Behavioral Causal Inference*

Ran Spiegler†

March 25, 2024

Abstract

When inferring causal effects from correlational data, a common practice — by professional researchers but also lay people — is to control for potential confounders. Inappropriate controls produce erroneous causal inferences. I model decision-makers who use observational data to learn actions' causal effect on payoff-relevant outcomes. Different decision-maker types use different controls. Their resulting choices affect the very correlations they learn from, thus calling for equilibrium analysis of the steady-state welfare cost of using bad controls. I obtain tight upper bounds on this cost. Equilibrium forces drastically reduce it when types' sets of controls contain one another.

*Financial support by ISF grant no. 320/21 and the Foerder Institute is gratefully acknowledged. I thank Alex Clyde, Nathan Hancart, Heidi Thysen, numerous seminar participants, and referees of a previous (rejected) version, for helpful comments. I am especially grateful to Omer Tamuz for his help with the proof of one of the results.

†Tel Aviv University and University College London

1 Introduction

Learning causal effects from observational data is an important economic activity. Indeed, applied economists do it for a living. However, even lay decision makers regularly perform this activity to evaluate the consequences of their actions. They obtain data about observed correlations among variables (via first- or second-hand experience, or from the media) and try to extract causal lessons from the data. Which college degree will improve their long-run economic prospects? Will wearing surgical masks on airplanes lower their chances of catching a virus? Is coffee drinking good for one's health?

There are two main differences between causal inference from observational data as practiced by professional researchers and lay decision makers. First, the researcher employs sophisticated inference methods that are subjected to stringent scrutiny by other professionals. In contrast, lay decision makers use intuitive, elementary methods, and they do not face pushback when they employ these methods inappropriately. Sometimes they simply follow the advice of methodologically flawed research (or the media's misrepresentation of sound research). Second, while the professional researcher is an outside observer, lay decision makers interact with the economic system in question; the aggregate behavior that results from their causal inferences can affect the very correlations from which they draw these inferences. It is therefore apt to refer to the kind of causal inference that lay decision makers engage in as "behavioral", in both senses of the word.

This paper is an attempt to model "behavioral causal inference". I study a decision maker (DM) who faces a choice between two actions (0 and 1). The DM's choice is based on his belief regarding the action's causal effect on a payoff-relevant outcome. Using an intuitive causal-inference method, the DM extracts this causal belief from long-run correlational data about actions, outcomes and a collection of exogenous variables. The data is generated by the behavior of other DMs in similar situations. In equilibrium, the DMs' behavior is consistent with best-replying to their causal belief.

The intuitive method of causal inference that the DM in my model employs is very simple: Measuring the observed correlation between actions and outcomes, while *controlling* for some set of exogenous variables. This is a basic and widespread procedure in scientific data analysis, but it is based

on a simple idea that lay people practice to some extent. For example, when an agent decides whether to wear a surgical mask for protection against viral infection, it is natural for him to look for infection statistics about people in his own age group. Likewise, when a student choosing a college major tries to evaluate future earnings by STEM and non-STEM graduates, it is natural for him to focus on people who share his high school math background. In both cases, when the agent consults data to estimate the consequences of various actions, he may try to focus on data points that share his own characteristics — if he has access to such fine-grained data. Indeed, we should expect heterogeneity in this regard: Agents may differ in what they feel a need to control for, as well as in their access to data that enable controlling.

In general, suppose that long-run observational data is given by some joint probability distribution p over actions a, outcomes y, and a collection of exogenous variables $x = (x_1, ..., x_K)$. The DM is able to control for the variables indexed by $C \subseteq \{1, ..., K\}$. His estimated causal effect of a on y is

$$E_p(y \mid a = 1, x_C) - E_p(y \mid a = 0, x_C) \tag{1}$$

When the set C of control variables differs from what an outside researcher would deem appropriate, the DM's causal inference can be wrong: he may misread the causal meaning of observed correlations, and consequently obtain a biased estimate of the causal effect of a on y (Angrist and Pischke (2009), Cinelli et al. (2022)). The bias can be large, if x is strongly correlated with both a and y.

However, when the correlation between a and x (as given by the conditional distribution $p(a \mid x)$) reflects the aggregate behavior of DMs facing the same situation, it also reflects their subjective optimization, induced by the causal inferences they draw from p. This raises the following question: What are the limits to the DM's errors of causal inference due to bad controls, when the data-generating process p has to be consistent with *equilibrium behavior* — i.e., when the DM's choice is optimal with respect to the subjective belief he extracts from p using his causal-inference procedure? Could this equilibrium condition change our conclusions regarding the maximal welfare cost of using bad controls?

I approach this question with a simple model, in which a DM chooses a

after the exogenous variables $t, x_1, ..., x_K$ are realized, where $t \in \{0, 1\}$ is the DM's preference type. The payoff-relevant outcome y lies in $[0, 1]$. The DM's vNM utility function is $u(a, t, y) = y - \theta \cdot \mathbf{1}[a \neq t]$. Thus, t indicates the DM's favorite action, and θ is the cost he incurs when he does not take it. The DM will only do so if he thinks that a has a beneficial causal effect on y. In the baseline model, I assume that the actual effect is *null*: y is determined only by the exogenous variables. This restriction is made for expositional convenience (Appendix II extends the analysis straightforwardly whenever a has an additively separable causal effect on y).

In addition to the variation in preference types, there is heterogeneity among DMs in terms of how they perform causal inference. The DM's "*data type*" is defined by his set of control variables C, which is drawn independently from some given set N. A DM of type C forms an estimated causal effect of a on y (given x) according to (1). The formula is evaluated according to an objective joint distribution p over all variables. The DM observes the realization of t prior to his decision. However, since t is a *private* preference type, I assume there is no statistical data about t, and therefore the DM never uses it for his causal estimates (this entails no loss of generality for my worst-case analysis, because one of the x variables can be a perfect proxy for t). The objective distribution of a conditional on (t, x) describes the *aggregate* behavior of the DM population, which arises from the strategies of all DM types. In equilibrium, each type's strategy prescribes best-replies to his causal belief.

The basic insight of this paper is that this equilibrium condition can drastically lower (and sometimes eliminate altogether) the welfare loss due to errors of causal inference. These errors consist of misreading the causal component of observed correlational patterns. Agents' response to their beliefs change these very patterns, and hence the causal lessons they draw from them. As a result, unlike academic researchers, DMs who perform faulty causal inference "in the field" need not suffer dire consequences.

Example 1.1: Investing in one's education
Suppose that $t = 0$ with certainty — i.e., preferences are homogeneous. There is a single x variable. There are two possible data types in this environment: type 1 controls for x (hence, he conditions his action on x); whereas type 2 does not (hence, he does not condition his action on x). Both types

rely on the same aggregate "database" given by the joint distribution p over x, a, y to form causal beliefs. An economic story behind this scenario is that a, y and x represent personal educational investment, future earnings, and a demographic characteristic. All agents in the population have an intrinsic disutility from education, and will only make the investment if they think it improves future earnings. Agents of type 1 are "sophisticated" in the sense that they control for demographic characteristics when trying to infer the causal effect of education on earnings, whereas agents of type 2 are "naive" in the sense that they do not control for any exogenous variable.

Since type 1 controls for x, he correctly estimates a null causal effect of a on y. This type plays $a = t = 0$ regardless of x — i.e., he ends up not varying his action with x. Type 2 potentially commits an error of causal inference because he fails to control for x, and therefore interprets any empirical correlation between a and y as a causal effect. However, this type, too, does not vary his action with x by definition. It follows that *none* of the two types vary their actions with x. Thus, if p is consistent with equilibrium, a and x are independent, thus destroying any possibility of x acting as a confounder of the relation between a and y. In the absence of confounding, failure to control for x does not result in an error of causal inference, as a and y are statistically independent. It follows that under the equilibrium restriction that $p(a \mid x)$ reflects data types' subjective optimization with respect to their causal beliefs, the DM incurs *no* welfare loss due to bad controls. By comparison, the maximal loss without the equilibrium condition is 1 (since the values that a and y take are bounded between 0 and 1). □

The main results in this paper explore the generality of this observation. I examine various families of joint distributions over $t, x_1, ..., x_K, y$, and characterize the upper bound on the DM's equilibrium welfare loss relative to the expected payoff from the rational-expectations strategy $a \equiv t$. When the objective causal effect of a on y is null, the welfare loss is simply $\theta \cdot \Pr(a \neq t)$.

The analysis involves some structure on the set of data types. In many areas of economic theory, we use typologies of economic agents that impose some "vertical" order on types. In mechanism design, we often order preference types according to the single-crossing property. Likewise, information types in economic models are often ordered by the monotone-likelihood-ratio property. In the present context, a natural vertical ordering of data types

is via *set inclusion*: $C_1 \supset C_2 \supset \cdots \supset C_n$. In this case, lower-indexed types control for larger sets of variables; they are closer to the ideal of controlling for all potential confounders. In a naive sense, they are more "sophisticated" (we will have opportunities to be reminded that controlling for more variables is not necessarily a good thing).

The distinction between vertically ordered and unordered data-type spaces turns out to be crucial for the upper bounds on the welfare loss due to bad controls. In Section 3, I consider the case of homogenous preferences (i.e., there is no variation in t). When data types are vertically ordered, the equilibrium welfare loss is *zero* — that is, the equilibrium condition fully protects the DM from choice errors arising from the use of bad controls. It does so by shutting down the channels through which the choice behavior of some types confound the relation between other types' actions and y. Conversely, when data types are not vertically ordered, the tight upper bound on the DM's welfare loss (when we are free to set the value of θ and the data-type distribution) is 1 — the same as when no equilibrium requirement is imposed.

In Section 4, I obtain a softer version of this "bang-bang" result when there *is* variation in t. Here I restrict attention to distributions in which t is the sole true cause of y (and x variables are thus observable proxies for t). For a vertically ordered data-type space, the tight upper bound on the DM's equilibrium welfare loss is $\Pr(t=1)\cdot\Pr(t=0)$. When types are not ordered, the tight upper bound is $\max\{\Pr(t=1), \Pr(t=0)\}$. When in addition we relax the assumption that y is independent of x conditional on t, the tight upper bound is 1.

The raw intuition behind these results is that when DMs systematically best-reply to their causal belief, they attenuate the correlational patterns that lend themselves to causal misinterpretation. In Example 1.1, this effect took a very simple form: The type that conditions on x can in principle correlate his actions with x, thus creating a confounding pattern that leads the other type (who fails to control for x) astray. However, his individual best-replying rules out this correlation, thus "protecting" the other type. The general lesson is that there is a big difference between errors of causal inference that are committed by an outside observer and those that are committed by agents with skin in the game. However, this lesson critically

relies on the vertical ordering of data types.

In Section 5, I enrich the notion of data types, such that DMs are allowed to control for variables they do not condition on. For instance, in the surgical-mask example mentioned in the opening paragraph, a DM may have access to statistical data about the prevalence of certain genes and their correlation with viral infection, without knowing his own genome. The DM can then control for genetic variables without conditioning on them, by adjusting for their correlation with the variables he does condition on. A data type in this extended environment consists of the set of variables on which the type has statistical data, and the subset of variables he conditions on. I extend the analysis of the homogenous-preference case to this environment, via an appropriate generalization of the notion of vertically ordered types.

This paper continues my line of research into decision making under imperfect causal reasoning (Spiegler 2016,2020). The problem it poses — quantifying the decision costs of faulty causal inference — is novel and lacks precedents in the literature. The paper also introduces a modeling innovation. While prior work has focused on DMs who misperceive the causal mapping from actions to consequences, the DM in this paper effectively fails to perceive that his *own* actions have direct causes that confound the relation between actions and consequences.[1] Moreover, DM types *differ* in their understanding of these causes, via their different sets of controls. This heterogeneity is what makes the problem of quantifying the decision costs of bad controls technically non-trivial.

Appendix III shows that existing frameworks for equilibrium modeling with non-rational expectations (Jehiel (2005), Spiegler (2016), Esponda and Pouzo (2016)) can be adapted to incorporate the novel features, thus recasting the present model in (modified) existing languages. The reason I chose to present the model in a *new* language is twofold. First, it is relatively simple and self-contained, and therefore does not require familiarity with previous frameworks. Second, by drawing a connection with the familiar and intuitive notion of "bad controls" and the work habits of empirical researchers, this paper will hopefully help inspiring new research about how everyday DMs perform causal inference.

[1] Clyde (2023) effectively shares this feature, by assuming that the DM forms equilibrium beliefs on the basis of data about *proxies* of states or actions, rather than data about the variables themselves.

2 A Model

Let $a \in A = \{0, 1\}$ be an *action* that a decision maker (DM) chooses. Let $t \in \{0, 1\}$ be the DM's *preference type*. Let $y \in Y \subset [0, 1]$ be an *outcome*. Let $x = (x_1, ..., x_K)$ be a collection of exogenous variables that are realized jointly with t, prior to the realization of a and y. Let X_k denote the finite set of values that x_k can take. For every $M \subseteq \{1, ..., K\}$, denote $x_M = (x_k)_{k \in M}$ and $X_M = \times_{k \in M} X_k$. I assume that x and t are the only potential causes of y — i.e., a has *no causal effect* on y. This assumption is made for expositional clarity; I will relax it in Appendix II.

The DM's vNM utility function is $u(t, a, y) = y - \theta \cdot \mathbf{1}[a \neq t]$, where $\theta \in (0, 1)$ is a constant. Thus, the DM has an intrinsic motive to match his action to his preference type; he will choose $a \neq t$ only if he believes this increases the expected value of y. If the DM understood that a has no causal effect on y, he would always choose $a = t$.

There is a set $N = \{1, ..., n\}$ of DM *data types*. Each type $i \in N$ is associated with a *distinct* subset $C_i \subseteq \{1, ..., K\}$. I refer to C_i as type i's set of control variables. The interpretation is that type i observes the realization of x_{C_i} prior to making his decision; he also has access to "public" data about the long-run statistical behavior of these variables (jointly with a and y); and he believes that in order to learn the causal effect of a on y, he should control for these variables.

As far as variables outside C_i are concerned, the DM type either lacks data on them, or he thinks they are irrelevant and therefore need not be controlled for. The model does not accommodate variables that are caused by a or y as possible controls — it only focuses on exogenous, "pre-treatment" controls. Note that t never belongs to the DM's set of control variables. The interpretation is that t is a *private* preference type, and as such it is unlikely to enter publicly available datasets. However, note that we can always allow one of the variables x_i to be a copy of t; in this sense, the assumption does not rule out the possibility of effectively controlling for t.

Let $\lambda \in \Delta(N)$ be the distribution of data types in the DM population. This distribution is *independent* of all variables (this assumption is immaterial for the results in Section 3 but plays a role in Section 4). A strategy for type (t, i) is a function $\sigma_{t,i} : X_{C_i} \to \Delta(A)$.

Let p be a joint probability distribution over t, x, a, y. I interpret p as a steady-state or long-run distribution. Data type i knows $p(x_{C_i}, a, y)$ — this is what "having access to long-run data" about x_{C_i}, a and y means. Denote $\gamma = p(t = 1)$. The assumption that a has no causal effect on y means that p satisfies the conditional-independence property $y \perp a \mid (t, x)$, and hence factorizable as follows:

$$p(t, x, a, y) = p(t, x)p(a \mid t, x)p(y \mid t, x)$$

where the term $p(a \mid t, x)$ represents the DM's average behavior across data types:

$$p(a \mid t, x) = \sum_{i \in N} \lambda_i \sigma_{t,i}(a \mid t, x_{C_i})$$

This term is endogenous, whereas $p(t, x)$ and $p(y \mid t, x)$ are exogenous.

Since a DM of data type i believes that C_i is a valid set of controls, he regards $p(y \mid a, x_{C_i})$ as a proper estimate of the probabilistic consequence of choosing a, given his observation of x_{C_i} His perceived causal effect of a on y given x is

$$\Delta_i(x) = E_p(y \mid a = 1, x_{C_i}) - E_p(y \mid a = 0, x_{C_i}) \tag{2}$$

If the DM had long-run data about all exogenous variables (including t), he could control for all of them, and thus correctly infer the action's null causal effect. This is the *rational-expectations benchmark* for this model. In contrast, our DM may end up believing that a has a non-zero causal effect on y because he fails to control for some exogenous variables. In this case, he misinterprets part of the correlation between a and y as a causal effect, whereas in reality this correlation is entirely due to confounding by t, x. What makes the model non-trivial is that these confounding patterns are *endogenous*, as they are affected by the strategy profile. Specifically, $E_p(y \mid a, x_{C_i})$ is not invariant to σ, since σ determines how a varies with the unconditioned exogenous variables, t and x_{-C_i}.

Expression (2) has the appearance of an expected-utility calculation by a standard Savage DM who receives a signal x_{C_i}. There is a fundamental difference, however, arising from the endogeneity of p and from its interpretation as empirical frequencies from which the DM draws causal inferences. Therefore, I refrain from referring to x_{C_i} as a signal, and further discuss the

connection to the Savage framework in Appendix III.

As controlling for all exogenous variables is an ideal of correct causal inference, it is natural to seek to order data types in terms of how far they are from this ideal.

Definition 1 (Vertically ordered types) *The set of data types N is vertically ordered if types can be enumerated such that $C_1 \supset \cdots \supset C_n$.*

When N is vertically ordered, lower-indexed data types control for a larger set of variables. In particular, type i controls for every variable that type $j > i$ conditions on.

I now introduce the notion of equilibrium behavior.

Definition 2 (Equilibrium) *Let $\varepsilon > 0$. A strategy profile $\sigma = (\sigma_1, ..., \sigma_n)$ is an ε-equilibrium if for every $i = 1, ..., n$ and every t, x, a', $\sigma_i(a' \mid t, x) > \varepsilon$ only if*

$$a' \in \arg\max_a \ \{E_p(y \mid a, x_{C_i}) - \theta \cdot \mathbf{1}[a \neq t]\}$$

An equilibrium is a limit of a sequence of ε-equilibria for $\varepsilon \to 0$.

The trembling-hand aspect of the equilibrium concept ensures that all the conditional probabilities it involves are well-defined. Trembles do not play a role in the characterization results, with the exception of Proposition 4.

The structure of u means that in equilibrium, type i will play $a \neq t$ with positive probability at x only if $|\Delta_i(x)| \geq \theta$. Since a has no causal effect on y, playing $a \neq t$ yields a welfare loss.

Definition 3 (Expected welfare loss) *Given a strategy profile σ, the DM's expected welfare loss is*

$$\theta \sum_{t,x} p(t,x) \sum_{i \in N} \lambda_i \sigma_i(a \neq t \mid t, x) \tag{3}$$

My main task in the next sections will be to derive *upper bounds* on this quantity when σ is required to be *an equilibrium*. Without this equilibrium

condition, the upper bound is 1. To illustrate why, suppose that $t = 0$ with certainty, and that $x \in \{0, 1\}$. Assume $y = x$ with certainty for every x, and consider the strategy σ that prescribes $a = x$ with probability one. By definition, the probability of error is $p(x = 1)$. If $p(x = 1) \approx 1$ and $\theta \approx 1$, the welfare loss is approximately 1. However, the strategy σ is inconsistent with equilibrium. If a data type i varies his action with x, then he controls for it and correctly estimates the null causal effect of a. As a result, he will always play $a = 0$, a contradiction. It follows that the requirement that σ is an equilibrium strategy can have bite.

Comment: The rationality benchmark. The rational-expectations benchmark for this model is a DM who controls for t and x. What would be a "rational" mode of behavior for a DM given that he only has data about a subset of potential confounders, given by C? The standard Bayesian model assumes that in this case, the DM has a subjective prior belief over (t, x) and updates this belief according to the signal x_C. If the DM correctly believes that the mapping from (t, x, a) to y is constant in a, then he will always play $a = t$, regardless of his signal — as in the rational-expectations benchmark. In contrast, the DM in our model ignores the variables he does not control for. Equivalently, he assumes they are independent of all other variables.

Comment: A "persuasion" interpretation. Worst-case analysis of the DM's welfare can be interpreted through the prism of the small literature on persuading boundedly rational agents (e.g., Glazer and Rubinstein (2012), Galperti (2019), Hagenbach and Koessler (2020), Schwartzstein and Sunderam (2021), Eliaz et al. (2021b), and De Barreda et al. (2022)). Under this interpretation, the DM is the receiver who takes an action. The sender's objective is to maximize the probability that the receiver plays $a \neq t$. Toward this end, he designs a distribution over the variables the receiver observes as signals. This is a conventional "information design" tool. The unconventional aspect of this tool is that it also determines the long-run statistical data that the receiver uses to form his belief. Worst-case analysis can thus be viewed as finding the sender's optimal information-cum-data provision strategy.

3 Analysis: Homogenous Preferences

This section characterizes the maximal welfare loss that is consistent with equilibrium behavior, when there is no variation in the preference type t. Specifically, assume that $t = 0$ with probability one (i.e., $\gamma = 0$), such that the DM's expected welfare loss is simply $\theta \cdot \Pr(a = 1)$. In this environment of preference homogeneity, the only potential source of variation in the DM's behavior is the way the various types condition their actions on x.

For any set N of data types, there is an equilibrium in which the DM plays $a = 0$ with probability one. To see why, construct the perturbation of this strategy: Each data type i plays $a = 1$ with probability $\varepsilon \approx 0$, independently of x_{C_i}. By construction, $a \perp x$ under this strategy profile, and therefore $\Delta_i(x) = 0$ for every type i, such that $a = 0$ is the type's unique best-reply. The question is whether there are additional equilibria, in which the DM commits an error with positive probability.

Example 1.1 presented a specification with homogenous preferences, in which the equilibrium requirement completely eliminated the possibility of decision errors. Our first result establishes that this is a general feature of vertically ordered data-type spaces. The results in this section are special cases of results reported in Section 5, which are based on a generalization of the notion of vertically ordered types. Like all the results in this paper, they are proved in Appendix I.

Proposition 1 *Let $\gamma = 0$. Suppose N is vertically ordered. Then, the unique equilibrium is for all DM types to play $a = 0$ with probability one. In particular, the DM's expected welfare loss is zero.*

Thus, when $\gamma = 0$ and data types are vertically ordered, the equilibrium requirement fully "protects" the DM from choice errors due to bad controls. It does so by shutting down the channels through which the choice behavior of some data types could confound the relation between other types' actions and y. Type 1 effectively controls for all sources of correlation between a and y. Even when he fails to control for some exogenous variables, this does not matter because no other type conditions on them, hence they generate no confounding effect. As a result, type 1's subjective best-replying generates no variation in choice behavior. This means that type 2 effectively controls

for all potential confounders — which would not be the case if we did not impose the equilibrium condition on type 1's behavior. This equilibrium effect spreads through the entire type space. The proof formalizes this intuition via induction on the set of data types.

How important is the vertical ordering of data types for Proposition 1? The following example begins to address this question.

Example 3.1: Analysts with diverse expertise
Let $K = 2$. All variables take values in $\{0, 1\}$, and their joint distribution satisfies:[2]

$$\begin{aligned} p(x_1 &= 1) = p(x_2 = 1) = \beta \in (0, 1) \\ p(x_2 &= 1 \mid x_1 = 1) = p(x_1 = 1 \mid x_2 = 1) = q \in [\frac{1}{2}, 1) \\ p(y &= 1 \mid x_1, x_2) \equiv x_1 x_2 \end{aligned}$$

Let $n = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, $C_i = \{i\}$.

The following is an economic story behind this specification. A firm's environment is defined by financial and technological factors (represented by x_1 and x_2). The firm is profitable as long as both factors are favorable. The firm's decision is guided by business analysis. There are two kinds of analysts, who specialize in different aspects. Some firms base their decisions on a financial analyst, while others base their decisions on a technological analyst. Firms' analysts rely on the same aggregate data arising from the decisions of both types of firms, but each analyst has tunnel vision and neglects the aspect outside his area of expertise. This is an instance of "horizontal" differentiation between data types.

Consider the following strategy profile: Each type $i = 1, 2$ always plays $a = x_i$. I will show that this profile constitutes an equilibrium. Begin by calculating type 1's subjective estimate of actions' causal effect on profits, given his information. Observe that since $y = x_1 x_2$ independently of a,

$$\begin{aligned} p(y &= 1 \mid a, x_1 = 1) = p(x_2 = 1 \mid a, x_1 = 1) \\ p(y &= 1 \mid a, x_1 = 0) = 0 \end{aligned}$$

[2] Presenting these marginal and conditional distributions suffices and is more convenient for our purposes; there is no need for full specification of p.

for every a. Note that these quantities never involve conditioning on a zero-probability event. For example, the combination $a = 0, x_1 = 1$ arises when $x_2 = 0$ and the DM is of type 2.

Thus, we only need to calculate two conditional probabilities, given the DM's postulated strategy. First, $p(x_2 = 1 \mid a = 1, x_1 = 1)$ is equal to

$$\begin{aligned} & \frac{p(x_1 = 1)p(x_2 = 1 \mid x_1 = 1)p(a = 1 \mid x_1 = 1, x_2 = 1)}{p(x_1 = 1)\sum_{x_2} p(x_2 \mid x_1 = 1)p(a = 1 \mid x_1 = 1, x_2)} \\ = & \frac{q(\lambda_1 \cdot 1 + \lambda_2 \cdot 1)}{q(\lambda_1 \cdot 1 + \lambda_2 \cdot 1) + (1-q)(\lambda_1 \cdot 1 + \lambda_2 \cdot 0)} \\ = & \frac{q}{q + \frac{1}{2}(1-q)} \end{aligned}$$

Second,

$$p(x_2 = 1 \mid a = 0, x_1 = 1) = 0$$

since the combination $(a = 0, x_1 = 1)$ cannot arise when $x_2 = 1$, given the strategy profile. It follows that

$$\Delta_1(x_1 = 1) = \frac{q}{q + \frac{1}{2}(1-q)} - 0 = \frac{2q}{1+q}$$

If $2q/(1+q) > \theta$, type 1 will prefer to play $a = 1$ when $x_1 = 1$. In addition, we established that $\Delta_1(x_1 = 0) = 0 - 0 = 0$. Therefore, type 1 will prefer to play $a = 0$ when $x_1 = 0$. The same calculations apply to type 2.

It follows that as long as $q > \theta/(2-\theta)$, the postulated strategy profile is an equilibrium. The equilibrium error probability (i.e., $\Pr(a = 1)$) is β, which can be arbitrarily close to one — hence, the equilibrium welfare loss can be as large as the non-equilibrium benchmark. In this case, equilibrium forces do not "protect" DMs from their errors of causal inference.

The intuition behind this result is that since type i varies his action with x_i yet fails to control for x_j, each type creates a confounding effect that "fools" the other type. For example, type 1 is vulnerable to interpreting the residual correlation between a and y after controlling for x_1 — which exists because of type 2's strategy — as a causal effect. This residual correlation can be seen from our calculation of $p(y = 1 \mid a, x_1 = 1)$.

The result does *not* necessitate correlation between x_1 and x_2. Even when $q = \frac{1}{2}$, the above equilibrium can be sustained as long as $\theta < \frac{2}{3}$. The

reason is that although the DM types in this case condition their actions on independent exogenous variables, their subjective causal estimates involve conditioning on a (a variable whose distribution records the DM's aggregate behavior). Since this variable is a common consequence of x_1 and x_2, conditioning on it creates correlation between otherwise independent variables.

The equilibrium welfare loss is non-monotone with respect to the data types' sets of control variables. For example, suppose $C_1 = \{1\}$ and $C_2 = \emptyset$ — i.e., type 2 now does not control for any variable. In this case, the type space is vertically ordered; and by Proposition 1, neither DM type will commit an error in equilibrium. It follows that expanding one type's set of control variables can be detrimental for all types' welfare. □

The following result generalizes this example: Whenever the data-type space is not vertically ordered, the tight upper bound on the DM's equilibrium welfare loss is 1.

Proposition 2 *Let $\gamma = 0$. Suppose N is not vertically ordered. Then, for any $\theta, \beta \in (0,1)$, there exist λ and $(p(x,y))$ such that $\Pr(a=1) > \beta$ in some equilibrium. In particular, when $\theta \approx 1$, the equilibrium welfare loss can be arbitrarily close to 1.*

Thus, when types are not vertically ordered, equilibrium forces do not curb the welfare loss due to faulty causal inference. The reason is that the equilibrium behavior of different types can create confounding patterns that mutually sustain their inference errors.

4 Analysis: Heterogeneous Preferences

In this section I reintroduce preference heterogeneity, by assuming $\gamma \in (0,1)$. The significance of this degree of freedom is that it implies an intrinsic motive for the DM to vary his behavior with an exogenous variable. By comparison, in the homogenous-preference case, the DM would vary his behavior with an exogenous variable only when he (erroneously) concluded that it influences the causal effect of a on y. Denote $\delta_t = E_p(y \mid t)$. Without loss of generality, assume $\delta_1 \geq \delta_0$.

Example 4.1: Choosing a college major
Let $y \in \{0,1\}$. Suppose $\delta_t = p(y = 1 \mid t) = t$. Thus, $y = 1$ if and only if $t = 1$. Let $K = 0$ and $n = 1$ — i.e., there is a unique data type, $C = \emptyset$. I assume $\gamma \neq \theta$, to rule out an annoying knife-edge case.

One economic story behind this specification is that a represents a student's decision whether to select a math-intensive major in college; t indicates whether he likes math; and y represents his subsequent earnings. The student learns the correlation between a and y. He has no access to control variables, and ends up treating the correlation as causal. The assumption that $\delta_t \equiv t$ means that fondness for math is perfectly correlated with math skills that determine earnings, independently of the student's decision.

I will now establish uniqueness of equilibrium in this setting, and characterize the DM's equilibrium welfare loss. The DM's estimated causal effect of a on y is

$$\Delta = p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0)$$

Denote $\alpha_t = \sigma(a = 1 \mid t)$. By the DM's preferences, $\alpha_1 \geq \alpha_0$. Now obtain explicit expressions for the terms that define Δ:

$$\begin{aligned} p(y &= 1 \mid a = 1) = \frac{\gamma \cdot \alpha_1 \cdot \delta_1 + (1-\gamma) \cdot \alpha_0 \cdot \delta_0}{\gamma \cdot \alpha_1 + (1-\gamma) \cdot \alpha_0} \\ p(y &= 1 \mid a = 0) = \frac{\gamma \cdot (1-\alpha_1) \cdot \delta_1 + (1-\gamma) \cdot (1-\alpha_0) \cdot \delta_0}{\gamma \cdot (1-\alpha_1) + (1-\gamma) \cdot (1-\alpha_0)} \end{aligned}$$

A simple calculation establishes that since $\delta_1 > \delta_0$ and $\alpha_1 \geq \alpha_0$, we must have $\Delta \geq 0$. This in turn implies that $\alpha_1 \geq 1 - \varepsilon$ in ε-equilibrium, because when $t = 1$, the DM perceives no conflict between his intrinsic taste for $a = t$ and the estimated effect of his choice on y. Plugging the known expressions for α_1 and δ_t and taking the $\varepsilon \to 0$ limit, we obtain

$$\Delta = \frac{\gamma}{\gamma + (1-\gamma) \cdot \alpha_0}$$

If $\alpha_0 \leq \varepsilon$ in ε-equilibrium, then $\Delta \to 1$ in the $\varepsilon \to 0$ limit. But then, $\Delta > \theta$, hence playing $a = 1$ at $t = 0$ is subjectively optimal, in contradiction to $\alpha_0 \leq \varepsilon$. It follows that $\alpha_0 > 0$ in equilibrium. There are two cases to consider. First, suppose $\alpha_0 \in (0,1)$. This requires $\Delta = \theta$ (and therefore

$\gamma < \theta$), such that

$$\alpha_0 = \frac{\gamma(1-\theta)}{(1-\gamma)\theta}$$

Since the DM only commits an error in equilibrium when $t = 0$, his expected equilibrium welfare loss is

$$\theta \cdot (1-\gamma) \cdot \alpha_0 = \gamma(1-\theta) < \gamma(1-\gamma)$$

By setting $\theta \approx \gamma$, we can get arbitrarily close to the upper bound of $\gamma(1-\gamma)$.

Second, suppose $\alpha_0 = 1$. This requires us to sustain this equilibrium with suitable trembles. Specifically, suppose $\alpha_1 = 1 - \varepsilon^2$ and $\alpha_0 = 1 - \varepsilon$. As $\varepsilon \to 0$, we obtain $p(y = 1 \mid a = 1) \approx \gamma$ and $p(y = 1 \mid a = 1) \approx 0$. If $\gamma > \theta$, this is consistent with equilibrium. The DM's welfare loss in this equilibrium is

$$\theta \cdot (1-\gamma) \cdot 1 < \gamma(1-\gamma)$$

Again, by setting $\theta \approx \gamma$, we can get arbitrarily close to the upper bound.

Thus, for any configuration of θ and γ, there is a unique equilibrium in this setting. The DM's equilibrium welfare loss in this equilibrium is always below $\gamma(1-\gamma)$. This bound can be approximated arbitrarily well by setting $\theta \approx \gamma$. The trembling-hand aspect of our equilibrium concept is not necessary for the upper bound.

As in earlier examples, equilibrium forces in Example 4.1 "protect" the DM from causal errors, by pushing his welfare loss below $\gamma(1-\gamma)$ — compared with the non-equilibrium benchmark of 1. The intuition is as follows. The DM mistakes the correlation between a and y for a causal effect. This correlation is large when a varies strongly with t; it hits the maximal level when a always coincides with t. However, that extreme case is precisely when the DM commits *no* error. At the other extreme, if the DM almost always plays $a = 1$ because his estimated causal effect of a on y is above θ, the frequency of the DM's error is maximal. However, since in this case a varies little with y, the estimated causal effect is small. In general, a larger estimated causal effect goes hand in hand with a lower equilibrium frequency of making a decision error. This is why equilibrium behavior limits the DM's expected cost of failing to control for x. □

Let us now turn to a characterization of the upper bound on the DM's equilibrium welfare loss, for a restricted domain of data-generating processes. Specifically, I assume that $p(y \mid t, x) \equiv p(y \mid t)$ — i.e., $y \perp x \mid t$. This fits situations in which the DM's preference type is a sufficient statistic for determining the outcome; the x variables are merely observable correlates of this statistic. For instance, whether a student regards studying as a costly or pleasurable activity is the cause of her school performance. This latent attitude may be correlated with observable characteristics, but these are only indirect causes, or mere proxies for the true cause.

Proposition 3 *Suppose N is vertically ordered. If $y \perp x \mid t$, then the DM's expected welfare loss in equilibrium is at most $\gamma(1-\gamma)$.*

Example 4.1 established the tightness of this upper bound. Proposition 3 also means that across all distributions that satisfy $y \perp x \mid t$, the expected welfare loss is at most $\frac{1}{4}$ — compared with the non-equilibrium upper bound of 1. When $\gamma \to 0$, the loss converges to zero. (This limit case is not a special case of Section 3, because it implies $x \perp y$.)

As in the case of Proposition 1, the proof of Proposition 3 proceeds by induction on the set of data types, starting with type 1, whose set of controls is the largest. Although this type controls for every x variable the other data types condition on, this does not mean he is immune to neglecting confounders, because he fails to control for the preference type t. Furthermore, since this type varies his behavior with t, he exerts a "confounding externality" on the other data types. This makes the inductive proof considerably more intricate. A key argument in the proof is that while data types may disagree on the magnitude of the causal effect of a on y, they all agree on its sign, which is positive (since $\delta_1 > \delta_0$). This feature holds in any equilibrium when the type space is vertically ordered.

Example 4.2: Choosing a college major, revisited
To further illustrate the heterogeneous-preference model, enrich Example 4.1 by letting $K = 1$ and $n = 2$. The exogenous variable is $x \in \{0, 1\}$. This is an observable proxy for t, whose conditional distribution is $p(x = t \mid t) = q \in (\frac{1}{2}, 1)$ for every t. The two data types are $C_1 = \{1\}$ and $C_2 = \emptyset$. That is, type 1 controls for x while type 2 does not. The DM population includes both

types: $\lambda_1, \lambda_2 > 0$. In the context of the college-major story, x may represent the student's high-school math performance. The "sophisticated" type 1 has data that enables him to control for x when estimating the correlation between major choice and subsequent earnings. The "naive" type either lacks the data or finds it irrelevant.

In this environment, it is natural to predict that unlike the naive type, the sophisticated type's ability to control for x may protect him from falsely inferring that a causes y. Thus, let us postulate the following strategy profile: Type 1 always plays $a = t$; while type 2 plays $a = 1$ with probability one (α) when $t = 1$ (0). Under this strategy profile, only the naive type can ever commit decision errors. Let us check whether this kind of strategy profile can be an equilibrium.

Note that $p(y = 1 \mid a = 0, x) = 0$ for every x — hence, $p(y = 1 \mid a = 0) = 0$ — since $y \equiv t$ and neither type ever plays $a = 0$ when $t = 1$. Therefore, $\Delta_2 = p(y = 1 \mid a = 1)$ and $\Delta_1(x) = p(y = 1 \mid a = 1, x)$. In addition,

$$p(y = 1 \mid a = 1) = \frac{\gamma \cdot 1}{\gamma \cdot 1 + (1 - \gamma) \cdot \lambda_2 \cdot \alpha} \tag{4}$$

and

$$p(y = 1 \mid a = 1, x = 1) = \frac{\gamma \cdot q \cdot 1}{\gamma \cdot q \cdot 1 + (1 - \gamma) \cdot (1 - q) \cdot \lambda_2 \cdot \alpha} \tag{5}$$

If $\alpha = 0$ (such that neither type ever plays $a \neq t$), $\Delta_2 = 1 > \theta$, hence type 2 wants to deviate to $a = 1$ at $t = 0$, a contradiction. Now suppose $\alpha > 0$. Then, $\Delta_2 \geq \theta$. Since type 1 plays $a = t$ at $x = 1$, $\Delta_1(x = 1) \leq \theta$. Plugging (4)-(5) in these inequalities, we obtain a contradiction.

It follows that there is no equilibrium in which type 1 never plays $a \neq t$. No matter how accurate x is as a proxy of t, it cannot fully protect the sophisticated type from his failure to control for t, when equilibrium effects are taken into account. □

When data types are not vertically ordered, the tight upper bound on the DM's expected welfare loss (under the restriction $y \perp x \mid t$) is significantly higher.

Proposition 4 *Suppose N is not vertically ordered. If $y \perp x \mid t$, then the DM's expected welfare loss in equilibrium is at most* $\max(\gamma, 1-\gamma)$. *When $|X_k| \geq 3$ for all k, this upper bound can be approximated arbitrarily well, by appropriately selecting θ, λ and $(p(x, y \mid t))$.*

This result carries the relevance of vertical ordering of data types to the heterogeneous-preferences setting. The gap between the upper bounds in the two cases — $\gamma(1-\gamma)$ vs. $\max(\gamma, 1-\gamma)$ — is significant, and gets more so as the preference type distribution becomes more unbalanced. To attain the upper bound given by Proposition 4, I use trembles and also require exogenous x variables to take at least three values. Whether these elements in the construction are indispensable is an open question. Finally, unlike the case of vertically ordered types, different data types may disagree on the causal effect's sign; indeed, this feature plays an important role in my implementation of the upper bound.

The final result in this section lifts all restrictions on $(p(x, y \mid t))$ and the set of data types, and shows that in this case, the gap between equilibrium and non-equilibrium upper bounds on the DM's welfare loss disappears.

Proposition 5 *Suppose N is not vertically ordered. For every $\gamma, \theta \in (0, 1)$, there exist λ and $(p(x, y \mid t))$ for which there is an equilibrium in which* $\Pr(a \neq t) = 1$.

The results in this section leave two open problems. First, does the upper bound $\gamma(1-\gamma)$ obtained for vertically ordered types extend to distributions p that violate $y \perp x \mid t$? Second, how do results change when the distribution over data types is allowed to be correlated with t and x?

5 Controlling without Conditioning

So far, we have assumed that the DM conditions on every variable he controls for. This is a natural assumption in many settings — e.g., when x variables are demographic or socioeconomic characteristics. Agents are likely to be informed of their own age, ethnicity and parental education, at least as much as they are likely to know the population-level distribution of these characteristics.

However, in some cases it makes sense to assume that the DM has access to statistical data about variables, without knowing their realization at the moment of choice. For example, a firm may know how its performance is correlated with macroeconomic indicators, yet it need not know their current value when making its business decisions because the indicators are published with delay. In such cases, the DM can still control for such variables, even when he cannot condition on their realization. I refer to this mode of controlling as *adjustment* as opposed to conditioning.

To accommodate this distinction, extend the definition of a data type, so that it consists of a *distinct* pair (C, D) of subsets of $\{1, ..., K\}$, where $C \subseteq D$. The set D represents the type's control variables — i.e., the variables on which he has long-run statistical data (such that he knows their joint distribution with a and y). The set C represents the variables whose realization he learns before making his decision. The assumption that $C \subseteq D$ means that if the DM conditions on a variable, he must have long-run data about it. In principle, one can imagine situations in which agents know the realization of a variable without having data about its long-run statistical behavior. For instance, the DM may know his height but lack access the statistics about how height is correlated with the outcome of interest. However, in the absence of such data, the DM cannot make use of his height information, and therefore, we might as well assume that he lacks it. This is the justification for the assumption that $C \subseteq D$.

The DM's estimated causal effect of switching from $a = 0$ to $a = 1$ (given x) is

$$\Delta_i(x) = \sum_{x_{D \setminus C}} p(x_{D \setminus C} \mid x_C) \left[E_p(y \mid a = 1, x_D) - E_p(y \mid a = 0, x_D) \right] \quad (6)$$

Thus, controlling for x_D involves conditioning on x_C and adjusting for $x_{D \setminus C}$.

Example 5.1: Adjusting for an irrelevant variable
This example illustrates the danger of excessive controlling for "pre-treatment" variables, independently of equilibrium considerations. It is adapted from Cinelli et al. (2022), a guide to "good and bad controls" that (following Pearl (2009)) makes use of the formalism of directed acyclic graphs (DAGs).

Suppose that the true causal structure underlying p is given by the DAG

$$a \leftarrow x_1 \rightarrow x_3 \leftarrow x_2 \rightarrow y$$

All variables take values in $\{0, 1\}$; x_1 and x_2 are uniformly distributed; $y = x_2$ and $x_3 = x_1 x_2$ with certainty; and $p(a = x_1 \mid x_1) = 1 - \varepsilon$ for all x_1, where $\varepsilon \approx 0$. The objective causal effect of a on y is null because there is no causal path from a to y. Moreover, $E_p(y \mid a = 1) - E_p(y \mid a = 0)$ is a correct formula for the objective (null) causal effect. In other words, there is no need to control for any of the x variables.

Suppose, however, that one of the DM types has $C = \emptyset$ and $D = \{3\}$ — i.e., he does not condition on any variable, while adjusting for x_3.[3] The type's estimated causal effect is

$$\sum_{x_3} p(x_3)[E_p(y \mid a = 1, x_3) - E_p(y \mid a = 0, x_3)] \tag{7}$$

Under the specification of p, we can calculate that $p(y = 1 \mid a, x_3 = 1) = 1$ for every a, whereas

$$\begin{aligned} p(y &= 1 \mid a = 1, x_3 = 0) \approx 0 \\ p(y &= 1 \mid a = 0, x_3 = 0) \approx \frac{1}{2} \end{aligned}$$

Plugging these values in (7), we obtain a non-null estimated causal effect. The intuition is as follows. Because x_3 is a common consequence of x_1 and x_2 (which are correlated with a and y, respectively), it is not necessarily true that $a \perp y \mid x_3$. Therefore, x_3 is a bad control, and the DM's estimate can be biased. □

The following definition adapts the concept of ε-equilibrium to the present setting (the definition of equilibrium is derived from ε-equilibrium, just as in Section 2).

[3] The absence of a direct link from x_3 into a in the DAG is consistent with no DM type conditioning on x_3 — i.e., this variable does not enter any data type's set C.

Definition 4 *Let $\varepsilon > 0$. A strategy profile $\sigma = (\sigma_1, ..., \sigma_n)$ is an ε-equilibrium if for every $i = 1, ..., n$ and every t, x, a', $\sigma_i(a' \mid t, x) > \varepsilon$ only if*

$$a' \in \arg\max_a \left\{ \sum_{x_{D_i \setminus C_i}} p(x_{D_i \setminus C_i} \mid x_{C_i}) E_p(y \mid a, x_{C_i}) - \theta \cdot \mathbf{1}[a \neq t] \right\}$$

I now extend the notion of vertically ordered types. Define a binary relation P over data types: iPj if $D_i \supseteq C_j$. The meaning of iPj is that data type i controls for every variable that type j conditions on. Since $D_i \supseteq C_i$ for every $i \in N$, P is reflexive. Let P^* be the asymmetric (strict) part of P — i.e., iP^*j if iPj and $j \not P i$. Following Sen (1969), P is *quasitransitive* if P^* is transitive. We say that type i is P^*-undominated in a set of types M, if there is no $j \in M$ such that jP^*i. The following observation is standard.

Remark 1 *Suppose P is complete and quasitransitive. Then, N can be partitioned into L classes, $N_1, ..., N_L$, such that: (i) N_1 consists of all P^*-undominated types in N; and (ii) for every $\ell > 1$, N_ℓ consists of all P^*-undominated types in $N \setminus (\cup_{h<\ell} N_h)$.*

The partition induced by a complete and quasitransitive P is the extended model's analogue of vertical ordering of types. When $C_i = D_i$ for all $i \in N$, it collapses to the original definition, which orders types via set inclusion.

The following results extend the worst-case analysis of Section 3 (homogenous preferences).

Proposition 6 *Let $\gamma = 0$. Suppose P is complete and quasitransitive. Then, the unique equilibrium is for all DM types to play $a = 0$ with probability one. In particular, the DM's expected welfare loss is zero.*

Proposition 7 *Let $\gamma = 0$. Suppose P violates completeness or quasitransitivity. Then, for any $\theta, \beta \in (0, 1)$, there exist λ and $(p(x, y))$ such that $\Pr(a = 1) > \beta$ in some equilibrium. In particular, when $\theta \approx 1$, the equilibrium welfare loss can be arbitrarily close to 1.*

The proof of Proposition 6 is by induction on the partition induced by P. Types in the top layer N_1 effectively control for all sources of correlation between a and y. Even when a top-layer type does not control for some exogenous variable, this does not matter because no other type conditions on this variable, hence it generates no confounding effect. As a result, top-layer types' subjective best-replying implies that they do not generate any variation in choice behavior. This means that types in the next layer N_2 effectively control for all potential confounders — which would not be the case if we did not impose the equilibrium condition on the behavior of top-layer types. This equilibrium effect infects all layers of the partition.

Proposition 7 shows the other side of the "bang-bang" characterization. When P is incomplete or not quasitransitive, the equilibrium requirement does not constrain the maximal possible welfare loss due to bad controls. The proof is constructive, involving more elaborate versions of Example 3.1. In particular, when P is complete but not quasitransitive, the construction involves data types.

Thus, as in Section 3, the distinction between type spaces that are vertically ordered and those that are not is crucial for the worst-case analysis. The contribution of this section is to provide the appropriate extension of the vertical ordering to settings in which the DM may control for variables he does not condition on. Extending this analysis to environments with heterogeneous preferences is an open problem.

6 Conclusion

When DMs draw causal inferences from observed correlations, they may commit errors if they fail to control for an appropriate set of confounding variables. This paper examined a model of this error, when DMs rely on endogenous datasets and may differ in their sets of control variables. Since DMs' causal inferences determine how they condition their actions on their signals, and since this response in turn shapes the very correlations from which DMs draw their inferences, equilibrium analysis is required to evaluate the decision cost of erroneous causal inference due to bad controls.

The general insight that emerged from this analysis was that when DM types are "vertically" differentiated in terms of the sets of their control

variables, the equilibrium cost of bad controls is significantly lower than the non-equilibrium benchmark, and sometimes it completely vanishes. I substantiated the role of vertical differentiation by showing that the upper bound on the welfare loss is significantly higher when types are not vertically ordered. Indeed, in some cases the distinction between equilibrium and non-equilibrium welfare losses completely vanishes.

Of course, worst-case analyses have a built-in limitation: The worse the worst case gets, the less useful it is. This is why the results on vertically ordered type spaces are more meaningful, whereas the role of the complementary results is to put the results on vertically ordered spaces in perspective. Of course, there are economic settings in which we want to assume different typologies of DMs in terms of how they perform causal inference. I hope to explore some of these in future work.

On a speculative note, the results on vertically ordered type spaces suggest that failure to use proper controls, which is a grave error for academic researchers, may not be such a big problem for everyday decision-making, thanks to the corrective equilibrium forces. Could this be one of the reasons this error of causal inference is so ubiquitous in real life?

References

[1] Angrist, J. and J. S. Pischke (2009). Mostly Harmless Econometrics: An Empiricists Guide. Princeton: Princeton University Press.

[2] Clyde, A. (2023). Proxy Variables and Feedback Effects in Decision Making. Mimeo.

[3] De Barreda, I., G. Levy and R. Razin (2022). Persuasion with Correlation Neglect: A Full Manipulation Result, American Economic Review: Insights 4, 123-138.

[4] Cinelli, C., A. Forney and J. Pearl (2020). A Crash Course in Good and Bad Controls, Sociological Methods & Research: 00491241221099552.

[5] Eliaz, K. , R. Spiegler and H. Thysen (2021b). Strategic Interpretations, Journal of Economic Theory 192, Article 105192.

[6] Eliaz, K., R. Spiegler and Y. Weiss (2021a). Cheating with Models, American Economic Review: Insights 3, 417-434.

[7] Galperti, S. (2019). Persuasion: The Art of Changing Worldviews, American Economic Review 109, 996-1031.

[8] Glazer, J. and A. Rubinstein (2012). A Model of Persuasion with Boundedly Rational Agents, Journal of Political Economy 120, 1057–1082.

[9] Jehiel, P. (2005). Analogy-Based Expectation Equilibrium, Journal of Economic theory 123, 81-104.

[10] Esponda. I. and D. Pouzo (2016). Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models, Econometrica 84, 1093-1130.

[11] Hagenbach, J. and F. Koessler (2020). Cheap Talk with Coarse Understanding, Games and Economic Behavior 124, 105-121.

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

[13] Sen, A. (1969). Quasi-transitivity, Rational Choice and Collective Decisions, Review of Economic Studies 36, 381-393.

[14] Schwartzstein, J. and A. Sunderam (2021). Using Models to Persuade, American Economic Review 111, 276-323.

[15] Spiegler, R. (2016). Bayesian Networks and Boundedly Rational Expectations, Quarterly Journal of Economics 131, 1243-1290.

[16] Spiegler, R. (2020). Behavioral Implications of Causal Misperceptions, Annual Review of Economics 12, 81-106.

[17] Spiegler, R. (2022). On the Behavioral Consequences of Reverse Causality, European Economic Review 149: 104258.

Appendix I: Proofs

The proofs are presented out of order, because Propositions 1 and 2 are special cases of Propositions 6 and 7.

Proposition 6

I will show that $a = 0$ with probability one in equilibrium. The proof is by induction with respect to the partition induced by P. Consider an arbitrary type i in the top layer N_1. This type satisfies $D_i \supseteq C_j$ for all $j \in N$. Hence, there is no x variable outside D_i that *any* DM type conditions his action on. Since t is constant, this means that $y \perp a \mid x_{D_i}$ — i.e., $p(y \mid a, x_{D_i}) = p(y \mid x_{D_i})$. Formula (6) then implies that $\Delta_i(x) = 0$. It follows that in equilibrium, type i plays $a = 0$ for all x.

Suppose the claim holds for all types in the top m layers in the partition, and now consider an arbitrary type i in the $(m+1)$-th layer. By definition, $D_i \supseteq C_j$ for every type j outside the top m layers of the partition. As to types in the top m layers, by the inductive step these types play a constant action $a = 0$ in any equilibrium — i.e., there is no variation in their action. It follows that if p is consistent with equilibrium, then $y \perp a \mid x_{D_i}$. Formula (6) then implies $\Delta_i(x) = 0$. It follows that in equilibrium, type i plays $a = 0$ for all x. ■

Proposition 7

Suppose first that P is incomplete. Then, there exist two types, denoted conveniently 1 and 2, such that $C_1 \setminus D_2$ and $C_2 \setminus D_1$ are non-empty. Select two variables in $C_1 \setminus D_2$ and $C_2 \setminus D_1$, and denote them 1 and 2 as well, respectively. Suppose that $\lambda_1 = \lambda_2 = \frac{1}{2}$. Construct p as follows. First, let $x_1, x_2, y \in \{0, 1\}$, and

$$\begin{aligned} p(x_1 &= 1, x_2 = 1) = 1 - \varepsilon \\ p(x_1 &= 0, x_2 = 1) = p(x_1 = 1, x_2 = 0) = \frac{\varepsilon}{2} \end{aligned}$$

where $\varepsilon > 0$ is arbitrarily small. Second, let $p(y = 1 \mid x_1, x_2) = x_1 x_2$. Thus, x_1 and x_2 are the only x variables that determine y, and so we can afford to ignore all other x variables. Given this specification of λ and $p(x, y)$, we can construct an equilibrium in which for each type $i = 1, 2$, $a_i = x_i$

with probability one — exactly as in Example 3.1 — such that $\Pr(a = 1)$ is arbitrarily close to one.

Now suppose that P is complete but not quasitransitive. This means that P^* must have a cycle of length 3 — that is, we can find three types, denoted $1, 2, 3$, such that $1P^*2$, $2P^*3$ and $3P^*1$ — that is, $D_1 \supseteq C_2$, $D_2 \supseteq C_3$ and $D_3 \supseteq C_1$. Since P^* is asymmetric by definition, this means that for each of the three types $i = 1, 2, 3$, there is a distinct variable in $\{1, ..., K\}$, conveniently denoted i as well, such that $1 \in C_1 \setminus D_2$, $2 \in C_2 \setminus D_3$ and $3 \in C_3 \setminus D_1$. Suppose $\lambda_1, \lambda_2, \lambda_3 > 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Let $x_1, x_2, x_3, y \in \{0, 1\}$. Construct p as follows: First,

$$p(x_1 = 1, x_2 = 1, x_3 = 1) = 1 - \varepsilon$$

and

$$p(x_i = 0, x_j = x_k = 1) = \frac{\varepsilon}{3}$$

for every $i = 1, 2, 3$ and $j, k \neq i$, where $\varepsilon > 0$ is arbitrarily small. Second, let $p(y = 1 \mid x_1, x_2, x_3) = x_1 x_2 x_3$. Thus, x_1, x_2, x_3 are the only x variables that determine y, and so we can afford to ignore all other x variables. Suppose each type $i = 1, 2, 3$ plays $a = x_i$ with probability one. Using essentially the same calculation as in the case of incomplete P, we can see that for every $i = 1, 2, 3$, $\Delta_i(x_i = 0) = 0$, whereas $\Delta_i(x_i = 1) \to 1$ as $\varepsilon \to 0$. Therefore, the postulated strategy profile is an equilibrium. ■

Proposition 3

The proof proceeds stepwise. Recall that since P is complete, it is a linear ordering. For convenience, enumerate the types according to P — i.e., $1P2P \cdots Pn$. For every x and every $C \subseteq \{1, ..., K\}$, denote $\gamma(x) = p(t = 1 \mid x)$ and $\gamma(x_C) = p(t = 1 \mid x_C)$.

Step 1: Deriving an expression for $\Delta_i(x)$
Proof: Since $y \perp (a, x) \mid t$, we can write

$$p(y \mid a, x_{C_i}) = \sum_t p(t \mid a, x_{C_i}) p(y \mid a, x_{C_i}, t) = \sum_t p(t \mid a, x_{C_i}) p(y \mid t)$$

Plugging this in (2), we obtain

$$\Delta_i(x) = [p(t = 1 \mid a = 1, x_{C_i}) - p(t = 1 \mid a = 0, x_{C_i})][\delta_1 - \delta_0] \tag{8}$$

We have thus derived an expression for $\Delta_i(x)$. $\square$

Step 2: For every x, $\Delta_1(x) \geq 0$ and $\sigma_1(a = 1 \mid t = 1, x_{C_1}) = 1$.
Proof: For every a, the terms $p(t = 1 \mid a, x_{C_i})$ in (8) can be written as

$$\frac{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i})}{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i}) + (1 - \gamma(x_{C_i}))p(a \mid t = 0, x_{C_i})} \tag{9}$$

Consider the terms $p(a \mid t, x_{C_1})$ in (9). Note that

$$p(a \mid t, x_{C_1}) = \sum_{x_{-C_1}} p(x_{-C_1} \mid t, x_{C_1})p(a \mid t, x_{C_1}, x_{-_{C_1}}) \tag{10}$$

By definition, $C_1 \supset C_j$ for every $j > 1$. This means that no data type j conditions his actions on x_{-C_1}. Therefore, (10) is equal to

$$\sum_{j=1}^{n} \lambda_j \sigma_j(a \mid t, x_{C_j})$$

By the DM's preferences, $\sigma_i(a = 1 \mid t = 1, x_{C_i}) \geq \sigma_i(a = 1 \mid t = 0, x_{C_i})$ in any equilibrium, for every i, x. It follows that $p(a = 1 \mid t = 1, x_{C_1}) \geq p(a = 1 \mid t = 0, x_{C_1})$ for every x_{C_1}. A simple calculation then confirms that the expression (9) is weakly increasing in a for $i = 1$. Since $\delta_1 - \delta_0 \geq 0$, $\Delta_1(x) \geq 0$. $\square$

Step 3: Extending Step 2 to all data types
Proof: The proof is by induction on P. Suppose that for every type $j = 1, ..., m$, $\Delta_j(x) \geq 0$ and $\sigma_j(a = 1 \mid t = 1, x_{C_j}) = 1$. Now consider type $i = m + 1$. We can write

$$p(a \mid t, x_{C_i}) = \sum_{x_{-C_i}} p(x_{-C_i} \mid t, x_{C_i}) \left[\sum_{j \leq m} \lambda_j \sigma_j(a \mid t, x_{C_j}) + \sum_{j > m} \lambda_j \sigma_j(a \mid t, x_{C_j}) \right]$$

By the inductive step,

$$\sigma_j(a = 1 \mid t = 1, x_{C_j}) = 1 \geq \sigma_j(a = 1 \mid t = 0, x_{C_j})$$

for every $j \leq m$. By definition, $C_j \subseteq C_i$ for every $j > m$, hence $\sigma_j(a \mid t, x_{C_j})$ is constant in x_{-C_i}. Therefore,

$$p(a = 1 \mid t = 1, x_{C_i}) = \sum_{j \leq m} \lambda_j \cdot 1 + \sum_{j > m} \lambda_j \sigma_j(a \mid t = 1, x_{C_j})$$

We already observed that

$$\sigma_j(a = 1 \mid t = 1, x_{C_j}) \geq \sigma_j(a = 1 \mid t = 0, x_{C_j})$$

for every x_{C_j}. It follows that

$$\begin{aligned} p(a &= 1 \mid t = 1, x_{C_i}) = \sum_{j \leq m} \lambda_j \cdot 1 + \sum_{j > m} \lambda_j \sigma_j(a \mid t = 1, x_{C_j}) \\ &\geq \sum_{x_{-C_i}} p(x_{-C_i} \mid t, x_{C_i}) \left[\sum_{j \leq m} \lambda_j \sigma_j(a \mid t = 0, x_{C_j}) + \sum_{j > m} \lambda_j \sigma_j(a \mid t = 0, x_{C_j}) \right] \\ &= p(a = 1 \mid t = 0, x_{C_i}) \end{aligned}$$

As in the proof of Step 2, applying this inequality to (9) implies that $\Delta_i(x) \geq 0$ and $\sigma_i(a = 1 \mid t = 1, x_{C_i}) = 1$. This completes the inductive proof. □

Interlude: Step 3 and Simpson's paradox

Before turning to the next step in the proof, it may be helpful to pause and discuss the significance of the proof of Step 3. In both Steps 2 and 3, the key to proving that the DM's estimated causal effect of a on y is always non-negative is showing that $p(a = 1 \mid t = 1, x_{C_i}) \geq p(a = 1 \mid t = 0, x_{C_i})$ for every x_{C_i} — i.e., that the DM's average behavior conditional on x_{C_i} is increasing in t, for every x, i. In general, this need not be the case, despite the fact that $p(a = 1 \mid t, x) = \sum_i \lambda_i \sigma_i(a = 1 \mid t = 0, x_{C_i})$ *is* increasing in t for every x. The reason is that $p(a \mid t, x_{C_i})$ marginalizes $p(a = 1 \mid t, x)$ over x_{-C_i}. The observation that monotonicity of conditional probabilities is not always preserved under marginalization is known as *Simpson's paradox* (see Pearl (2009)). The challenge of the proof of Steps 2 and 3 is to ensure that Simpson's paradox is moot in the present context.

Step 4: An upper bound on the expected equilibrium welfare loss given x

Proof: We have established that in any equilibrium, all data types play $a = 1$ with probability one when $t = 1$. Therefore, they only commit an error

if they play $a = 1$ with positive probability when $t = 0$. Fix the realization of x. Let $i(x)$ be the lowest-indexed type j for which $\sigma_j(a = 1 \mid t = 0, x_{C_j}) > 0$. Then, the DM's expected welfare loss given x is

$$\theta(1 - \gamma(x)) \sum_{j=i(x)}^{n} \lambda_j \sigma_j(a = 1 \mid t = 0, x_{C_j})$$

In order for type $i(x)$ to play $a = 1$ given x and $t = 0$, it must be the case that $\theta \leq \Delta_{i(x)}(x)$. By Step 3, $\sigma_j(a = 1 \mid t = 1, x_{C_j}) = 1$ for all j, hence $p(a = 1 \mid t = 1, x_{C_{i(x)}}) = 1$. Plugging this identity into (8)-(9) and recalling that $0 \leq \delta_1 - \delta_0 \leq 1$, we obtain

$$\Delta_{i(x)}(x) \leq \frac{\gamma(x_{C_{i(x)}})}{\gamma(x_{C_{i(x)}}) + (1 - \gamma(x_{C_{i(x)}}))p(a = 1 \mid t = 0, x_{C_{i(x)}})}$$

Since $C_j \subseteq C_i$ for every j for which $\sigma_j(a = 1 \mid t = 0, x_{C_j}) > 0$, it follows that none of these types j condition on $x_{-C_{i(x)}}$. Therefore,

$$p(a = 1 \mid t = 0, x_{C_{i(x)}}) = \sum_{j=i(x)}^{n} \lambda_j \sigma_j(a = 1 \mid t = 0, x_{C_j})$$

Denote this quantity by α. This means that the DM's expected welfare loss given x is at most

$$\frac{\gamma(x_{C_{i(x)}})}{\gamma(x_{C_{i(x)}}) + (1 - \gamma(x_{C_{i(x)}}))\alpha} \cdot (1 - \gamma(x)) \cdot \alpha$$

This expression attains its maximal value when $\alpha = 1$. Therefore, the following expression

$$(1 - \gamma(x))\gamma(x_{C_{i(x)}}) = (1 - \gamma(x)) \cdot \sum_{x'} p(x' \mid x'_{C_{i(x)}} = x_{C_{i(x)}})\gamma(x')$$

is an upper bound on the DM's expected welfare loss given x. $\square$

Step 5: Deriving the upper bound on the DM's ex-ante expected equilibrium welfare loss

Proof: By Step 4, the ex-ante welfare loss is at most

$$\sum_{x} p(x)(1-\gamma(x)) \cdot \sum_{x'} \beta(x',x)\gamma(x') \tag{11}$$

where $\beta(x',x) = p(x' \mid x'_{C_{i(x)}} = x_{C_{i(x)}})$. The coefficients $\beta(\cdot)$ constitute a system of convex combinations. Expression (11) is a concave function of $(\gamma(x))_x$. By Jensen's inequality, it attains a maximum when $\gamma(x) = \gamma$ for all x, such that the upper bound on the DM's expected equilibrium welfare loss is $\gamma(1-\gamma)$. ■

Proposition 4

(i) Deriving the upper bound

Let $\gamma \geq \frac{1}{2}$, without loss of generality, such that $\max\{\gamma, 1-\gamma\} = \gamma$. Suppose there is an equilibrium in which the DM's expected welfare loss exceeds γ. To reach a contradiction, the proof proceeds stepwise.

Step 1: Deriving a necessary condition

Proof: If the expected equilibrium welfare loss exceeds γ, then $p(a = 1 \mid t = 0) > 0$. Thus, there exist x and i such that $\sigma_i(a = 1 \mid t = 0, x) > 0$. Denote

$$X_i^* = \{x \mid \sigma_i(a = 1 \mid t = 0, x) > 0\}$$

Define

$$B_t(x,i) = \begin{cases} \sum_{x' \mid x'_{C_i} = x_{C_i}} p(x' \mid t) p(a = 1 \mid t, x') & \text{if} \quad X_i^* \neq \emptyset \\ 0 & \text{if} \quad X_i^* = \emptyset \end{cases}$$

Note that whether $x \in X_i^*$ only depend on x_{C_i}. Likewise, $B_t(x,i)$ is effectively a function of x_{C_i}.

By the equilibrium condition, every $x \in X_i^*$ must satisfy

$$\begin{aligned} p(t &= 1 \mid a = 1, x_{C_i}) - p(t = 1 \mid a = 0, x_{C_i}) \geq p(t = 1 \mid a = 1, x_{C_i}) \\ &= \frac{\gamma B_1(x,i)}{\gamma B_1(x,i) + (1-\gamma) B_0(x,i)} \geq \theta \end{aligned}$$

which can be written equivalently as

$$B_0(x,i) \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} B_1(x,i) \tag{12}$$

Summing $B_t(x,i)$ over x_{C_i} yields

$$\bar{B}_t(i) = \sum_{x \in X_i^*} p(x \mid t) p(a = 1 \mid t, x) \tag{13}$$

Performing this summation over x_{C_i} on both sides of (12) implies

$$\bar{B}_0(i) \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} \bar{B}_1(i)$$

for every i for which $X_i^* \neq \emptyset$. (Note that $\bar{B}_t(i) = 0$ when $X_i^* = \emptyset$.) It follows that a necessary condition for the welfare loss to exceed γ is

$$\max_i \bar{B}_0(i) \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} \max_i \bar{B}_1(i) \tag{14}$$

Note that

$$p(a = 1 \mid t, x) = \sum_{j=1}^{n} \lambda_j \sigma_j(a = 1 \mid t, x)$$

Using this observation and (13), we can reformulate (14) as follows. Every x is assigned a subset of types $M(x) = \{i \mid x \in X_i^*\}$. The joint distribution p over (t, x) and the strategy profile σ induce a distribution μ over M, such that

$$\mu(M) = p(\{i \mid x \in X_i^*\} = M \mid t = 0)$$

Denote

$$\lambda_j^* = \lambda_j \sum_x p(x \mid t = 0, x \in X_j^*) \sigma_j(a = 1 \mid t = 0, x)$$

Then, (14) can be rewritten as

$$\max_i \sum_{M \mid i \in M} \mu(M) \sum_{j \in M} \lambda_j^* \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} \max_i \bar{B}_1(i) \tag{15}$$

This inequality is a necessary condition for the equilibrium welfare loss to exceed γ. $\square$

Step 2: The following inequality holds:

$$\max_i \sum_{M \mid i \in M} \mu(M) \sum_{j \in M} \lambda_j^* \geq \left(\sum_M \mu(M) \sum_{j \in M} \lambda_j^* \right)^2 \tag{16}$$

Proof:[4] If we prove that

$$\sum_{M|i\in M}\mu(M)\sum_{j\in M}\frac{\lambda_j^*}{\sum_k\lambda_k^*}\geq\left(\sum_M\mu(M)\sum_{j\in M}\frac{\lambda_j^*}{\sum_k\lambda_k^*}\right)^2$$

then this will immediately imply (16) because $\sum_k\lambda_k^*\leq 1$. Therefore, we can assume without loss of generality that $\sum_j\lambda_j^*=1$. Moreover, I will prove a more demanding inequality:

$$\sum_i\lambda_i^*\sum_{M|i\in M}\mu(M)\sum_{j\in M}\lambda_j^*\geq\left(\sum_M\mu(M)\sum_{j\in M}\lambda_j^*\right)^2 \tag{17}$$

The L.H.S of this inequality can be written equivalently as

$$\sum_M\mu(M)\sum_{i\in M}\lambda_i^*\sum_{j\in M}\lambda_j^*=\sum_M\mu(M)\left(\sum_{j\in M}\lambda_j^*\right)^2$$

Denote

$$z(M)=\sum_{j\in M}\lambda_j^*$$

We can regard $z(M)$ as a real-valued random variable whose distribution is determined by the distribution μ. The expression

$$\sum_M\mu(M)\left(z(M)\right)^2-\left(\sum_M\mu(M)z(M)\right)^2$$

is the variance of this random variable, which is non-negative by definition. This proves (17), and consequently the result. □

Step 3: Reaching a contradiction
Denote

$$\beta=\max_i\bar{B}_1(i)$$

By the definition of $\bar{B}_1$ given by (13), β is a lower bound on $\Pr(a=1\mid t=1)$. Therefore,

$$\Pr(t=1,a=0)\leq\gamma-\gamma\beta$$

[4]This proof is due to Omer Tamuz.

Furthermore, $\Pr(a = 1 \mid t = 0)$ is by definition

$$\sum_x \Pr(x \mid t = 0) \Pr(a = 1 \mid t = 0, x) = \sum_M \mu(M) \sum_{j \in M} \lambda_j^*$$

Applying Step 2, the DM's expected equilibrium welfare loss is bounded from above by

$$\theta \cdot \left[\gamma - \gamma\beta + (1 - \gamma) \sqrt{\frac{\gamma(1-\theta)\beta}{\theta(1-\gamma)}} \right]$$

which by assumption exceeds γ. Rewriting this inequality as

$$\theta \cdot \left[\gamma - \gamma\beta + \sqrt{\frac{\gamma(1-\gamma)(1-\theta)\beta}{\theta}} \right] - \gamma > 0$$

and regarding it as a quadratic function of $\sqrt{\beta}$, we can check that this inequality has no solution whenever $\gamma > \frac{1}{5}$, a contradiction. ■

(ii) Implementing the upper bound

Since P is incomplete, $K \geq 2$. Moreover, there exist two data types, 1 and 2, and two exogenous variables, conveniently denoted x_1 and x_2, such that $1 \in C_1 \setminus C_2$ and $2 \in C_2 \setminus C_1$. Suppose $\lambda_1 + \lambda_2 = 1$. Without loss of generality, let $\gamma \geq \frac{1}{2}$, such that $\max\{\gamma, 1 - \gamma\} = \gamma$. Suppose that $x_1, x_2 \in \{0, 1, \#\}$. Construct the following distribution over triples (t, x_1, x_2):

Pr	t	x_1	x_2
β	1	1	1
β^2	0	1	0
β^2	0	0	1
$1 - \gamma - 2\beta^2$	0	#	#
$\gamma - \beta$	1	0	0

Suppose that p is constant over the other x variables, such that they can be ignored. Complete the exogenous components of p by letting $\delta_1 = 1$ and $\delta_0 = 0$. Since there are no relevant x variables other than x_1 and x_2, we can set without loss of generality $C_1 = \{1\}$ and $C_2 = \{2\}$.

Let each type i play $a_i = x_i$ with probability one whenever $x_i \in \{0, 1\}$.[5]

[5]This involves some imprecision: The definition of ε-equilibrium requires the DM's strategy to be fully mixed. I chose to include no perturbation when $x_i = 0, 1$ in order to

In addition, suppose each type i plays $a = 0$ with probability $1 - \varepsilon$ when $x_i = \#$, where ε and β are arbitrarily small. Let us calculate the terms in $\Delta_1(x_1 = 1)$:

$$\begin{aligned} p(t &= 1 \mid a = 1, x_1 = 1) = \frac{\beta}{\beta + \lambda_1 \beta^2} \approx 1 \\ p(t &= 1 \mid a = 0, x_1 = 1) = 0 \end{aligned}$$

such that $\Delta_1(x_1 = 1) \approx 1$. Let us now calculate the terms in $\Delta_1(x_1 = 0)$:

$$\begin{aligned} p(t &= 1 \mid a = 1, x_1 = 0) = 0 \\ p(t &= 1 \mid a = 0, x_1 = 0) = \frac{\gamma - \beta}{\gamma - \beta + \lambda_1 \beta^2} \approx 1 \end{aligned}$$

such that $\Delta_1(x_1 = 0) \approx -1$. It follows that $\Delta_1(x_1 = 1) > \theta$ and $\Delta_1(x_1 = 0) < -\theta$, such that type 1 strictly prefers to play $a_i = x_i$ for all $x_i \in \{0, 1\}$. This is consistent with the postulated strategy.

Finally, note that $p(t = 1 \mid a, x_1 = \#) = 0$ for both $a = 0, 1$, hence $\Delta_1(x_1 = \#) = 0$. It is therefore optimal for type 1 to play $a = 0$ when $x_1 = \#$. Since he follows this prescription with probability $1 - \varepsilon$, this completes the confirmation that type 1's behavior is consistent with ε-equilibrium. By symmetry, the same calculation holds for type 2. We have thus constructed an ε-equilibrium in which the DM commits an error with probability arbitrarily close to γ. Since θ can be arbitrarily close to 1, this completes the proof. ■

Proposition 5

Since P is incomplete, $K \geq 2$. Moreover, there exist two data types, 1 and 2, and two exogenous variables, conveniently denoted x_1 and x_2, such that $1 \in C_1 \setminus C_2$ and $2 \in C_2 \setminus C_1$. Let $\lambda_1 = \lambda_2 = 0.5$. Construct a distribution p over t, x_1, x_2, y given by the following table (suppose that p is constant over the other x variables, such that they can be ignored), where $\beta > 0$ is

clarify the role of trembles when $x_i = \#$. This imprecision can be fixed by introducing trembles on the order of ε^2 when $x_i = 0, 1$.

arbitrarily small:

$p(t, x_1, x_2, y)$	t	x_1	x_2	y
$1-\gamma-\beta$	0	1	1	1
$\gamma-\beta$	1	0	0	1
β	0	1	0	0
β	1	0	1	0

Suppose data type i plays $a_i \equiv x_i$. Let us calculate $\Delta_1(x_1)$ for each x_1. First,

$$\begin{aligned} p(y &= 1 \mid a = 1, x_1 = 1) = \frac{1-\gamma-\beta}{1-\gamma-\beta+\beta \cdot 0.5} \approx 1 \\ p(y &= 1 \mid a = 0, x_1 = 1) = 0 \end{aligned}$$

where the second equation holds because the combination of $a = 0$ and $x_1 = 1$ occurs only when $x_2 = 0$, in which case $y = 0$ with certainty.

Second,

$$\begin{aligned} p(y &= 1 \mid a = 0, x_1 = 0) = \frac{\gamma-\beta}{\gamma-\beta+\beta \cdot 0.5} \\ p(y &= 1 \mid a = 1, x_1 = 0) = 0 \end{aligned}$$

where the second equation holds because the combination of $a = 1$ and $x_1 = 0$ occurs only when $x_2 = 1$, in which case $y = 0$ with certainty.

Plugging these terms into the definition of $\Delta_1(x_1)$ yields $\Delta_1(x_1 = 1) \approx 1$ and $\Delta_1(x_1 = 0) \approx -1$. The calculation for type 2 is identical due to symmetry. Therefore, for every $\theta < 1$, we can set β such that each data type i will indeed prefer to play $a \equiv x_i$. Furthermore, for both types i, $x_i = 1 - t_i$ with probability arbitrarily close to one. Therefore, the DM plays $a = 1 - t$ with arbitrarily high probability, such that the expected welfare loss is arbitrarily close to one. ■

Appendix II: Consequential Actions

Throughout the paper, we focused on the extreme case in which the DM's action has no causal effect on the outcome. This facilitated the definition of the DM's equilibrium welfare loss due to poor controls, relative to the rational-expectations benchmark. This appendix extends the analysis to

situations in which actions do influence outcome. I build on the extended notion of types presented in Section 5.

Define a variable z that takes values in $[0,1]$, This variable is a consequence of (t,x), *independently* of a — just as y was in the baseline model. The outcome y is purely caused by a and z, such that

$$E_p(y \mid a,z) = \beta a + (1-\beta)z$$

where $\beta \in (0,1)$ quantifies the true causal effect of a on y.

This formulation implies that for every type $i \in N$, the perceived z-outcome of actions is

$$\sum_{x_{D_i \setminus C_i}} p(x_{D_i \setminus C_i} \mid x_{C_i}) E_p(z \mid a, x_{D_i})$$

The type's estimated causal effect of switching from $a=0$ to $a=1$ on z given x is

$$\Delta_i^z(x) = \sum_{x_{D_i \setminus C_i}} p(x_{D_i \setminus C_i} \mid x_{C_i}) \left[E_p(z \mid a=1, x_{D_i}) - E_p(z \mid a=0, x_{D_i})\right]$$

Therefore, the type's estimated causal effect of switching from $a=0$ to $a=1$ on y given x is $\beta + (1-\beta)\Delta_i^z(x)$. Since $z \perp a \mid (t,x)$, the equilibrium analysis of $\Delta_i^z(x)$ and how it relates to the DM's strategy is the same as the analysis of $\Delta_i(x)$ in the baseline model.

It follows that the only thing that needs adjustment is the definition of the DM's welfare loss. The optimal rational-expectations action maximizes $\beta a - \theta \cdot \mathbf{1}[a \neq t]$, because a has no causal effect on z, such that the only effect of a on y is via the direct channel parameterized by β. Therefore, the expected welfare loss given a joint distribution p is

$$\gamma \cdot p(a=0 \mid t=1) \cdot (\theta+\beta) + (1-\gamma) \cdot p(a=1 \mid t=0) \cdot (\theta-\beta) \qquad (18)$$

The DM chooses $a=0$ at $(t=1,x)$ only if $\theta + \beta \leq -(1-\beta)\Delta_i^z(x)$. Likewise, he chooses $a=1$ at $(t=0,x)$ only if $\theta - \beta \leq (1-\beta)\Delta_i^z(x)$. Consequently, by (18), the upper bounds on the DM's equilibrium welfare loss are the same as in Sections 3-4, multiplied by $1-\beta$.

Appendix III: Other Modeling Frameworks

The model of behavioral causal inference presented in this paper poses a new question. However, it can be formulated by adapting existing modeling frameworks. To make the comparison complete, I make use of the extended formalism of Section 5.

Subjective state spaces

The perceived causal effect given by (6) can be interpreted traditionally in terms of the Savage framework, where the state space itself is *subjective*. According to this interpretation, X_{D_i} is type i's subjective state space and X_{C_i} is his set of signals. The novelty here is that while the state space is subjective, the DM's belief is a projection of the *objective* distribution p on his subjective state space. Moreover, unlike the standard Savage model, the stochastic mapping from the DM's subjective states to outcomes is affected by the DM's strategy, hence it is an endogenous object. These deviations from the Savage framework are so drastic that they justify avoiding the Savage terminology in the paper's formal exposition.

Analogy-based expectations

Jehiel's (2005) concept of analogy-based expectations equilibrium captures the idea that players' perception of other players' strategies is coarse. In the present context, we can regard y as the action taken by a fictitious opponent of the DM after observing the history $(a, t, x_1, ..., x_n)$. In this context, x_{C_i} is type i's information set, whereas D_i determines his "analogy partition". Two histories belong to the same partition cell if they share the same value of x_{D_i}. My definition of equilibrium is consistent with Jehiel's assumption that type i believes that the fictitious player's strategy is measurable with respect to type i's analogy partition, and that the equilibrium belief is consistent with the average objective behavior of y conditional on each partition cell.[6]

Bayesian networks

The model can be cast in the Bayesian-network language of Spiegler (2016). When a has no causal effect on y, the objective distribution p is consistent

[6] A minor difference is that I use trembles to handle conditioning on null events, whereas Jehiel (2005) relies on the sequential-equilibrium conceptual baggage.

with the following DAG:

$$\begin{array}{ccccc} a & \leftarrow & t & \rightarrow & y \\ & \nwarrow & \uparrow & \nearrow & \\ & & x & & \end{array}$$

Using the DAG language, the distinction between data types in the present model can be redefined in terms of subjective causal models. Specifically, type i's causal model is

$$\begin{array}{ccc} x_{D_i \setminus C_i} & \longrightarrow & y \\ \uparrow & \nearrow & \uparrow \\ x_{C_i} & \longrightarrow & a \end{array}$$

According to Spiegler (2016), the belief generated by this subjective model obeys the Bayesian-network factorization formula

$$p(x_{C_i})p(x_{D_i \setminus C_i} \mid x_{C_i})p(a \mid x_{C_i})p(y \mid a, x_{C_i}, x_{D_i})$$

The DM's perceived causal effect of a on y given x_{C_i} is thus given by (6). Equilibrium in the present model is consistent with the notion of personal equilibrium in Spiegler (2016), with the modification that the DM's subjective causal model itself is random.

Previous applications of the Bayesian-network framework contain precedents for two of this paper's ingredients. Eliaz et al. (2021a) characterize the worst-case distortion of pairwise correlations generated by misspecified Gaussian Bayesian networks. Spiegler (2022) illustrates how equilibrium effects can ameliorate the cost of a reverse-causality error.

Berk-Nash equilibrium

The Bayesian-network framework in Spiegler (2016) can be subsumed into the more general concept of Berk-Nash equilibrium (Esponda and Pouzo (2016)). According to this concept, the DM best-replies to a conditional belief (over outcomes given actions and signals), which minimizes a weighted version of Kullback-Leibler divergence with respect to the objective conditional distribution. Proper adaptation of this concept to the present context requires the weights to be given by the DM's *ex-ante* equilibrium strategy.