{n+1\} \cup J'} c^\pi$$

and attach probability one to $\bar{x}_{N \setminus J'}$ for some $\bar{x} \in \mathcal{X}$. Clearly $\hat{a}^{\pi'}, \hat{b}^{\pi'}$ are part of a Markovian family of menus $\{\{\hat{a}^{\hat{\pi}}, \hat{b}^{\hat{\pi}}\}\}_{\hat{\pi}}$ for J' and $\hat{a}^{\pi'}(\pi'(1)) > \hat{b}^{\pi'}(\pi'(1))$. Letting π^{**}

satisfy $\pi^{**}(i) = \pi^*(i+i^*)$ set

$$q = \alpha\hat{a}^{\pi'} + (1-\alpha)\hat{b}^{\pi'} \ \& \ \rho' = \alpha\hat{a}^{\pi^{**}} + (1-\alpha)\hat{b}^{\pi^{**}}.$$

Then, IH implies that $q_{\hat{R}}(n+1|i^*+1) > q_{\hat{R}}(n+1|-(i^*+1))$ and that $\rho'_{\hat{R}}(n+1|i^*+1) \geq q_{\hat{R}}(n+1|i^*+1)$. Since $p^\pi(i^*+1|i^*) > p^\pi(i^*+1|-i^*)$ by applying strict MLRP multiple times and $p^\pi(x_{i+1}|x_i) = q(x_{i+1}|x_i)$ for $i > i^*$, we have $p^\pi_R(n+1|x_{i^*})$ equal to

$$p^\pi(i^*+1|x_{i^*})[q_{\hat{R}}(n+1|i^*+1) - q_{\hat{R}}(n+1|-(i^*+1))] + q_{\hat{R}}(n+1|-(i^*+1)).$$

Conclude that $p^\pi_R(n+1|i^*) > p^\pi_R(n+1|-i^*)$. Moreover, let k be such that $\pi(i^*+k) = i^*$. For $i \geq 1$ we have $\rho'(\pi^{**}(i+1)|x_{\pi^{**}(i)})$ equals

$$\begin{cases} p^\pi(\pi(i^*+i+1) \mid x_{\pi(i^*+i)}) = \rho(i^*+i+1 \mid x_{(i^*+i)}) & \text{if } i+1 < k \\ p^\pi(\pi(i^*+i+2) \mid x_{\pi(i^*+i)}) = \rho(i^*+i+2 \mid x_{(i^*+i)}) & \text{if } i+1 = k \\ p^\pi(\pi(i^*+i+2) \mid x_{\pi(i^*+i)}) = \rho(i^*+i+2 \mid x_{(i^*+i)}) & \text{if } i+1 > k \end{cases}$$

so $\rho'_{\hat{R}}(n+1|\pi^{**}(i)) = \rho_R(n+1|\pi^{**}(i))$ for any i. Letting $\hat{\imath}$ be the first index so that $\pi'(\hat{\imath}) \neq \pi^{**}(\hat{\imath})$,

$$\begin{aligned} &\rho_R(n+1|i^*) \\ =&\rho(\pi^{**}(\hat{\imath})|i^*)[\rho'_{\hat{R}}(n+1|\pi^{**}(\hat{\imath})) - \rho'_{\hat{R}}(n+1|-\pi^{**}(\hat{\imath}))] + \rho'_{\hat{R}}(n+1|-\pi^{**}(\hat{\imath})) \\ >&\rho(\pi^{**}(\hat{\imath})|i^*)[q_{\hat{R}}(n+1|\pi^{**}(\hat{\imath})) - q_{\hat{R}}(n+1|-\pi^{**}(\hat{\imath}))] + q_{\hat{R}}(n+1|-\pi^{**}(\hat{\imath})) \\ =&p^\pi(n+1|i^*) \end{aligned}$$

since $\rho'(n+1) = q(n+1)$ by construction and $\rho'_{\hat{R}}(n+1|\pi^{**}(\hat{\imath})) > q_{\hat{R}}(n+1|\pi^{**}(\hat{\imath}))$ by IH. Conclude the Lemma is true when $|J| = m+1$, establishing the Lemma by induction for all J. □

Proof of Lemma 2. Let ρ have a SCR (R, u), and J be a ρ-MAP. We show that there is a unique π^* so that for any Markovian family of menus $\{\{a^\pi, b^\pi\}\}_\pi$ for J, either $\rho(a^\pi, \{a^\pi, b^\pi\}) = \frac{1}{2}$ for all π or π^* maximizes $|\rho(a^\pi, \{a^\pi, b^\pi\}) - \rho(b^\pi, \{a^\pi, b^\pi\})|$ across all indexes π. This π^* has the same ordering as R on J, so any other DAG the represents ρ must agree with R on $J \cup \{0, n+1\}$.

Relabel the variables in J to be the first $m = |J|$ integers, such that π^* is the identity map on J, and $a = a^{\pi^*}$ and $b = b^{\pi^*}$. Write $a(i+1 \mid i) = r_i$ and $a(i+1 \mid i) = s_i$, noting $r_i > s_i$ for every i by definition.

Consider the Markovian family of menus $\{\{a^\pi, b^\pi\}\}_\pi$ for $J \in\subset N$. Consider $a = a^{\pi^*}$ and $b = b^{\pi^*}$ where $a(1) = x = 1 - a(-1)$, $b(1) = y = 1 - b(-1)$, and $x > y$ (WLOG after relabeling since $x = y$ implies $a = b$).

Fix any index $\pi \neq \pi^*$. Since $a^\pi(\pi(1)) - b^\pi(\pi(1)) = (x - y)$ and $a^\pi(\pi(k+1)) - b^\pi(\pi(k+1)) = [a^\pi(\pi(k)) - b^\pi(\pi(k))](r_k - s_k)$ for $k \geq 1$, $[a^\pi(i) - b^\pi(i)] < [a(i) - b(i)]$ for every $i \in J$. Furthermore,

$$\rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid c^\pi) = c^\pi(i)\rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid i) + c^\pi(-i)\rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid -i)$$

and

$$\rho_R^{\{a,b\}}(n+1 \mid x_0 = c) = c(i)\rho_R^{\{a,b\}}(n+1 \mid i) + c(-i)\rho_R^{\{a,b\}}(n+1 \mid -i).$$

Let i be the first index such that $\pi(i) \neq i$. Combining the above with the results from Lemma 4 we get

$$\begin{aligned}\rho_R^{\{a,b\}}(n+1 \mid a) - \rho_R^{\{a,b\}}(n+1 \mid b) &= [a(i) - b(i)][\rho_R^{\{a,b\}}(n+1 \mid i) - \rho_R^{\{a,b\}}(n+1 \mid -i)]\\ &> [a^\pi(i) - b^\pi(i)][\rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid i) - \rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid -i)]\\ &= \rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid a^\pi) - \rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid b^\pi).\end{aligned}$$

Label $\operatorname{supp}\operatorname{marg}_{n+1} a = \{x^1, x^2\}$. Now,

$$\begin{aligned}&\left|\int_{\mathcal{X}_{n+1}} u(x)d\rho_R^{\{a,b\}}(x_{n+1}|a) - \int_{\mathcal{X}_{n+1}} u(x)d\rho_R^{\{a,b\}}(x_{n+1}|b)\right|\\ &= \left(\rho_R^{\{a,b\}}(n+1 \mid a) - \rho_R^{\{a,b\}}(n+1 \mid b)\right)|u(x_1) - u(x_2)|\\ &> \left(\rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid a^\pi) - \rho_R^{\{a^\pi,b^\pi\}}(n+1 \mid b^\pi)\right)|u(x_1) - u(x_2)|\\ &= \left|\int_{\mathcal{X}_{n+1}} u(x)d\rho_R^{\{a^\pi,b^\pi\}}(x_{n+1}|a^\pi) - \int_{\mathcal{X}_{n+1}} u(x)d\rho_R^{\{a^\pi,b^\pi\}}(x_{n+1}|b^\pi)\right|\end{aligned}$$

whenever $u(x^1) \neq u(x^2)$, and all are equal to zero when $u(x_1) = u(x_2)$. Hence, the difference has a unique maximum at π^*, establishing the result. □

The result is completed by noting that the minimal active paths are sufficient to identify all fundamental links in the subjective DAG.

A.3. **Preliminary Results for Proof of Theorem 1.** For a DAG Q, let $N^*(Q)$ be the minimal set of nodes such that $\text{marg}_{n+1} p_Q(\cdot \mid a) = \text{marg}_{n+1} p_{R\cap N^*(Q)^2}(\cdot \mid a)$ and $N^*(\rho) = \{i \in N : i \in I \text{ for some } \rho\text{-MAP } I\} \cup \{0, n+1\}$. A DAG Q' is equivalent to Q if and only if $p_Q = p_{Q'}$ for all $p \in \Delta(\mathcal{X})$. Let the skeleton of Q be $\tilde{Q} = \{(i,j) : iQj \text{ or } jQi\}$.

Proposition 2 (Theorem 1 of Verma and Pearl (1991))**.** *Two DAGs are equivalent if and only if they have the same skeleton and the set of v-colliders.*

Let the set of free-will DAGs behaviorally equivalent to Q be

$$\mathcal{B}_Q = \{Q' : p_Q(\cdot) = p_{Q'}(\cdot) \text{ for all } p \in \Delta(\mathcal{X}) \text{ and } Q'(0) = \emptyset\}.$$

Definition 9. Consider two nodes $i, j \in N^*$. If iGj for all $G \in \mathcal{B}_Q$, then the link iGj is called a *fundamental link* in Q and denoted by $i\hat{Q}j$. An Q-AP $(i_0, ..., i_m)$ is *Q-fundamental* if $i_j\hat{Q}i_{j+1}$ for all $j < m$.

Proposition 3 (Prop 6 of Schumacher and Thysen (2020))**.** *Given be a perfect DAG Q, $N^*(Q) = \{i \in N \mid i$ is part of a Q-fundamental active path between 0 and $n+1\}$.*

The fundamental links are characterized in the next proposition. Before that we need another definition.

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

Proposition 4 (Prop 7 of Schumacher and Thysen (2020))**.** *Let Q be a perfect, free will DAG, and consider a link $(i,j) \in Q$. Then, $i\hat{Q}j$ if and only if at least one of the following conditions is satisfied:*

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

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

A.4. **Proof of Theorem 1.** Let $N^*(\rho) = \{i \in N : i \in I \text{ for some } \rho\text{-MAP } I\} \cup \{0, n+1\}$.

Lemma 5. *If ρ has a perfect SCR (R, u), then $N^*(\rho) \subseteq N^*(R)$.*

Proof. It is enough to show that all the links between nodes in $I \cup \{0, n+1\}$ are fundamental for every ρ-MAP I. Suppose for contradiction that I is a ρ-MAP and $I \cup \{0, n+1\}$ contains non-fundamental links. Let iRj be the non-fundamental link in $I \cup \{0, n+1\}$ that is closest to the node 0.

If $i = 0$, then we have an immediate contradiction, since all links from 0 are fundamental by definition. Therefore suppose that $i \geq 1$ and let $k \in I \cup \{0, n+1\}$ such that kRi. By assumption kRi is fundamental. Since iRj is non-fundamental this implies (by Proposition 4.b) that kRj. However, this contradicts that I is a ρ-MAP as $I \setminus \{i\}$ contains an R-AP. □

Lemma 6. *If ρ has a perfect SCR (R, u), then $N^*(R) \subseteq N^*(\rho)$.*

Proof. Let $i \in N^*(R)$.

Step 1: [$k\hat{R}l$ and $k \neq 0$ implies $\exists j \in N$ s.t. $j\hat{R}k$ and $j \not\hat{R} l$.]

Let $k, l \in N$ so that $k\hat{R}l$ and $k \neq 0$. We want to show that there exists a node $j \in N$ so that $j\hat{R}k$ and $j \not\hat{R} l$. As $k \neq 0$ and R is a perfect free-will DAG there exists at lest one node j so that $j\hat{R}k$. Assume for contradiction that for every node j so that $j\hat{R}k$ it holds that jRl. As this rules out condition Proposition 4.b, it must be that $d(0,k) = d(0,l) - 1$. Since $d(0,k) > 0$, then by definition there exists a node j so that jRk and $d(0,j) = d(0,k) - 1$. By Proposition 4.a $j\hat{R}k$ and by assumption jRl. But then $d(0,l) \leq d(0,j) + 1 = d(0,k)$, a contradiction.

Step 2: [There exists a fundamental active path between 0 and $n+1$ that contains i so that there are no links in R between the nodes that precede i and the nodes that succeed i on the path.]

Let j and k be two nodes in the same fundamental active path, and let j be closer to 0 on the path than k. Note, that if there exists a link between j and k in R, then $j\hat{R}k$. If kRj, then there is a cycle in R as $j\hat{R}\dots\hat{R}k$ by their position on the fundamental path. Furthermore, if there is no link between j and k, then j is not linked to any nodes further down the path. Suppose that $j\hat{R}\cdots k\hat{R}k+1$, $j \not\hat{R} k$ and $jRk+1$.

Then there is a v-collider $(j, k, k+1)$, a contradiction since R is perfect. Inductively applying this argument completes the claim.

Pick some fundamental active path between 0 and $n+1$ that contains i. We fix the part of the path that succeeds i: $i\hat{R} \dots \hat{R}n+1$. We reconstruct the path preceding i as follows. Delete all the nodes that precede i. Let k and l be the first two nodes on the path. When $k \neq 0$ pick a node j so that $j\hat{R}k$ and $j \not{\hat{R}} l$ (by Step 1 such a node exists), and add the link $j\hat{R}k$ to the path. When we add a link that contains 0 we are done. The resulting fundamental active path satisfy the claim.

Let I be the nodes on this fundamental active path.

Step 3: [I contains a ρ-MAP containing i]

Let $I_-(I_+)$ be the set of nodes that precede (succeed) i on the path constructed in Step 2. By construction all the link in R between I_- and I_+ involve i. Thus, $I \setminus \{i\}$ cannot contain an R-AP. However, as I contains an active path between 0 and $n+1$, then I contains a ρ-MAP containing i, and by definition $i \in N^*(\rho)$. □

By Lemmas 5 and 6, we can restrict attention to DAGs on $N^*(\rho)$. For a DAG Q, say that $I \subseteq N$ is a Q-MAP-set if $I = \{i_2, ..., i_{m-1}\}$ and $(i_1, i_2, ..., i_m)$ is a Q-MAP-set. If ρ has an SCR (Q, u), then I is a ρ-MAP if and only if I is a Q-MAP-set. For a Q-MAP-set I, the I-ancestors and I-descendants of $i \in I \cup \{0, n+1\}$ are the sets $A_I(i) = \{j \in I : j\bar{Q}i\}$ and $D_I(i) = \{j \in I : i\bar{Q}j\}$ where $\bar{Q}$ is the transitive closure of Q.

Lemma 7. *For a perfect, free-will DAG Q and $i, j \in N^*(Q)$, $i\hat{Q}j$ if and only if there exist Q-MAP-sets I and J so that $i \in I$ and $j \in J$, $A_I(i) \cap D_J(j)$, $A_I(i) \cup D_J(j)$ contains a Q-MAP-set, and $[A_I(i) \cup D_J(j)] \setminus \{i\}$ does not contain a Q-MAP-set or $i = 0$ ($j = n+1$) and there exists a Q-MAP-set J (I) so that $D_J(j) = J$ ($A_I(i) = I$).*

Proof. Let Q be a perfect, free-will DAG and $i, j \in N^*(Q)$. Suppose first that $i\hat{Q}j$. If $i = 0$ or if there exists a Q-MAP-set I that contains both i and j then the above conditions are satisfied: $A_I(i) \cap D_I(j) = \emptyset$, $A_I(i) \cup D_I(j) = I$, and either $i \neq 0$ in which case $[A_I(i) \cup D_I(j)] \setminus \{i\}$ does not contain a Q-MAP-set or $i = 0$ and $[A_I(i) \cup D_I(j)] \setminus \{i\} = I$. Therefore, suppose that there does not exists a Q-MAP-set containing both

i and j. By Step 2 of Lemma 6, we can find a Q-MAP-sets I and J so that $j \in J$, $i \in I$, and any link from $A_I(i)$ to $D_J(j)$ involves i. Observe that $A_I(i) \cap D_J(j) = \emptyset$, since $i' \in A_I(i) \cap D_J(j)$ implies $i'Q \dots Qi\hat{Q}jQ \dots Qi'$, contradicting that Q is acyclic. Furthermore, $A_I(i) \cup D_J(j)$ contains a Q-AP and therefore a Q-MAP-set. However, $[A_I(i) \cup D_J(j)] \setminus \{i\}$ does not contain a Q-MAP-set by construction.

Now, suppose $i = 0$ ($j = n + 1$) and there exists a Q-MAP-set J (I) so that $D_J(j) = J$ ($A_I(i) = I$). This implies that $0Qj$ ($iQn + 1$), and thus $0\hat{Q}j$ ($i\hat{Q}n + 1$) by definition. Next, fix Q-MAP-sets I and J so that $i \in I$, $j \in J$, $A_I(i) \cup D_J(j)$ contains a Q-MAP-set, and $[A_I(i) \cup D_J(j)] \setminus \{i\}$ does not. Since $[A_I(i) \cup D_J(j)] \setminus \{i\}$ does not contain a Q-MAP-set, $i \notin D_J(j)$ and no $i' \in A_I(i) \setminus \{i\}$ has a link to anything in $D_J(j)$. We claim that iQj. As $A_I(i) \cup D_J(j)$ contains a Q-MAP-set by hypothesis, iQj^* for some $j^* \in D_J(j) \setminus \{j\}$. Let $j' \in D_J(j)$ be a parent of j^* and $i' \in I$ be a parent of i. As Q is perfect, $i\tilde{Q}j'$. If $j'Qi$, then since Q is perfect and $i'Qi$ we must have $j'\tilde{Q}i'$; by hypothesis, $j'Qi'$. This same logic requires that j' be a parent of i''s parent, the parent of the parent of i', and so on. Inductively, it requires $j'Q0$, a contradiction of Q having free-will. Hence, iQj', and by the same arguments iQj'' for any $j'' \in D_J(i)$ so that $j''Qj'$. Successively applying the same argument implies that iQj. If iQj, then $i\hat{Q}j$ as there exists a node in $i' \in A_I(i)$ with $i'\hat{Q}i$ and $i' \not{Q}j$. □

Lemma 8. *In a perfect, free will DAG Q with $i, j \in N^*(Q)$, if iQj and $i \not\hat{Q}j$, then there exists k such that $i\hat{Q}k$ and $j\hat{Q}k$.*

Proof. Let $i, j \in N^*(Q)$. Suppose $i\tilde{Q}j$, $i \not\hat{R}j$, and $j \not\hat{R}i$. We show first that iQk and jQk for some k. By Lemma 7, there exist Q-MAP-sets I and J such that i contains I (j contains J). Let k be the maximum length in $D_I(i) \cup D_J(i)$ from i or j to a common descendant. The payoff relevant node $n + 1$ is a descendant of both i and j so $k < \infty$. If $k = 1$, the claim is proved. Let k^* be the closest common descendant. For contradiction, assume $k > 1$. Let $A_i^0 = A_j^0 = k^*$, A_i^z be the immediate ancestor in $D_I(i)$ of A_i^{z-1} $[A_i^z Q A_i^{z-1}]$, and A_j^z be the immediate ancestor in $D_J(i)$ of A_j^{z-1} .

Since Q is perfect, $A_i^1 \tilde{Q} A_j^1$. WLOG, $A_i^1 Q A_j^1$ (otherwise repeat with roles of i and j interchanged). But then $A_i^1 Q A_j^1$ and $A_j^2 Q A_j^1$ is a v-collider. So $A_i^1 \tilde{Q} A_j^2$ and $A_i^1 Q A_j^{z'}$ for all $z' < 2$. For $z \geq 2$, assume $A_i^1 \tilde{Q} A_j^z$ and $A_i^1 Q A_j^{z'}$ for all $z' < z$. If $A_j^z Q A_i^1$, then A_i^1 is a common descendant of i and j and is closer to either than k^*. Then, $A_i^1 Q A_j^z$

and $A_j^{z+1} Q A_j^z$, so $A_i^1 \tilde{Q} A_j^{z+1}$ by perfection and moreover $A_i^1 Q A_j^{z'}$ for all $z' < z+1$. Inductively conclude that $A_i^1 Q A_j^z$ for all $z \geq 2$. By arguments as above we must have $A_i^1 Q j$.

Since jQi implies the DAG has a cycle, we must have iQj. Then, iQj and $A_i^1 Qj$, so iQA_i^1 by perfection. Since iQj is not fundamental, i's predecessor i^* must have i^*Qj [$i^*Q_I i$ implies i^*Qj]. But then i^*Qj and $A_i^1 Qj$, so $i^*QA_i^1$ by perfection. But this contradicts that I is a Q-MAP-set! Conclude $k^* = 1$.

By the above arguments, there is no loss in assuming that k^* is successor of j in the Q-MAP-set J. Since the link between i and j is not fundamental, let Q be the equivalent DAG with jQi. Proposition 4 gives that $d(j,0) \neq d(i,0) - 1$ and that $j'\hat{Q}j$ implies $j'Qi$. Let j^* be j's predecessor in J. Since $k \in D_J(j)$, $j^* \not R k$. Then, $j^* \not R k$, j^*Qi, and iQk, so $i\hat{Q}k$ by Proposition 4. (If $j^* \not\hat{R} i$, then the predecessor of j^*, j^{**} has $j^{**}Qi$. Repeat with $j^* = j^{**}$ until $j^*\hat{Q}i$ for $j^* \in J$. This terminates because J is finite, and the j^* we find must not be 0 since if $j^{**} = 0$, then J is not an Q-MAP-set.) □

Lemmas **??** and **??** establish necesssity. Let ρ have a perfect SCR (R, u). To establish sufficiency, consider any perfect, free-will DAG $R' \subseteq N^*(\rho)$ so that $(i_1, ..., i_m)$ is a R'-MAP if and only if it is also a R-MAP. By Lemma 7, $i\hat{R}j$ if and only if $i\hat{R}'j$. By Lemma 8, $(i,j) \in R' \setminus \hat{R}'$ if and only if $i\hat{R}'k$ and $j\hat{R}'k$ for some k. Since $\hat{R}' = \hat{R}$, either iRj or jRi. Hence, $R^* = R \cap N^*(\rho)^2$ and R' have the same skeleton and v-colliders. By Proposition 2, $\rho^S_{R^*} = \rho^S_{R'}$, so ρ has an SCR (R', u) if and only if it has an SCR (R^*, u). By Lemmas 5 and Lemma 6, ρ has an SCR (R^*, u). Uniqueness of u follows from the uniqueness results for Logit and EU since the DM having a logit-EU representation on R-Markov menus. □

A.5. **Preliminary Results for Proof of Theorem 2.**

Lemma 9. *If ρ has a perfect SCR, then R^ρ is a perfect, free-will DAG.*

Proof. Let ρ have a perfect SCR (R, u). Note that by construction R^ρ is a directed graph and 0 is ancestral. It has already been shown that if there is a fundamental link in R from node i to node j, then $i\hat{R}^\rho j$ (henceforth iR^*j) and so $iR^\rho j$. It remains to be R^ρ does not have any cycles or v-colliders. Denote by iQj a pair (i,j) with $i \in R^\rho(j)$

for which $i \not R^* j$ in order to distinguish the fundamental links from the non-fundamental links.

Step 1: R^ρ does not contain any cycles.
First we list some immediate facts, that will be useful later on:

(1) Any cycle contains at least one fundamental link.
This follows immediately from the construction of R^ρ from the non-fundamental links.

(2) Any cycle contains at least one non-fundamental link.
Otherwise, the set of fundamental links is not consistent with the agent's DAG.

(3) Any cycle contains at least two non-fundamental links.
If there is a cycle $iR^*jQkR^*...R^*i$, then there is also a DAG R' in the behavioral equivalence class of R where $kR'j$ since the link jRk is non-fundamental. For this DAG, $kR'i$ or $iR'k$ (otherwise, (k, i, j) is a v-collider). To avoid a cycle, only $kR'i$ is possible, so kR^*i. But then $jR'k$ creates a cycle, so it must be that kR^*j, a contradiction.

(4) Any nodes connected by a path of non-fundamental links have the same distance to node 0.
This follows immediately from Proposition 4.

(5) The shortest cycle is of length 3.
Suppose that the shortest cycle is of length 4 or more. Then there exists a fundamental link iR^*j and a non-fundamental link jQk in this cycle such that $k \not R^\rho i$ and $i \not R^\rho k$. Otherwise this is not the shortest cycle. However, this contradicts that jQk is a non-fundamental link: any perfect DAG R' behaviorally equivalent to R must have $iR'j$, $i \not R' k$ and $k \not R' i$, so $kR'j$ would constitute a v-collider.

Suppose for contradiction that R^ρ has at least one cycle. By the above if there exists a cycle in R^ρ, then there are $i, j, k \in N$ such that iR^*jQkQi. Since iR^*j, then there exists a node $l \in N$ such that lR^*i and $l \not R j$ (and $l \not R' j$ for any R' behaviorally equivalent to R). Let R', R'' be perfect DAGs on $N^*(\rho)$ behaviorally equivalent to R with $iR'k$ and $kR''i$. Since R'' has no cycles, $kR''j$. Since R'' is perfect, $kR''i$, and lR^*i, $lR''k$ or $kR''l$; hence either lQk or kQl by Proposition 2. Therefore, either $lR'k$

or $kR'l$. If $lR'k$, then (l, i, k) is a v-colllider in R' (since l and i cannot be linked), but if $kR'l$, then $iR'kR'lR'i$ is a cycle. Conclude that kR^*i, contradicting that kQi.

Step 2: R^ρ does not contain any v-colliders.
Suppose for contradiction that R^ρ has a v-collider (i, j, k). Let $D(l)$ be the R^ρ descendants of l, i.e. for any $j \in D(l)$ there is a directed path from l to j in R^ρ. By Step 1 and that there are a finite number of nodes, there is no loss in picking (i, j, k) to be such that the set $D(v) = D(i) \cup D(j) \cup D(k)$ does not contain any v-colliders in $\mathcal{R}^*$. We must have iQk and jQk; if iR^*k, then jR^*k by definition, and if iR^*k, then jR^*k.

Since $n + 1 \in D(i), D(j), D(k)$ by definition, there exists nodes l and l' such that iR^*l, kR^*l, jR^*l' and kR^*l'. Note that lRj implies lR^*j as R^ρ does not contain any cycles by Step 1. This in turn implies kR^*j, contradicting kQj. Similarly, jRl implies iRj or jRi, which contradicts that (i, j, k) is a v-collider. Similarly, neither iRl' or $l'Ri$ can hold. As a result, we get a contradiction from $lR^\rho l'$ or $l'R^\rho l$ since $D(v)$ does not contain a v-collider.

Let $l^* \in \arg\min_{l'' \in D(l) \cup D(l')} d(l, l'') + d(l', l'')$. Either $l^* = l$ or $l^* = l'$. Suppose not, then there exists $m, m' \in N$ such that $m \in D(l)$, $mR^\rho l^*$, $m' \in D(l')$, $m'R^\rho l^*$, and either $m \neq l$ or $m' \neq l'$. Since $l^*, m, m' \in D(v)$, (m, m', l^*) is not a v-collider, so $mR^\rho m'$ or $m'R^\rho m$. In the former case, $d(l, m') + d(l', m') < d(l, l^*) + d(l', l^*)$, and in the latter case $d(l, m') + d(l', m') < d(l, l^*) + d(l', l^*)$. This contradicts the definition of l^*, so $l^* = l$ or $l^* = l'$.

Suppose $l^* = l'$, so there is a directed path from l to l'. Let m be the parent of l' on that path. Since jR^*l', (j, m, l') cannot be a v-collider, so either $jR^\rho m$ or $mR^\rho j$. The latter would cause a cycle, so $jR^\rho m$. Similarly, $kR^\rho m$. Let m^* be the parent of m on the path. The same argument implies that $jR^\rho m^*$ and $kR^\rho m^*$. By continuing this way, we see $jR^\rho l$, a contradiction. A similar contradiction obtains when $l^* = l$. Conclude R^ρ does not contain a v-collider, completing the proof. □

Lemma 10. *If R^ρ is a perfect, free-will DAG and ρ satisfies Axioms 1, 4, 5, and 6 , then there exists non-constant $u : \mathcal{X}_{n+1} \to \mathbb{R}$ so that*

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

for every R^ρ-Markov $S \in \mathcal{S}$, and u is unique up to adding a constant.

Proof of Lemma 10. Say that ρ has a Luce representation with index u on a subset of menus $\Sigma \subset \mathcal{S}$ if for every $S \in \Sigma$, $\rho(a, S) = u(a)/\sum_{b\in S} u(b)$ for every $a \in S$.

For any finite $Y \subset \mathcal{X}_{n+1}$, let $P(Y, \epsilon) = \{p \in \Delta\mathcal{X}_{n+1} : p(Y) = 1, p(y) \geq \epsilon \forall y \in Y\}$ for $\epsilon \in (0, \frac{1}{M})$ with $M = |Y|$. For any $p_1, \ldots, p_m \in P(Y, \epsilon)$, there is an R^ρ-Markov menu $S = \{a_1, \ldots, a_m\}$ so that $\operatorname{marg}_{n+1} a_i = p_i$. To do so, let $(i_0 = 0, i_1, \ldots, i_k = n + 1)$ be a minimal active path in R^ρ, i.e. $i_j R^\rho i_{j+1}$ for $j = 0, 1, ..., k - 1$, and label $Y = \{y_1, \ldots, y_M\}$. Take $A(Y, \eta) \subset \Delta X$ to be the lotteries so that X_{i_j} takes values between 1 and M for $j = 1, \ldots, k-1$, $x_{i_{j+1}} = x_{i_j}$ with probability $(1-\eta)$ and equals every other value with equal probability, $X_{n+1} = y_i$ with probability 1 whenever $X_{i_{k-1}} = i$, and $X_j = 0$ with probability 1 for every $j \notin \{i_0, \ldots, i_k\}$. Observe that $A(Y, \eta)$ is convex, and that there is an $a \in A(Y, \eta)$ so that $a(X_{n+1} = y_i) \geq a(X_1 = i)(1 - \eta)^k$, which approaches 1 as $\eta \to 0$ and $a(X_1 = i) \to 1$, and if $a(X_1 = j) = a(X_1 = j')$ for all $j, j' \neq i$, then $a(X_{n+1} = y_j) = a(X_{n+1} = y_j') < a(X_1 = i)(1 - \eta)^k$. In particular, given $\epsilon > 0$, there exists $\eta > 0$ so that for any $p_i \in P(Y, \epsilon)$ there is an $a_i \in A(Y, \eta)$ so that $\operatorname{marg}_{n+1} a_i = p_i$ by convexity. In particular $\{\operatorname{marg}_{n+1} a : a \in A(Y, \eta)\} = P(Y, \epsilon)$ for appropriately chosen η.

Let $P(Y) = \cup_{\epsilon>0} P(Y, \epsilon)$. We claim that there exists is a Luce representation u_Y on all R^ρ-Markov S for which $\operatorname{marg}_{n+1} a \in P(Y)$ for every $a \in S$, and $u_Y(a) = u_Y(b)$ whenever $\operatorname{marg}_{n+1} a = \operatorname{marg}_{n+1} b$. By the above, there exists an $\eta > 0$ so that for every $a \in S$ there exists $a' \in A(Y, \eta)$ with $\operatorname{marg}_{n+1} a' = \operatorname{marg}_{n+1} a$. Let S' be these actions. By Axiom 4 and standard results, there is a Luce representation when restricted to $A(Y, \eta)$; let u_η be its index. By Axiom 6, $u_\eta(a)/u_\eta(b) = u_{\eta'}(a')/u_{\eta'}(b')$ whenever $\operatorname{marg}_{n+1} a = \operatorname{marg}_{n+1} a'$ and $\operatorname{marg}_{n+1} b = \operatorname{marg}_{n+1} b'$. Pick η^* and $a \in A(Y, \eta^*)$. Normalize $u_{\eta'}$ for each $\eta' < \eta^*$ so that $u_{\eta'}(a) = u_{\eta^*}(a)$. Since there is one degree of freedom, $u_{\eta'} = u_{\eta''}$ on their common domain. Let this be u_Y, and then Axiom 6 establishes the claim.

Next, we extend to any other $P(Y')$ where $Y' \supset Y$. Pick Y and $p^* \in P(Y)$. Define $U = u_Y$ on $P(Y)$. Now, take $Y' \supset Y$ and $r \in P(Y')$. Set

$$\lambda = \exp\left[2 \ln u_{Y'}\left(\frac{1}{2}p^* + \frac{1}{2}r\right) - \ln u_{Y'}(r)\right] / U(p^*)$$

and then take $v(p) = \lambda u_{Y'}(p)$ for any $p \in P(Y')$. There is a unique continuous extension of v to $cl(P(Y'))$, which includes $P(Y)$, and this extension of v coincides with U. Hence v extends U to $cl(P(Y'))$. Then, the set on which we define the Luce index is irrelevant for its value and hence we can extend U to all of $\Delta(\mathcal{X}_{n+1})$.

To see that this is consistent, consider the sequence $p_m = \frac{m-1}{m}p^* + \frac{1}{m}r$, noting $p_{m+1} = \frac{1}{m+1}p^* + \frac{m}{m+1}p_m$. For each m, there is $\eta > 0$ and $r, p^*_m, p^*_{m+1} \in A(Y', \eta)$ so that $\operatorname{marg}_{n+1} p^*_m = p_m$, $\operatorname{marg}_{n+1} r^* = r$, and $\operatorname{marg}_{n+1} p^*_{m+1} = p_{m+1}$. Then by Axiom 5,

$$\begin{aligned} \frac{m-1}{m} \ln \frac{\rho(\{p^*_{m+1}, r^*\})(p^*_{m+1})}{\rho(\{p^*_{m+1}, r^*\})(r^*)} &= \frac{m}{m+1} \ln \frac{\rho(\{p^*_m, r^*\})(p_m)}{\rho(\{p^*_{m+1}, r^*\})(r^*)} \\ \iff \frac{m-1}{m}[\ln v(p_{m+1}) - \ln v(r)] &= \frac{m}{m+1}[\ln v(p_m) - \ln v(r)] \end{aligned}$$

since the menus are R^ρ-Markov. Recursively substituting yields that

$$\ln v(p_{m+1}) = \prod_{j=2}^{m} \frac{j^2}{j^2-1}[\ln v(\frac{1}{2}p + \frac{1}{2}r) - \ln v(r)] + \ln v(r) \to 2\ln v(\frac{1}{2}p + \frac{1}{2}r) - \ln v(r)$$

so $\lim v(p_{m+1}) = u(p^*)$. Pick any q, and set $q_m = \frac{m-1}{m}q + \frac{1}{m}r$ Since $\operatorname{marg}_{n+1} p_{m+1} \to \operatorname{marg}_{n+1} p^*$ and $\operatorname{marg}_{n+1} q_{m+1} \to \operatorname{marg}_{n+1} q$, $\frac{\rho(\{p_m,q_m\})(p_m)}{\rho(\{p_m,q_m\})(q_m)} \to \frac{\rho(\{p^*,q\})(p^*)}{\rho(\{p^*,q\})(q)}$ by Axiom 6. Then, $v(p_m)/v(q_m) \to U(p)/U(q)$, implying that $v(q_m) \to U(q)$. Moreover, for any other q'_m so that $\operatorname{marg}_{n+1} q'_m \to q$, we must have $v(q'_m) \to U(q)$ by Axiom 6 and that $v(q_m)$ converges.

Now, let U be the overall Luce index and $V = \ln U$. V is affine (by Axiom 5), continuous (by Axiom 6), and ranks every lottery in $\Delta(\mathcal{X}_{n+1})$. Conclude there exists $u : \mathcal{X}_{n+1} \to \mathbb{R}$ so that $V(p) = \int u(c)dp(c)$ for any $p \in \Delta\mathcal{X}_{n+1}$. Then, $U(p) = \exp V(p) = \exp \int u(c)dp(c)$, completing the proof. □

A.6. **Proof of Thoerem 2. Necessity:** Suppose that ρ has a perfect SCR. Lemma 9 implies that R^ρ is a perfect, free-will DAG, and Theorem 1 implies that ρ has an SCR (R^ρ, u). Axioms 2 and 6 are obviously necessary. Axiom 4 follows from continuity of the expected utility functional and $\rho^{S_m}|R \to \rho^S|R$ implies $\rho^{S_m}_{R^\rho}(\cdot|a) \to \rho^S_{R^\rho}(\cdot|a)$. Axiom 5 follows since for an R^ρ-Markov menu S,

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

so if $a = \alpha a' + (1-\alpha)b$,

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

Axiom 3 follows from

$$\int u(c) dp_R(c_{n+1}|a') \in \left[\min_{x \in \mathcal{X}_{-0}} \int u(c) dp(c_{n+1}|x_N), \max_{x \in \mathcal{X}_{-0}} \int u(c) dp(c_{n+1}|x_N)\right]$$

by iterated expectations.

Sufficiency: Suppose that ρ satisfies the axioms and that R^ρ is a perfect free-will DAG. Given Lemma 10, we approximate choice in every menu by a sequence of R^ρ-Markov menus. Pick any $S = \{a_1, a_2, \ldots, a_m\} \in \mathcal{S}$. We show that

$$\frac{\rho(a_i,S)}{\rho(a_j,S)} = \frac{\exp[\int_{\mathcal{X}_{n+1}} u(c) d\rho^S_{R^\rho}(c_{n+1}|a_i)]}{\exp[\int_{\mathcal{X}_{n+1}} u(c) d\rho^S_{R^\rho}(c_{n+1}|a_j)]} \tag{2}$$

for any i, j. The DM then has a SCR (R^ρ, u), since the probabilities add up to one.

Define $a'_k(y) = \rho^S_{R^\rho}(y|a_k)$ for every k and every $y \in \mathcal{X}_{-0}$. Pick any i, j, and let $a = a_i$, $b = a_j$, $a' = a'_i$, and $b' = a'_i$. Since $\{a', b'\}$ is R^ρ-Markov, $\rho(a', \{a', b'\})/\rho(b', \{a', b'\})$ has the desired form by Lemma 10. If $a' = b'$, then $a(x_F) = b(x_F)$ where $F = \{k \in N^* : 0 \in R^\rho(k)\}$, so $\rho(a,S) = \rho(b,S)$ by Axiom 2, and the formula 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 R^ρ-Markov by construction, and each has $m+1$ distinct alternatives. By Axiom 3, there exists $K > 0$ so that for any $a'', b'' \in S'' \in \mathcal{S}$, $\frac{\rho(a'',S'')}{\rho(b'',S'')} \leq K$. In particular, for $S_m \setminus S = \{s_1, \ldots, s_{M(m)}\}$ (noting $M(m) \geq m + 1 - |S|$), $a'' \in S$, and $i \leq M(m)$, we have $K^{-1}\rho(a'', S_m \cup S) < \rho(s_i, S_m \cup s)$. Then,

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

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

For $p_m = \rho^{S_m \cup S}$ and arbitrary $i \in N^*$ with $0 \not{R}^\rho i$ and $E = R^\rho(i)$, we have

$$p_m(x_i|x_E) = \frac{1}{p_m(x_E)} \left[\sum_{a'' \in S} p_m(a'') p_m(x_E|a'') a''(x_i|x_E) + p_m(S_m) p_m(x_E|x_0 \in S_m) a'(x_i|x_E)\right]$$

for p-a.e. $x \in \mathcal{X}_{-0}$ since $\hat{a}(x_i|x_E) = a'(x_i|x_E)$ for all $\hat{a} \in S_m$. This converges to $\rho^{S_1}(x_i|x_E) = a'(x_i|x_E)$ because $p_m(a'') \to 0$ for all $a'' \in S$. Since i and $x_{R^\rho(i)}$ were arbitrary, $\rho^{S_m \cup S}|R^\rho \to \rho^{S_1}|R^\rho = \rho^S|R^\rho$.

Axiom 2 gives that $\rho(S_m \cup S, a) = \rho(S_m \cup S, a')$ and $\rho(S_m \cup S, b) = \rho(S_m \cup S, b')$. Axiom 4 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 (2) holds for i, j. Since i, j and S were arbitrary, ρ has an SCR. □

Appendix B. Proofs from Section 5

B.1. **Deterministic Identification.** First identify u. Given a c-MAP I and $p, q \in \Delta(\mathcal{X}_{n+1})$ let $\{a^\pi, b^\pi\}$ be a Markovian family of menus for I where $\operatorname{supp} \operatorname{marg}_j a^\pi = \{h, l\}$ for $h > l$, $c^\pi(x_{n+1}|h_{\pi(|I|)}) = p(x)$ and $c^\pi(x_{n+1}|l_{\pi(|I|)}) = q(x)$ for all $c \in \{a, b\}$, $a^\pi(h_{\pi(1)}) > b^\pi(h_{\pi(1)})$. Observe that

$$p^X_R(x_{n+1}|a) - p^X_R(x_{n+1}|b) = [p(x) - q(x)][p^X_R(h_j|h_i) - p^X_R(h_j|l_i)][a^\pi(h_i) - b^\pi(h_i)] + q(x)$$

for any $p \in \Delta(\{a^\pi, b^\pi\})$ when $i \in I$ is s.t. $0Ri$, and $j \in I$ is s.t. $jR(n+1)$. Since $p^X_R(h_j|h_i) > 0$, $c(\{a^\pi, b^\pi\}) = \{\delta_{a^\pi}\}$ whenever $\int u dp > \int u dq$, $c(\{a^\pi, b^\pi\}) = \{\delta_{b^\pi}\}$ whenever $\int u dp < \int u dq$, and $c(\{a^\pi, b^\pi\}) = \Delta(\{a^\pi, b^\pi\})$ whenever $\int u dp = \int u dq$.

Let u_c by the utility index that represents c above. For any u and DAG R, define $\rho_{(R,u)}$ to be a choice rule with a perfect SCR (R, u). We generalize Markovian family of menus as follows.

Definition 11. Given a DAG R, $u_0 \in (0, 1)$ and $\epsilon > 0$, say that $\{\{\hat{a}^\pi, \hat{b}^\pi, \hat{e}^\pi\}\}_\pi$ is a (R, u_0, ϵ)-DMarkovian family of menus for $J \subset N$ if there exists a Markovian family of menus for J $\{a^\pi, b^\pi\}$ where $\operatorname{supp} \operatorname{marg}_j a^\pi = \{h, l\}$ for $h > l > 0$, $u_c(h) > u_c(l)$, and all $j \in J \cup n+1$, $\operatorname{marg}_{j'} a^\pi = \delta_0$ for $j' \in N \setminus J$, $a^\pi(h_{n+1}|h_{\pi(|I|)}) = a^\pi(l_{n+1}|l_{\pi(|I|)}) = 1$, and

$a^\pi(h_{\pi(1)}) > b^\pi(h_{\pi(1)})$, and for each π,

$$\begin{aligned}
\hat{a}^\pi &= (1-\epsilon^2)a^\pi + \epsilon^2\iota \\
\hat{b}^\pi &= (1-\epsilon^2)b^\pi + \epsilon^2\iota \\
\hat{e}^\pi &= (1-\epsilon-\epsilon^2)d_z + \epsilon\rho_{(R,u_c)}^{\{a^\pi,b^\pi\}} + \epsilon^2\iota \\
d_z &= u_0\delta_{\{(0,\dots,0,h)\}} + (1-u_0)\delta_{\{(0,\dots,0,l)\}} \\
\iota &= (x_I, 0_{N\setminus I}, x_{n+1}) \mapsto \frac{1}{2}3^{-\#I}\chi_{\{h,l,0\}^I\times 0^{N\setminus I}\times\{h,l\}}(x_I, 0_{N\setminus I}, x_{n+1}).
\end{aligned}$$

Setting $\epsilon = 0$, $\{\{a^\pi, b^\pi\}\}_\pi$ is a Markovian family of menus (without full-support). The menu $\{a^\pi, b^\pi, e^\pi\}$ also has a Markovian structure (but not binary), and has the same conditional distributions for h, l as $\{a^\pi, b^\pi\}$. It adds an extra value, 0, that is an absorbing state and leads to payoff of h with probability u_0. For $\epsilon > 0$, the above is true only approximately, and so we need to take the limit. Moreover, for any $q, q' \in co\{\hat{a}^\pi, \hat{b}^\pi, \hat{e}^\pi\}$, $q(x_{\pi(i+1)}|x_{\pi(i)}) = q'(x_{\pi(i+1)}|x_{\pi(i)})$ and $X_{\pi(i+1)}$ is conditionally independent of $X_0, \dots, X_{\pi(i-1)}$.

Proposition 5. *Suppose that c has a DSCR (R, u) and that $\{\{\hat{a}^\pi, \hat{b}^\pi, \hat{e}^\pi\}\}_\pi$ is a (R, u_0, ϵ)-DMarkovian family of menus for a c-MAP I.*
If $\epsilon > 0$ is small enough and $u_0 < \hat{a}^\pi(h_{n+1})$ is large enough, then π agrees with R if and only if $(1, \hat{e}^\pi) \notin c(\{\hat{a}^\pi, \hat{b}^\pi, \hat{e}^\pi\})$.

For each ϵ define $p^\epsilon \in \Delta\mathcal{X}$ to be so that

$$p^\epsilon(\hat{c}^\pi, y) = \frac{\epsilon\rho_{(R,u_c)}(\{a^\pi, b^\pi\})(c^\pi)c^\pi(y) + \epsilon^2\frac{1}{2}^{|I|+1}}{\epsilon+\epsilon^2}$$

when $y_i \in \{h, l\}$ for $i \in I$ and $y_i = 0$ for all other i. Then, $p^\epsilon(\hat{c}^\pi, y) \to \rho_{(R,u_c)}^{\{a^\pi,b^\pi\}}(c^\pi, y)$, and when π does not agree with R and I contains a ρ-MAP,

$$p_R^\epsilon(h_{n+1}|\hat{a}^\pi) \to \left(\rho_{(R,u_c)}^{\{a^\pi,b^\pi\}}\right)_R(h_{n+1}|a^\pi) < a^\pi(h_{n+1}) \leftarrow p^\epsilon(h_{n+1}|\hat{a}^\pi)$$

by Proposition **??**.

Proof. Take any index π. Let R_π be a perfect free-will DAG that agrees with π on I. Pick ϵ and u_0 so that

$$p_{R_\pi}^\epsilon(h_{n+1}|\hat{a}^\pi) > \hat{e}^\pi(h_{n+1}) > \max_{\{Q:Q\cap I^2\neq R_\pi\cap I^2\}} p_Q^\epsilon(h_{n+1}|\hat{a}^\pi).$$

As $\epsilon \to 0$, $p^{\epsilon}_{R_\pi}(h_{n+1}|\hat{a}^\pi) \to a^\pi(h_{n+1})$ and the lower bound is below $a^\pi(h_{n+1})$ by Proposition **??**, so a u_0 exists for every sufficiently small $\epsilon > 0$.

Suppose that π that does not agree with R. Consider

$$p^t = \Big((1-t^{-1}), \hat{e}^\pi; t^{-1}/2, \hat{a}^\pi; t^{-1}/2, \hat{b}^\pi\Big)$$

noting that

$$p^t(x_j|x_i) \to^{t\to\infty} p^\epsilon(x_j|x_i) = \frac{\epsilon\rho^{\{a^\pi,b^\pi\}}_{(R,u_c)}(x_ix_j) + \epsilon^2\frac{1}{9}}{\epsilon\rho^{\{a^\pi,b^\pi\}}_{(R,u_c)}(x_i) + \epsilon^2\frac{1}{3}} \to^{\epsilon\to 0} \rho^{\{a^\pi,b^\pi\}}_{(R,u_c)}(x_j|x_i)$$

for $x \in \{h,l\}^{n+1}$ and $i,j \in I$. Since $p^t(\hat{e}^\pi) \to 1$, $p^t_R(h_{n+1}|\hat{e}^\pi) \to \hat{e}^\pi(h_{n+1})$ by Spiegler (2017). and hence $p^t_R(h_{n+1}|\hat{e}^\pi) > p^t_R(h_{n+1}|\hat{a}^\pi) > p^t_R(h_{n+1}|\hat{b}^\pi)$ for all t sufficiently large. Hence (p^t) is a sequence of $1/t$-personal equilibria that converges to $(1, \hat{e}^\pi)$.

Now suppose that π that agrees with R. For contradiction, p^t be a $(R, u, 1/t)$-personal equilibrium that converges to $(1, \hat{e}^\pi)$. As above, $p^t(x_j|x_i) \to \rho^{\{a^\pi,b^\pi\}}_{(R,u_c)}(x_j|x_i)$. But since π agrees with R, this implies that $p^t_R(h_{n+1}|\hat{a}^\pi) \to a^\pi(h_{n+1}) > p^t_R(h_{n+1}|\hat{e}^\pi)$, requiring that $p^t(\hat{e}^\pi) \leq 1/t$, a contradiction. □

B.2. **Proof of Proposition 1.** Let ρ_1 and ρ_2 have perfect SCRs and ρ_2 have a coarser model than ρ_1. This implies that $\rho_1(\cdot, S) = \rho_2(\cdot, S)$ whenever $X_i \perp_S X_{-i}$ for all $i \in N \setminus K$ and K is ρ_2-MAP. We show first that if K is a ρ_2-MAP, then K is also a ρ_1-MAP. Pick $\{a,b\}$ where $X_i \perp_S X_{-i}$ for all $i \in N \setminus K$ and $\rho_2(a, \{a,b\}) \neq \frac{1}{2}$. By hypothesis, $\rho_1(a, \{a,b\}) \neq \frac{1}{2}$ and so $N \setminus K$ does not separate for ρ_1. Hence there is $K' \subseteq K$ that is a ρ_1-MAP. Since $N \setminus K'$ does not separate for ρ_1, there exists $\{a,b\}$ where $X_i \perp_S X_{-i}$ for all $i \in N \setminus K'$ and $\rho_1(a, \{a,b\}) \neq \frac{1}{2}$. But since $N \setminus K' \supseteq N \setminus K$ hypothesis, $\rho_2(a, \{a,b\}) = \rho_1(a, \{a,b\}) \neq \frac{1}{2}$ and so $N \setminus K'$ does not separate for ρ_2 either. Since $K' \subseteq K$ and K is the smallest set that does not separate, $K = K'$.

We now show that $\hat{R}^{\rho_2} \subseteq \hat{R}^{\rho_1}$. The above gives that $N^*(\rho_2) \subseteq N^*(\rho_1)$. Let $\{\{a^\pi, b^\pi\}\}_\pi$ be a Markovian family of menus for a ρ^2-MAP K. By construction, $X_i \perp_{\{a^\pi,b^\pi\}} X_{-i}$ for all $i \in N \setminus K$ and all indexes π. Hence the ρ_2-subjective ordering is the same as the ρ_1-subjective ordering, clearly implying that $\hat{R}^{\rho_2} \subseteq \hat{R}^{\rho_1}$. This completes the proof after taking $R_i = R^{\rho_i}$.

Conversely, let ρ_1 and ρ_2 have perfect SCRs (R_1, u_1) and (R_2, u_2) where $u_2 = u_1+\beta$ and $R_2 = R_1 \cap [N' \times N']$ for some $N' \subset \{0, ..., n+1\}$. Pick any ρ_2-MAP $K = \{i_1, ..., i_m\}$ where $(i_0 = 0, i_1, \ldots, i_m, i_{m+1} = n + 1)$ is an R_2-MAP. Then, it is also an R_1-AP since $R_2 \subset R_1$. If $i_j R_1 i_k$ and $k \neq j + 1$, then $k \notin N'$ and $(i_0, ..., i_{m+1})$ cannot be an R_2-MAP, so $(i_0, ..., i_{m+1})$ is also a R_1-MAP and K is a ρ_1-MAP. Now, pick any S so that $X_i \perp_S X_{-i}$ for all $i \in N \setminus K$ and any $p \in \Delta S$ with full support. Viewing p as a member of $\Delta\mathcal{X}$, observe that for $i \in K \cup \{n + 1\}$ and p-a.e. $x \in \mathcal{X}$,

$$\begin{aligned} p\left(x_i | x_{R_1(i)}\right) =& p\left(x_i | x_{R_2(i)}, x_{R_1(i)\setminus N'}\right) \\ =& p\left(x_i | x_{R_2(i)\cap K}, x_{(R_1(i)\setminus N')\cap K}\right) = p\left(x_i | x_{R_2(i)}\right) \end{aligned}$$

where the second equality follows from independence, and the third from $K \subset N'$. Hence $p_{R_1} = p_{R_2}$, and the set of R_1-personal equilibriums for S equals the set of R_2-personal equilibriums for S. □

References

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

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

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

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

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

Cerreia-Vioglio, S., Dillenberger, D., Ortoleva, P., and Riella, G. (2019). Deliberately Stochastic. *American Economic Review*, 109(7):2425–2445.

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

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

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

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

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

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

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

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

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

Griffith, G. J., Morris, T. T., Tudball, M. J., et al. (2020). Collider bias undermines our understanding of covid-19 disease risk and severity. *Nat Commun*, 11:5749.

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

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

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

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

Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic theory*, 123(2):81–104.

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

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

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

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

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

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

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

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

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

Miyara, M., Tubach, F., Pourcher, V., Morelot-Panzini, C., et al. (2020). Low incidence of daily active tobacco smoking in patients with symptomatic covid-19. *Qeios.*

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

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

Ritov, I. and Baron, J. (1992). Status-quo and omission biases. *J Risk Uncertainty*, 5:49–61.

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

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

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

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

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

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

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

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

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