Analysis Plan for "Models of Causal Belief Systems and Misperceptions: Experimental Tests"

Theoretical Framework

The theoretical framework is Spiegler's (2016) model of boundedly rational expectations with Bayesian networks, extended to allow noisy choice. In Spiegler's framework, agents fit a subjective causal model to the data they observe. If the subjective causal graph is misspecified, the fitted model can transform correlations in the data into incorrect causal beliefs. We extend this framework by replacing deterministic best response with logit quantal response in the spirit of McKelvey and Palfrey (1995). In this extension, subjective payoff differences determine choice probabilities rather than deterministic choices. Adapting McKelvey and Palfrey's terminology, we refer to fixed points of this noisy-choice system as QRE, short for quantal-response equilibrium.

Definitions. A *DAG* is a directed acyclic graph. A *DGP* is the data-generating process. The three main variables are binary variables A, X, and Y. Subjects can choose or intervene on A, while X and Y are generated by the experimental environment. The outcome Y is payoff-relevant. A *collider* $A \to X \leftarrow Y$ means that both A and Y affect X. A *chain* $A \to X \to Y$ means that A affects X, which affects Y. We call DGPs that do not include a causal path from A to Y *non-causal*. Each treatment seeks to induce a particular misspecified subjective model through which the agent interprets the data and that involves a causal path from A to Y. A *matched causal treatment* is a treatment in which the causal structure of the true DGP is changed to that subjective model, and which is parameterized to match the joint distribution or fitted conditional probabilities generated by the corresponding noncausal treatment under that subjective model. In the reverse-causality treatments, ε denotes DGP noise in the generation of X. In a noisy-OR treatment, $P(X = 1 \mid A, Y) = 1 - \varepsilon$ if $A = 1$ or $Y = 1$, and $P(X = 1 \mid A, Y) = \varepsilon$ otherwise. In a noisy-AND treatment, $P(X = 1 \mid A, Y) = 1 - \varepsilon$ if $A = 1$ and $Y = 1$, and $P(X = 1 \mid A, Y) = \varepsilon$ otherwise (these definitions slightly differ from the usual uses of the terms noisy-OR and noisy-AND). *Logit precision* refers to the sensitivity of choice probabilities to subjective payoff differences; higher precision means choices are closer to deterministic best responses.

Experimental Environments

The design has two main three-variable problem classes and a two-variable calibration task.

1. **Reverse Causality Problem.** The true DGP has collider structure $A \to X \leftarrow Y$, so A has no causal effect on Y. It seeks to induce the subjective model $A \to X \to Y$ by displaying the realizations of variables in the corresponding temporal order. Treatments vary DGP noise, the functional form of $P(X|A, Y)$ (our versions of noisy-OR and noisy-AND), and the order in which variables are displayed to subjects.

2. **Confounded Choice Problem.** Subjects observe X, choose A, and then observe Y. The true DGP has structure $Y \to X \to A$, so A has no causal effect on Y. It seeks to induce the subjective model $X \to A \to Y$ by displaying the realizations of variables in the corresponding temporal order. Treatments vary the informativeness of X about Y.
3. **Two-variable calibration.** Subjects face environments involving only A and Y with different magnitudes of the actual causal effect of A on Y. These tasks benchmark how beliefs about causal-effect magnitudes map to observed effect sizes when the causal structure is known and simple.

Primary Outcomes

The primary behavioral outcomes are causal belief systems and action choices . The former consist of vectors of elicited beliefs about the causal effect of one variable on another potentially fixing (interventionally) the third variable. They also include the graphs subjects draw to represent their beliefs about the causal influences of variables on each other. The latter consist of the choices subjects made in the last 30 of the total of 80 rounds in each condition ($P(A = 1)$ in reverse-causality and two-variable tasks, and $P(A = 1 \mid X = 0)$ and $P(A = 1 \mid X = 1)$ in confounded-choice tasks), as well as in the WTP/WTA they reveal in the corresponding high-stakes decision. The focus on the last 30 rounds is due to the fact that learning may take time and the fact that the environment presents subjects with an explore/exploit dilemma. The focus on the last 30 rounds makes exploration motives less important; the WTP/WTA elicitations are constructed such that exploration motives cannot play a role.

Hypotheses

We test two classes of hypotheses. The first class concerns rationalizability of subjects' elicited causal belief systems: can the beliefs be represented as if subjects fit one DAG, or a mixture of DAGs, to the data they observed? The second class concerns the behavioral comparative statics predicted by Spiegler's (2016) model, extended to allow noisy choice.

Part I: Rationalizability of Causal Belief Systems

To define rationalizability, for subject i, let g_i be the vector of elicited beliefs about causal effects. Let P_i be the empirical joint distribution over (A, X, Y) observed by that subject in the relevant task. Let $\mathcal{D}$ denote the set of admissible subjective DAGs. In the Reverse Causality case, a subjective DAG is admissible if A is ancestral. In the Confounded Choice Problem, a subjective DAG is admissible if X is not a descentant of A and A is not a descendant of Y.

For each admissible subjective DAG D, compute the projection of P_i onto D:

$$P_i^D(A, X, Y) = \prod_{V \in \{A,X,Y\}} P_i(V \mid \mathrm{Pa}_D(V)),$$

where $\mathrm{Pa}_D(V)$ is the set of parents of V in DAG D. Applying do-calculus to P_i^D gives a vector s_i^D of DAG-implied belief moments corresponding to the entries of g_i. Let S_i be the matrix whose columns are these s_i^D vectors.

Single-DAG rationalizability asks whether the subjects' beliefs can be explained as if by fitting a single DAG to the data. Formally, it asks whether g_i is close to one column of S_i. DAG-mixture rationalizability asks whether the subjects' beliefs can be explained as if by fitting a mixture of

DAGs to the data (as, for instance, a Bayesian structure learner would do). Formally, it asks whether g_i is close to the convex hull of the columns of S_i, i.e. whether there exists a vector q of nonnegative DAG weights summing to one such that $g_i \approx S_i q$.

Single-DAG rationalizability.
In the reverse-causality OR environment, if a subject reports a positive total effect of A on Y, the fitted chain $A \to X \to Y$ implies that the effect is blocked when X is fixed. In the confounded-choice environment, if a subject reports a nonzero effect of A on Y, the fitted graph should imply that A screens off residual dependence between X and Y.
Mixture-DAG rationalizability.
We test mixture-DAG rationalizability by measuring the distance from elicited belief systems to the set of belief systems that are rationalizable by mixtures of DAGs and by comparing these to noise and random-choice benchmarks. Noise benchmarks use repeated or otherwise redundant elicitations to measure the distance generated by elicitation noise. Random-choice benchmarks compare observed distances to distances generated by belief reports that are randomly reassigned or otherwise made uninformative about the subject's own data. In the representative-agent specification, the distance to the rationalizable set is:

$$\min_{q \in \Delta(\mathcal{D})} \sum_i \|g_i - S_i q\|^2.$$

The minimized value, rather than the minimizing q, is the object of interest. The mixture weights q are generally not identified: distinct mixtures can imply the same belief vector or the same distance to the rationalizable set. This non-identification does not create a problem for the rationalizability question, because rationalizability is a question about whether the elicited beliefs lie in or near the convex set $\{S_i q : q \in \Delta(\mathcal{D})\}$, not about which mixture generated them.
To address issues of overfitting, the main DAG-mixture exercise uses subject-level split-sample validation with shrinkage toward a uniform mixture. Let q^0 denote the uniform distribution over admissible DAGs. For each random split r, subjects are divided into training subjects T_r and validation subjects V_r. All observations for a subject stay on the same side of the split. For each shrinkage value $\tau \in [0, 1]$, estimate

$$\widehat{q}_r(\tau) = \arg\min_{q \in \Delta(\mathcal{D})} \left\{ \frac{1-\tau}{|T_r|} \sum_{i \in T_r} \|g_i - S_i q\|^2 + \tau \|q - q^0\|^2 \right\}.$$

Then evaluate the held-out validation loss

$$L_r^{\text{val}}(\tau) = \frac{1}{|V_r|} \sum_{i \in V_r} \|g_i - S_i \widehat{q}_r(\tau)\|^2.$$

We will consider the shrinkage parameter that minimizes the held-out validation loss, and focus on distance estimates given that parameter.

Part II: Comparative statics predicted by Spiegler's (2016) model

This part tests whether choices and beliefs move in the directions predicted by Spiegler's model once the agent's subjective model is fitted to the data generated by her own behavior. The noisy-choice extension modifies the deterministic personal-equilibrium predictions by replacing best response with logit response. In the reverse-causality problem, the payoff assumption is that taking A is costly and that Y is payoff-relevant, so action is attractive only to the extent that the subject

perceives a sufficiently large beneficial effect of A on Y. In the confounded-choice problem, Y remains payoff-relevant, and the direct payoff from taking A depends on the observed signal X: the design has a high-cost state at $X = 0$ and a rewarded or low-cost state at $X = 1$.

We will perform these analyses on the whole sample and on two theory-relevant subsamples. First, we will repeat the analyses on the subset of subjects who draw a DAG that includes a causal path from A to Y. Second, we will repeat the analyses on the subset of subjects who draw the specific subjective models emphasized by the theory: in Reverse Causality, the chain $A \to X \to Y$; in Confounded Choice, either the chain $X \to A \to Y$ or the DAG with $A \to Y$ and X isolated.

Noisy-OR reverse causality.
In the OR reverse-causality problem, the QRE is unique and stable. Increasing DGP noise lowers $P(A = 1)$. This is assessed using low- versus high-noise OR reverse-causality comparisons.

Noisy-AND reverse causality.
In the AND reverse-causality problem, sufficiently low choice noise allows two stable QRE, one near low action and one near high action; sufficiently high choice noise gives a unique stable interior QRE. At any regular stable interior QRE, meaning an equilibrium away from the boundary whose local adjustment dynamics are stable, increasing DGP noise lowers $P(A = 1)$. This is assessed in the AND-connector low- and high-noise treatments. Multiplicity is assessed from the distribution of subject-level action frequencies; the noise comparative static is assessed by low- versus high-noise comparisons.

Confounded-choice QRE.
For nonzero and finite logit precision, confounded-choice QRE satisfy $1 > P(A = 1 \mid X = 1) > P(A = 1 \mid X = 0) > 0$. In the rewarded-signal case, meaning the case in which taking A is directly rewarded or less costly at $X = 1$, $P(A = 1 \mid X = 1) > 1/2$. Increasing the informativeness of X about Y raises both conditional action probabilities on stable regular branches. This is assessed using confounded-choice low- and high-informativeness comparisons.

Matched causal comparisons.
If behavior in noncausal treatments is driven by fitting a subjective chain to observed correlations, our matched causal treatments in which the DGP is changed to conform to the subjects' intended subjective DAGs should yield the same behavior.

Structural Estimation and Prediction Exercise

To test the predictive validity of the noisy version of the Spiegler (2016) model, we do not only rely on qualitative treatment effect comparisons, but we also perform a structural out-of-sample prediction exercise that tests the quantitative accuracy of the model. The impact of noise is nontrivial because noisy choice does not only exert a direct effect on action probabilities, but also an indirect effect that arises due to the fact that the action probability affects the strength of the causal effects that the subject infers. Through this mechanism, equilibrium mechanics can amplify the impact of noise.

We estimate two empirical attenuation channels and then imposes a fixed-point condition that determines equilibria in the Reverse Causality and Confounded Choice problems. The first channel maps the data the subject has actually observed into the subject's reported beliefs about these effects. The second channel maps perceived causal effects into action frequencies.

We perform this exercise in two ways. In the first, we use the data from the two-variable case to estimate both maps, and derive the corresponding equilibrium predictions about the three-variable case (Reverse Causality and Confounded Choice). In the second, we use data from within the Reverse Causality case to estimate the mappings and make predictions in that case, and we use data from within the Confounded Choice case to estimate the mappings and make predictions in that

case. To estimate the mapping from beliefs to actions we use instrumental variable approaches to address the attenuation bias that would otherwise occur due to potential noise in belief elicitations.

References

[1] McKelvey, Richard D., and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10(1): 6–38.

[2] Spiegler, Ran. 2016. "Bayesian Networks and Boundedly Rational Expectations." *Quarterly Journal of Economics* 131(3): 1243–1290.