Behavioral Causal Inference[*]

Ran Spiegler[†]

March 8, 2023

Abstract

When inferring the causal effect of one variable on another from correlational data, a common practice by professional researchers and lay decision makers alike is to control for some set of exogenous confounding variables. Choosing an inappropriate set of control variables can lead to erroneous causal inferences. This paper presents a model of lay decision makers who use long-run observational data to learn the causal effect of their actions on a payoff-relevant outcome. Different types of decision makers use different sets of control variables. I obtain upper bounds on the equilibrium welfare loss due to wrong causal inferences, for various families of data-generating processes. The bounds depend on the structure of the type space. When types are "ordered" in a certain sense, the equilibrium condition greatly reduces the cost of wrong causal inference due to poor controls.

[*] Financial support by ISF grant no. 320/21 is gratefully acknowledged.

[†] Tel Aviv University and University College London

1 Introduction

Learning the causal effect of one variable on another 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 in order to evaluate the consequences of their actions. They obtain data about observed correlations among variables (via first-hand experience or through the media) and try to extract lessons from the data concerning the consequences of their actions. For example, which college degree will improve their long-run economic prospects? Will wearing surgical masks on airplanes their chances of catching a virus? Will drinking more coffee cause health problems?

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 any pushback when they employ these methods inappropriately. The second difference is that 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 their inferences. Thus, it is 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 binary choice between two actions, denoted 0 and 1. The DM's choice is based on his belief regarding the action's causal effect on a binary payoff-relevant outcome (which also takes the values 0 or 1). 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 deciding whether to choose a STEM major tries to evaluate the labor-market outcomes of STEM and non-STEM graduates, it is natural for him to focus on people who share his highschool math background. In both cases, when the agent consults data to estimate the consequences of various actions, he tries to focus on data points that share his own characteristics — if he has access to such fine-grained data. This type of controlling consists of *conditioning* on the realization of some exogenous variables.

Another type of controlling involves *adjustment*. For example, in the above-mentioned surgical-mask example, the agent may have access to data about the prevalence of certain genes and their correlation with viral infection. If he does not know his own relevant genetic background, he can nevertheless adjust his beliefs according to the available data.

In general, suppose that long-run correlational data is given by some joint probability distribution p over actions a, outcomes y, and a collection of exogenous variables is $x_1, ..., x_K$. The DM is able to control for the variables indexed by $D \subseteq \{1, ..., K\}$ and condition on a subset $C \subseteq D$ (such that he only adjusts for the variables in $D \setminus C$). The DM's estimated causal effect of a on y is given by the formula

$$\sum_{x_D} p(x_D \mid x_C) \left[p(y = 1 \mid a = 1, x_D) - p(y = 1 \mid a = 0, x_D) \right] \tag{1}$$

When the set D of control variables differs from the set that a normative outside observer 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.

Erroneous causal inference due to "bad (exogenous) controls" may take various forms, which are easy to illustrate with directed acyclic graphs (DAGs), following Pearl (2009). For instance, suppose that in reality, a has no causal effect on y and that every observed correlation between these variables is due to confounding by the exogenous variable x. These objective causal relations are represented by the DAG $a \leftarrow x \rightarrow y$. Given the observed joint distribution p over a, x, y, the proper measurement of the average causal effect of a on y is given by the formula

$$\sum_{x} p(x)[p(y=1 \mid a=1, x) - p(y=1 \mid a=0, x)]$$

This formula will correctly yield a zero causal effect. If, however, the DM fails to control for x, he will regard $p(y=1 \mid a=1) - p(y=1 \mid a=0)$ as the causal effect of a on y — in other words, he will mistake correlation for causation — and potentially measure an erroneous, non-zero effect.

Bad controls can also involve *excessive* controlling for exogenous variables. The following example is taken from Cinelli et al. (2022). The true causal model is given by the DAG $a \leftarrow x_1 \rightarrow x_2 \leftarrow x_3 \rightarrow y$. Thus, as in the previous example, the objective causal effect of a on y is null. However, in this case the quantity $p(y=1 \mid a=1) - p(y=1 \mid a=0)$ is the 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 the DM adjusts for x_2. Then, his estimated causal effect will be

$$\sum_{x_2} p(x_2)[p(y=1 \mid a=1, x_2) - p(y=1 \mid a=0, x_2)]$$

In this case, the variable x_2 is a bad control, and the DM's estimate can end up being non-null.

This paper poses 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 of actions given his information maximizes his subjective expected payoff with respect to the belief he extracts from p using his causal-inference procedure?

I study this question with a simple model, in which a DM chooses an action $a \in \{0,1\}$ after a collection of exogenous variables $t, x_1, ..., x_K$ is realized, where $t \in \{0,1\}$ is the DM's preference type. The DM's vNM utility function is $u(a,t,y) = y - c \cdot \mathbf{1}[a \neq t]$. Thus, the DM will only choose $a \neq t$ if he thinks that a has a causal effect on y. In the baseline model, however, I assume that the objective causal effect of a on y is null: y is determined only by the exogenous variables according to some conditional probability distribution (I relax this assumption in Section 5).

The DM's control variables are given by a "data type" (drawn randomly, independently of his preference type), which is defined by a distinct pair of subsets (D, C), where: $C \subseteq D \subseteq \{1, ..., K\}$; C represents the variables the type conditions on, and $D \setminus C$ represents the variables he only adjusts for, leading to an estimated causal effect of a on y (given x) as described by (1). The formula is evaluated according to a joint distribution over all variable. The DM observes the realization of t, but he has no long-run data about t and therefore does not use it for causal estimates. In equilibrium, the distribution of a conditional on the exogenous variables is consistent with each DM type best-replying to his causal belief. (Section 6 explains the relation between this equilibrium concepts and earlier modeling frameworks by Jehiel (2005), Spiegler (2016,2020) and Esponda and Pouzo (2016).)

The basic insight of this paper is that this equilibrium condition can restrict the magnitude of the DM's welfare loss due to errors of causal inference. These errors consist of misreading the causal component of observed correlational patterns. When agents act on these errors, they change these very patterns, and hence the causal effects they deduce from them.

The above pair of examples of "bad controls" provide an extreme illustration of this insight. Suppose that $t = 0$ with certainty — i.e., there are no preference shocks. In the first example, x is a direct cause of a. For this relation to be non-null, it must be the case that some DM data types condition their action on x. However, this means that these types correctly measure the null objective causal effect of a on y. Since $t = 0$ for sure, these types will play $a = 0$ with certainty. By definition, the same lack of variation of a with

x extends to the types who cannot condition their action on x. It follows that no DM type will vary his action with x, which destroys the confounding effect of x, and therefore any causal error due to failure to control for x.

The same reasoning applies to the second example. If a DM data type conditions on x_1, his causal inference is sound and therefore his action is constant (since there are no preference shocks); whereas if he does not condition on x_1, his behavior is independent of x_1 by definition. Thus, no type varies his behavior with x_1, such that the link $a \leftarrow x_1$ that makes x_2 a "bad control" is effectively severed.

The main results in this paper — presented in Section 4 — 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 rational-expectations strategy $a \equiv t$). It turns out that a simple binary relation over the data types is critical for this upper bound. Say that one type (C, D) dominates another (C', D') if $D \supseteq C'$ (i.e., the former type's set of control variables contains the latter type's set of conditioning variables). When t is constant, the upper bound is 0 when the domination relation is complete and quasitransitive, and 1 when it is not.[1] Thus, when data types are ordered in some sense, the equilibrium condition eliminates all welfare loss due to causal errors. Conversely, when data types are not ordered, the upper bound on the DM's welfare loss is the same as when we do not impose any restriction on the conditional action distribution.

I obtain partial characterization results when there is variation in t and $D = C$ for all data types. When the domination relation is complete and the y is purely a function of t, the upper bound on the DM's equilibrium welfare loss is $\Pr(t = 1) \cdot \Pr(t = 0)$. When the relation is incomplete, the upper bound must be higher. When in addition there is no restriction on the conditional outcome distribution, the upper bound is 1. Once again, the structure of the domination relation over data types plays a key role in how equilibrium forces constrain the cost of flawed causal inference.[2]

[1] A binary relation is quasitransitive if its asymmetric part is transitive (following Sen (1969)).

[2] Spiegler (2022) presents an example of how equilibrium forces can restrict the cost of committing a reverse-causality misperception.

2 A Model

Let a, t and y be three binary variables that take values in $\{0, 1\}$, where: a is an *action* that a decision maker (DM) chooses; y is an *outcome*; and t is the DM's *preference type*. Let $x = (x_1, ..., x_K)$ be a collection of additional exogenous variables, which are realized jointly with t, and prior to the realization of a and y. Let $A = \{0, 1\}$ denote the set of values that a can take. Let X_k be the set of values that the variable x_k can take. For every $M \subset \{1, ..., K\}$, 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 sole 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 Section 5.

The DM is a subjective expected utility maximizer, whose vNM utility function is

$$u(t, a, y) = y - c \cdot \mathbf{1}[a \neq t]$$

where $c \in (0, 1)$ 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 that this increases the probability of the outcome $y = 1$. If the DM understood that a has no causal effect on y, he would always choose $a = t$.

The DM's *data type* is defined by a pair (C, D), where $C \subseteq D \subseteq \{1, ..., K\}$. The interpretation is that C defines the set of x variables that the type can condition on, because he observes their realization before taking an action; and D is the set of exogenous variables about which he has long-run data (note that t is never among these variables). There are n data types. Denote $N = \{1, ..., n\}$. Each data type $i \in N$ is associated with a distinct pair (C_i, D_i). We say that type i is *simple* if $C_i = D_i$ — i.e., the DM only has long-run data about the variables he conditions on. Let $\lambda \in \Delta(N)$ be a prior distribution over data types. This distribution is independent of all other variables. A strategy for type (t, i) is a function $\sigma_{t,i} : X \rightarrow \Delta(A)$. By definition, this strategy is measurable with respect to X_{C_i}.

Let p be a joint long-run probability distribution over t, x, a, y. Denote $\gamma = p(t = 1)$. The assumption that a has no causal effect on y means that p satisfies the conditional-independence property $y \perp a \mid (t,x)$.[3] The distribution p can thus be factorized 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:

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

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

I assume that a DM of data type i forms the following belief regarding the causal effect of a on y given his observation of x_{C_i}:

$$\tilde{p}_i(y \mid x_{C_i}, do(a)) = \sum_{x_{D_i \backslash C_i}} p(x_{D_i} \mid x_{C_i}) p(y \mid a, x_{D_i}) \qquad (2)$$

The $\tilde{p}$ notation indicates that this is a subjective belief, which may be incorrect. The *do* notation follows Pearl (2009). Its role here is merely to indicate that (2) is a causal quantity, to be distinguished from purely probabilistic conditioning. The DM's attempt to evaluate the causal effect of a on y impels him to *control* for every exogenous variable about which he has data. For some of these variables (represented by C_i), he also learns their realization prior to taking his action, and therefore he *conditions* on them. For the other variables (represented by $D_i \backslash C_i$), he has data about their long-run correlation with a, x_{C_i} and y, yet he does not learn their realization prior to taking an action, and therefore he *adjusts* his belief by summing over them.

Definition 1 *Data type i's perceived causal effect of switching from $a = 0$ to $a = 1$ given x is*

$$\Delta_i(x) = \tilde{p}_i(y = 1 \mid x_{C_i}, do(a = 1)) - \tilde{p}_i(y = 1 \mid x_{C_i}, do(a = 0))$$

Plugging (2) into this definition, we obtain:

[3] Throughout the paper, I use the symbol $\perp$ to denote statistical independence.

$$\Delta_i(x) = \sum_{x_{D_i \setminus C_i}} p(x_{D_i \setminus C_i} \mid x_{C_i})[p(y = 1 \mid a = 1, x_{D_i}) - p(y = 1 \mid a = 0, x_{D_i})] \tag{3}$$

This formula will serve us throughout this paper.

If the DM had long-run data about all exogenous variables (including t), then he could control for all of them. Doing so, he would correctly infer a null causal effect of a on y. In contrast, the DM in this model may end up believing that a has a non-zero causal effect on y because he fails to control for all the 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.

The preceding paragraph may give the impression that the only case of "bad controls" that the model captures is *insufficient* controls. However, note that while controlling for all $K + 1$ exogenous variables is always correct, it is possible that a strict subset of these variables is a sufficient set of controls. In this case, controlling for additional variables may induce errors, as in the example by Cinelli et al. (2022) described in the Introduction. Thus, the present model allows for both insufficient and excessive controlling. However, the model does not accommodate variables that are caused by a or y as possible controls — it only focuses on so-called "pre-treatment" variables.

Definition 2 *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*

$$a' \in \arg\max_a \sum \tilde{p}_i(y \mid x_{C_i}, do(a))u(t, a, y)$$

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

The structure of u means that in equilibrium, type i will play $a \neq t$ with positive probability at x only if

$$|\Delta_i(x)| \geq c$$

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*

$$c \sum_{t,x} p(t,x) \sum_{i \in N} \lambda_i \sigma_i(a = 1 - t \mid t, x)$$

Our 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 on the DM's expected welfare loss is 1. To see why, suppose that $t = 0$ with certainty, and that $x \in \{0, 1\}$. Assume $y = x$ with probability one for every x, and consider the strategy σ that prescribes $a = x$ with probability one. Then, the expected welfare loss is

$$c \cdot [p(x = 1) \cdot 1 + p(x = 0) \cdot 0]$$

Then, if we set c and $p(x = 1)$ to be arbitrarily close to one, the expected welfare loss is also arbitrarily close to one. However, the strategy σ is inconsistent with our equilibrium definition. For the DM to vary a with x, he must be able to *condition* on x — i.e., $C_i \neq \emptyset$. But this means the DM correctly *controls* for x when estimating the causal effect of a on y, which means that he correctly estimates it to be zero, contradicting the assumption that he plays $a = 1$ for some realization of x. It follows that the requirement that σ is an equilibrium strategy can have bite.

Comment. The assumption that $C \subseteq D$ means that if a DM conditions on a variable, he must have long-run data about it. In principle, one can easily imagine situations in which agents know the realization of a variable without having data about the long-run behavior of this variable. For instance, the DM may know his height but lack access the statistics about how height is correlated with the outcome of interest. 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

$C \subseteq D$. Note that the DM knows the realization of t, and he makes use of this information to calculate his utility, but this does not require access to any long-run statistical data.

3 Examples

In this section I present three examples that illustrate the upper-bound problem for various specifications of the model. Throughout the examples, the set of types is simple, such that a DM's data type i can be identified with C_i. In each example, I hold $\gamma = p(t = 1)$ fixed and derive upper bounds on the DM's expected welfare loss, allowing $p(x, y \mid t)$ to vary.

Example 3.1
Let $K = 0$ and $n = 1$. This means that there is a unique data type, $C = \emptyset$. The DM's estimated causal effect of a on y is

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

Without loss of generality, $\Delta \geq 0$. The DM's subjective expected payoff from a given t is

$$p(y = 1 \mid a) - c \cdot \mathbf{1}[a \neq t]$$

It follows that in equilibrium, $\sigma(a = 1 \mid t = 1) = 1$. Denote $\sigma(a = 1 \mid t = 0) = \alpha$. In order for the DM to commit an error with positive probability, it must be that $\alpha > 0$. For this to be consistent with the DM's subjective optimization, it must be the case that $c \leq \Delta$. Therefore, the DM's expected welfare loss is at most

$$(1 - \gamma) \cdot \alpha \cdot [p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0)] \tag{4}$$

Denote $p(y = 1 \mid t) = \delta_t$. Observe that

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

It follows that to find an upper bound on (4), we should select $\delta_0 = 0$ and $\delta_1 = 1$. The expression then becomes

$$(1-\gamma)\cdot \alpha \cdot \frac{\gamma}{\gamma + (1-\gamma)\cdot \alpha}$$

This expression is increasing in α, and therefore bounded from above by $\gamma(1-\gamma)$.

To see how this upper bound can be approximated arbitrarily well, suppose that indeed, $y = t$ with probability one. Let $c = k\gamma$, where k is arbitrarily close to 1 from above. The DM plays the strategy $\sigma(a = 1 \mid t = 1) = 1$ and $\sigma(a = 1 \mid t = 0) = (1 - k\gamma)/(k - k\gamma)$. It can be checked that given this σ, $\Delta = c$, hence σ is an equilibrium. As $k \rightarrow 1$, the expected welfare loss approaches the upper bound $\gamma(1-\gamma)$.

This example demonstrates how the equilibrium assumption can "protect" the DM against causal errors, by limiting his welfare loss below $\gamma(1-\gamma)$. 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 c, the frequency of the DM's error is maximal. However, since in this case a does not vary with y, the estimated causal effect is negligible, hence c must be close to zero for this to be consistent with equilibrium behavior. In this case, too, the average welfare loss vanishes. More generally, a larger estimated causal effect implies a lower equilibrium frequency of making an error. This is why equilibrium behavior limits the DM's expected welfare loss due to wrong causal inference. □

In the next two examples, I assume $\gamma = 0$ — that is, the preference type is constant, $t = 0$, such that the DM's type is pinned down by his data type. In this case, $y \perp a \mid x$, hence the rational course of action is to play always $a = 0$.

Example 3.2
Let $K \geq 1$ and $n = 2$, where $C_1 \subset C_2$. I will establish that in this case, both types must play $a = 0$ with probability one in equilibrium. To see why, recall that by definition, type i does not condition his action on any variable outside C_i — that is, $\sigma_i(a \mid x) \equiv \sigma_i(a \mid x_{C_i})$. Since $C_1 \subset C_2$, this means that $p(y \mid a, x_{C_2}) \equiv p(y \mid a, x)$ — that is, $y \perp a \mid x_{C_2}$. As a result,

$$\tilde{p}_2(y = 1 \mid x_{C_2}, do(a = 1)) - \tilde{p}_2(y = 1 \mid x_{C_2}, do(a = 0)) = 0$$

such that in equilibrium, $\sigma_2(a = 1 \mid x_{C_2}) = 0$ for every x_{C_2}. This means that $p(y \mid a, x_{C_1}) \equiv p(y \mid a, x)$ — in other words, the variables in $C_2 \backslash C_1$ do not confound the joint distribution of a and y. As a result,

$$\tilde{p}_1(y = 1 \mid x_{C_1}, do(a = 1)) - \tilde{p}_2(y = 1 \mid x_{C_1}, do(a = 1)) = 0$$

such that in equilibrium, $\sigma_1(a = 1 \mid x_{C_1}) = 0$ for every x_{C_1}.

The intuition behind this result is as follows. Type 1 is vulnerable to interpreting correlation between a and y as a causal effect, because he does not control for variables in $C_2 \backslash C_1$. However, this problem does not arise in equilibrium because a does *not* vary in response to fluctuations in $x_{C_2 \backslash C_1}$ — type 1 never conditions on $x_{C_2 \backslash C_1}$ by definition, whereas type 2 correctly controls for it and thus never deviates from the correct action $a = 0$. □

Example 3.3
Let $K = 2$. The two exogenous variables x_1 and x_2 take values in $\{0, 1\}$, and their joint distribution is as follows:

$$\begin{aligned} p(x_1 &= 1) = \frac{1}{2} \\ p(x_2 &= 1 \mid x_1) = \rho \cdot x_1 \end{aligned}$$

for every x, where $\rho \in (0,1)$. That is, the marginal distribution of each variable is uniform, and ρ measures the extent to which they are (positively) correlated. Assume $y = x_1x_2$ with probability one.

Let $n = 2$, where $C_i = \{i\}$. That is, each type conditions his action on one of these variables. Assume $\lambda_1 = \lambda_2 = \frac{1}{2}$. I will establish that in this case, the equilibrium probability of $a = 1$ can be arbitrarily close to one.

Suppose that each type $i = 1, 2$ always plays $a_i = x_i$. Let us calculate type 1's subjective estimate of the causal effect of his action given his information. First, observe that since $y = x_1x_2$ independently of a,

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

for every a. Therefore, we only need to calculate the following quantities, which also make use of the DM's postulated strategy:

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

Based on these calculations, we obtain

$$\tilde{p}_1(y = 1 \mid x_1 = 1, do(a = 1)) - \tilde{p}_1(y = 1 \mid x_1 = 1, do(a = 0)) = \frac{\rho}{\rho + \frac{1}{2}(1-\rho)}$$

In addition, we established that

$$\tilde{p}_1(y = 1 \mid x_1 = 0, do(a = 1)) - \tilde{p}_1(y = 1 \mid x_1 = 0, do(a = 0)) = 0$$

Therefore, for every $c < 1$, we can find ρ sufficiently close to one such that the postulated strategy is consistent with equilibrium. By symmetry, the same holds for type 2. We have thus established that the upper bound on the equilibrium error is the same as without the equilibrium restriction.

Unlike Example 3.1, the two types in this example are not hierarchically ordered in terms of their sophistication. Type 1 fails to control for x_2, whereas type 2 fails to control for x_1. Therefore, each type i is vulnerable to interpreting the residual correlation between a and y after controlling for x_i as a causal effect. This misperception can be sustained in equilibrium. As a result, the equilibrium condition per se does not limit the maximal welfare loss that the DM may incur as a result of his false causal inferences. □

4 Analysis

In this section I derive upper bounds on the welfare loss that is consistent with equilibrium behavior, for various specifications of the space of possible data types and the space of possible joint distributions over t, x, y.

The characterization results will make use of the following binary relation P over data types.

Definition 4 *For data types $i, j \in N$, iPj if $D_i \supseteq C_j$.*

The meaning of iPj is that the set of variables that data type i controls for (via conditioning or adjustment) is a weak superset of the set of variables that type j conditions on. Note that by our definition of types, P is reflexive, since $D_i \supseteq C_i$ for every $i \in N$. 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.

Lemma 1 *Suppose a binary relation P over N is complete and quasitransitive. Then, N can be partitioned into L classes, $N_1, ..., N_L$, as follows: For every $\ell = 1, ..., L$,*

$$N_\ell = \{i \notin \cup_{h<\ell} N_h \mid j \not P^* i \text{ for all } j \notin \cup_{h<\ell} N_h\}$$

Moreover, for every $i \in N_\ell$, iPj for all $j \in \cup_{h \geq \ell} N_h$.

Proof. By definition, P^* does not contain cycles. Hence, the set of data types $i \in N$ such that $j \not P^* i$ for all $j \in N$ (i.e., the set of P^*-undominated data types) is non-empty. Define this set by N_1. Since P is complete, iPj for every $i \in N_1$ and every $j \in N$. The other cells in the partition are defined inductively: After $N_1, ..., N_\ell$ are removed from N, let $N_{\ell+1}$ be the set of P^*-undominated types in the remaining set. Since none of the sets N_ℓ is empty, the procedure terminates after at most n steps. ■

The lemma confirms that when P is complete and quasitransitive, it partitions N into layers, such that the first (top) layer consists of all P^*-undominated, the second layer consists of all P^*-undominated element outside the first layer, and so forth.

When all data types are simple, the structure of P is simplified: iPj means $C_i \supseteq C_j$, hence P is automatically transitive. The relevant distinction is thus between the case of complete or incomplete P. Moreover, since I assumed that all data types are distinct, it follows that for every pair of distinct types i, j, iPj implies iP^*j. Therefore, the requirement that P is complete is reduced to the requirement that P^* is a *linear ordering.*

4.1 The Case of $\gamma = 0$

In this sub-section, I derive upper bounds on the DM's equilibrium welfare loss when $t = 0$ with probability one (i.e., when playing $a = 1$ always inflicts a cost c on the DM). In this case, any correlation between a and y is due to mutual correlation with x. Since t does not vary, the DM's type coincides with his data type.

Proposition 1 *Let $\gamma = 0$. Suppose that P is complete and quasitransitive. Then, the DM's equilibrium expected welfare loss is zero.*

Proof. I will show that $a = 0$ with probability one in any equilibrium. The proof is by induction with respect to the partition defined by Lemma 1. 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. This means that $y \perp a \mid x_{D_i}$ — i.e.,

$p(y=1 \mid a, x_{D_i}) = p(y=1 \mid x_{D_i})$. Therefore, $\Delta_i(x) = 0$. 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 $a = 0$ for all x 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 (3) then implies $\Delta_i(x) = 0$. It follows that in equilibrium, type i plays $a = 0$ for all x. This completes the inductive proof. ■

Thus, when $\gamma = 0$ and the binary relation P is complete and quasitransitive — i.e., the data types are ordered in a certain sense — the equilibrium requirement fully "protects" the DM from choice errors due to flawed causal inference. It does so by shutting down the channels through which the choice behavior of some types could confound the relation between other types' actions and y. Types in the top layer of the P-based partition effectively control for all sources of correlation between a and y. As a result, their subjective best-replying implies that they do not generate any variation in choice behavior. This means that types in the next layer effectively control for all relevant x variables. This would not be the case if we did not impose the equilibrium condition on the behavior of top-layer types. This equilibrium effect spreads through all layers of the partition.

When P violates completeness or quasitransitivity, the picture is diametrically opposed.

Proposition 2 *Let $\gamma = 0$. Suppose that P violates completeness or quasitransitivity. Then, for any $\beta > 0$, there exist c, λ and $(p(x,y))$ such that an expected welfare loss above $1-\beta$ can be sustained in equilibrium.*

Proof. 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 $\lambda_1, \lambda_2 > 0$ and $\lambda_1 + \lambda_2 = 1$. Construct p as follows. First, let $x_1, x_2 \in \{0, 1\}$, 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 $p(x, y)$, let us now construct an equilibrium in which for each type $i = 1, 2$, $a_i = x_i$ with probability one. Without loss of generality, consider type 1's reasoning. This type's perceived causal effect of a on y given x_1 is

$$\Delta_1(x_1) = p(y = 1 \mid a = 1, x_1) - p(y = 1 \mid a = 0, x_1)$$

because all other variables are either not in D_1 or irrelevant for the determination of y and therefore can be ignored. Note that since $y = x_1 x_2$ with probability one,

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

for every a. Neither of these two formulas ever conditions on null events, because all four combinations of (a, x_1) occur with positive probability. (Note that the combination $a = 1$ and $x_1 = 0$ arises when $x_2 = 1$, because of the DM's strategy.) By our construction of $p(x_1, x_2)$ and the DM's strategy,

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

It follows that $\Delta_1(x_1 = 0) = 0$, while $\Delta_1(x_1 = 1) \to 1$ as $\varepsilon \to 0$. Therefore, type 1's postulated strategy can be consistent with equilibrium for values of

c that are 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 $3P^*1$. By the definition of P^*, 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 $i \in C_i \setminus D_j$ for all other $j \in \{1, 2, 3\}$. Suppose $\lambda_1, \lambda_2, \lambda_3 > 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Let $x_1, x_2, x_3 \in \{0, 1\}$. 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 $j, k \in \{1, 2, 3\} \setminus \{i\}$, where $\varepsilon > 0$ is arbitrarily small. Second, let $p(y = 1 \mid x_1, x_2, x_3) = x_1x_2x_3$. 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. Showing that this strategy is an equilibrium proceeds essentially along the same lines as in the case of an incomplete P — namely, we show that $\Delta_i(x_i = 1) \rightarrow 1$ as $\varepsilon \rightarrow 0$ and that $\Delta_i(x_i = 0) = 0$. Since the calculation is essentially the same (it is based on deriving expressions for $p(x_2x_3 = 1 \mid a, x_1)$ for all a, x_1), I omit it for brevity. ■

Thus, the upper bound on the DM's equilibrium welfare loss due to wrong causal inferences critically depends on whether the binary relation P is complete and quasitransitive. When it is, it is impossible that the equilibrium behavior of some data types will generate a variation that will produce confounding patterns, which other data types will misinterpret as causal. When it is not, the equilibrium behavior of different types can create such confounding patterns that mutually sustain their causal-inference errors. In that case, the equilibrium assumption does not constrain the maximal possible welfare loss due to these errors.

For an economic example that illustrates the relevance of the findings of this sub-section, suppose that the DM is a company executive contemplating a business decision. Suppose further that x has a technological dimension (x_1) and a financial dimension (x_2). The executive may have expertise in either of these two dimensions. This expertise means that the executive has data about the variable and how it correlates with business decisions and business outcomes.

In this setting, suppose that there are two types of executives: sophisticates, who possess full data about a, x, y; and simpletons, who posses data about a, y but lack expertise about any of the x variables. Then, Proposition 1 implies that the executive will commit no error in equilibrium. Alternatively, suppose that the two types differ in their area of expertise: one type only has technological expertise, while the other type only has financial expertise. Then, by Proposition 2, we can construct a joint distribution over x, y such that each executive type will sometimes take the wrong business decision in equilibrium.

4.2 The Case of $\gamma > 0$ with Simple Data Types

In this sub-section, I obtain partial results for families of distributions for which $\gamma > 0$ — i.e., when there is variation in the DM's preference type. Denote $\delta_t = p(y = 1 \mid t)$. Without loss of generality, assume $\delta_1 \geq \delta_0$. Throughout this subsection, I restrict attention to simple data types, as defined in Section 2.

I begin by imposing the domain restriction 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, and the x variables are only potential correlates of this statistic. For instance, whether a student regards studying as a costly or pleasurable activity is the best predictor of her school performance. This attitude (which is not observable to others) may be correlated with observable socioeconomic indicators, which nevertheless contribute no additional predictive power.

Proposition 3 *Suppose that all data type are simple and that P is complete. If $y \perp x \mid t$, then the DM's expected welfare loss in equilibrium is at most $\gamma(1-\gamma)$.*

Proof. The proof proceeds stepwise. Recall that since P is complete, P^* is a linear ordering. For convenience, enumerate the types according to P^* — i.e., $1P^*2P^* \cdots P^*n$. 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

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

Plugging this in (3), we obtain

$$\Delta_i(x) = [p(t = 1 \mid a = 1, x_{C_i}) - p(t = 1 \mid a = 0, x_{C_i})][\delta_1 - \delta_0] \tag{5}$$

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 (5) can be written as

$$\frac{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i})}{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i}) + (1 - \gamma(x_{C_i}))p(a \mid t = 0, x_{C_i})} \tag{6}$$

We will now focus on the term $p(a \mid t = 1, x_{C_1})$. Note that

$$p(a \mid t, x_{C_1}) = \sum_{x_{-C_1}} p(x_{-C_1} \mid t, x_{C_1})p(a \mid t, x_{C_1}, x_{{}_{-C_1}}) \tag{7}$$

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, (7) 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 and every 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 (6) is weakly increasing in a for $i = 1$. Since $\delta_1 - \delta_0 \geq 0$, it follows that $\Delta_1(x) \geq 0$.

Step 3: Extending the property of 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, $\sigma_j(a = 1 \mid t = 1, x_{C_j}) = 1$ for every $j \leq m$. By definition, $C_j \subseteq C_i$ for every $j \geq m+1$, hence $\sigma_j(a \mid t, x_{C_j})$ is constant in x_{-C_i}. 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 $p(a = 1 \mid t = 1, x_{C_i}) \geq p(a = 1 \mid t = 0, x_{C_i})$. As in the proof of Step 2, applying this inequality to (6) implies that $\Delta_i(x) \geq 0$ and $\sigma_i(a = 1 \mid t = 1, x_{C_i}) = 1$. This completes the inductive proof.

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

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

In order for type $i(x)$ to play $a = 1$ given x and $t = 0$, it must be the case that $c \leq \Delta_{i(x)}(x)$. 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 (5)-(6) 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 $C_i \supseteq C_j$ 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 DM's expected welfare loss given x is bounded from above by

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

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

$$\sum_{x} p(x)(1 - \gamma(x)) \cdot \sum_{x'} \beta(x', x)\gamma(x') \tag{8}$$

where $\beta(\cdot)$ is a system of convex combinations, $\beta(x', x) = p(x' \mid x'_{C_{i(x)}} = x_{C_{i(x)}})$. Expression (8) 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 $\gamma(1 - \gamma)$. ■

Thus, when the set of types is simple and y is only determined by t, the DM's expected equilibrium welfare loss is at most $\gamma(1 - \gamma)$. Example 3.1

established the tightness of this bound. This result also means that across all distributions that satisfy $y \perp (x,a) \mid t$, the expected welfare loss is at most $\frac{1}{4}$ — compared with the non-equilibrium upper bound of 1. This is yet another demonstration of how the equilibrium condition restricts the cost of faulty causal inferences. As $\gamma \to 0$, this loss converges to zero.

When completeness of P is relaxed, finding the upper bound on the DM's expected welfare loss when $y \perp x \mid t$ is an open problem. However, the following result establishes that this bound must exceed $\gamma(1-\gamma)$ whenever $\gamma \neq \frac{1}{2}$. Without loss of generality, let $\gamma < \frac{1}{2}$.

Proposition 4 *For every γ, there exist c, a distribution λ over simple data types and a distribution $(p(x,y \mid t))$ satisfying $y \perp x \mid t$, for which there is an equilibrium in which the DM's expected welfare loss is arbitrarily close to $\frac{1}{2}\sqrt{\gamma(1-\gamma)}$.*

Proof. Let K be arbitrarily large. Let $n = K$, and $C_k = \{k\}$ for every $k = 1, ..., K$. Let $\lambda_k = \frac{1}{K}$ for every k.

Suppose $x_k \in \{0,1\}$ for every $k = 1, ..., K$. For every subset $M \subset \{1, ..., K\}$, define $x(M)$ as follows: $x_k = 1$ if and only if $k \in M$. Fix $m \in \{1, ..., K-1\}$. Define $(p(x \mid t))$ and $(p(y \mid t))$ as follows:

$$\begin{aligned} p(y &= t \mid t) \equiv 1 \\ p(x &= (1, ..., 1) \mid t = 1) = 1 \\ p(x(M) &\mid t = 0) = \frac{1-\gamma}{\binom{K}{m}} \text{ for every } M \text{ such that } |M| = m \end{aligned}$$

Suppose each data type i plays $a = 1$ if and only if $x_k = 1$. Let us calculate $\Delta_i(x)$ for every x and i. Without loss of generality, consider type 1, for whom $C_1 = \{1\}$. Since $y = t$ with certainty, it suffices to calculate $p(t = 1 \mid x_1, a)$. First,

$$p(t = 1 \mid x_1 = 1,\ a = 1) = \frac{\gamma}{\gamma + (\frac{m}{K})^2(1-\gamma)}$$

Let us elaborate on this expression. Consider the numerator. The probability of $t = 1$ is γ, and in this case $x_k = 1$ for all k, and all data types play $a = 1$.

As to the second term in the denominator, the probability that $t = 1$ and $x_1 = 1$ is $(1 - \gamma)m/K$; and the fraction of data types who play $a = 1$ given any realization of $t = 0$ is exactly m/K.

Next, $p(t = 1 \mid x_k = 1, a = 0) = 0$ because given the DM's assumed strategy, the realization $a = 0$ occurs with positive probability at $x_1 = 1$ only when $t = 0$ (in this case, it was one of the other data types who play $a = 0$). It follows that

$$\Delta_1(x_1 = 1) = \frac{\gamma}{\gamma + (\frac{m}{K})^2(1 - \gamma)}$$

Finally, $p(t = 1 \mid x_1 = 0, a) = 0$ for all a because by construction, the realization $x_1 = 0$ occurs with positive probability only when $t = 0$. It follows that $\Delta_1(x_1 = 0) = 0$.

Therefore, if $c < \Delta_1(x_1 = 1)$, the DM's strategy is consistent with equilibrium. In this case, the expected welfare loss of data type 1 can be arbitrarily close to

$$\frac{m}{K}(1 - \gamma)\frac{\gamma}{\gamma + (\frac{m}{K})^2(1 - \gamma)} \tag{9}$$

because this type plays $a \neq t$ when $t = 0$ and $x_1 = 1$. Since this same calculation applies to all data types, the DM's expected welfare loss is given by (9). The value of m/K that maximizes this expression is $\sqrt{\gamma/(1 - \gamma)}$. We can find values of m, K that get arbitrarily close to this value. Plugging it into (9), we obtain $\frac{1}{2}\sqrt{\gamma(1 - \gamma)}$. ■

The final result in this sub-section lifts all restrictions on $(p(x, y \mid t))$ and P 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 that all data types are simple and that P is incomplete. Then, for every γ, there exist c, λ and $(p(x, y \mid t))$ for which there is an equilibrium in which the DM's expected welfare loss is arbitrarily close to 1.*

Proof. Since P is incomplete, there are two data types, denoted conveniently 1 and 2, and two variable indices, also denoted 1 and 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 variables, such that they can be ignored), where $\varepsilon > 0$ is arbitrarily small:

$p(t, x_1, x_2, y)$	t	x_1	x_2	y
$1 - \gamma - \varepsilon$	0	1	1	1
$\gamma - \varepsilon$	1	0	0	1
ε	0	1	0	0
ε	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,

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

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,

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

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 $c < 1$, we can set ε 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. ■

The partial results in this section leave a few open problems. First, is the upper bound obtained in Proposition 4 tight when y is purely a function of t? Second, does the upper bound obtained for complete P in Proposition 3 extend to arbitrary distributions p? Finally, do the results extend to general data types?

5 Consequential Actions

So far, we focused on the extreme case in which the DM's action has a null objective causal effect on the outcome. This facilitated the definition of the DM's equilibrium welfare loss due to poor controls. In this section I extend the analysis to situations in which actions do influence outcome.

Define a variable z that takes values in 0 and 1, such that the objective causal model behind the joint distribution over t, x, z, a, y is given by the DAG

$$\begin{tikzcd} (t,x) \arrow[r] \arrow[d] & a \arrow[d] \\ z \arrow[r] & y \end{tikzcd}$$

That is, t and x are exogenous, as before. The action a is a consequence of (t, x), via the DM types' strategies. The variable z is also 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, according to the following conditional probability:

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

where $\beta \in (0, 1)$.

This formulation implies that for every type i, the perceived outcome of actions is given by

$$\tilde{p}_i(y = 1 \mid x_{C_i}, do(a)) = \beta a + (1 - \beta)\tilde{p}_i(z = 1 \mid x_{C_i}, do(a))$$

where the last term is defined just as in the baseline model:

$$\tilde{p}_i(z = 1 \mid x_{C_i}, do(a)) = \sum_{x_{D_i}} p(x_{D_i} \mid x_{C_i})p(z = 1 \mid a, x_{D_i})$$

The type's estimated causal effect of a on z given x is

$$\Delta_i^z(x) = \tilde{p}_i(z = 1 \mid x_{C_i}, do(a = 1)) - \tilde{p}_i(z = 1 \mid x_{C_i}, do(a = 0))$$

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

$$\beta a - c \cdot \mathbf{1}[a \neq t]$$

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

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

note that in equilibrium, the DM chooses $a = 0$ at $t = 1$ and x only if $c + \beta < -(1 - \beta)\Delta_i^z(x)$. Likewise, the DM chooses $a = 1$ at $t = 0$ and x only if $c - \beta < (1 - \beta)\Delta_i^z(x)$. Consequently, the upper bounds on the DM's equilibrium welfare loss are the same as in the previous section, multiplied by $1 - \beta$.

An example: Partying during a pandemic

Although the paper emphasized upper bounds on the equilibrium costs of using bad controls for causal inference, in economic applications we wish to restrict the objective process so that it can capture an underlying economic reality. I now present a simple example of such an application.

Suppose that $a = 1$ means that the DM chooses to socially distance himself during a pandemic — specifically, avoiding parties. The outcome $y = 1$ represents good health. Let x represent the DM's age ($x = 1$ indicates an old DM). Let t represent the DM's intrinsic taste for partying — $t = 1$ means that the DM prefers not to go to parties. Let $c < \frac{1}{2}$.

The objective distribution p satisfies: $p(x = 1) = \frac{1}{2}$; $p(t = x \mid x) = q$ for all x, where $q \in (\frac{1}{2}, 1)$; and $p(y = 1 \mid a, x) = \frac{1}{2}(a + 1 - x)$. This distribution is consistent with the DAG

$$\begin{tikzcd}
t \arrow[d] & x \arrow[l] \arrow[dl] \arrow[d] \\
a \arrow[r] & y
\end{tikzcd}$$

That is, y is only caused by a and x. When an old DM goes to parties, his health outcome is bad with certainty; when a young DM avoids parties, his health outcome is good with certainty; in all other cases, the DM's health outcome is equally like to be good or bad.

Data type 1 controls for x. This type correctly estimates the causal effect of switching from $a = 0$ to $a = 1$ on y to be $\frac{1}{2}$. Since $c < \frac{1}{2}$, this DM data type will rationally play $a = 1$, independently of t and x.

Data type 2 does not control for x (recall that even if it is obviously natural to assume that the DM knows his age group, the DM may lack statistics about the age dependence of the correlation between a and y, and therefore cannot use his knowledge of his age). This DM chooses a to maximize

$$p(y = 1 \mid a) - c \cdot \mathbf{1}[a \neq t] = \frac{1}{2} + \frac{1}{2}a - \frac{1}{2}p(x = 1 \mid a) - c \cdot \mathbf{1}[a \neq t]$$

Let us analyze the equilibrium in this example. As we saw, data type 1's strategy is $\sigma_1(a = 1 \mid t) = 1$ for all t, x. Denote $\sigma_2(a = 1 \mid t) = \alpha_t$ (recall that type 2 does not condition his action on x). Then,

$$\begin{aligned}
p(x &= 1 \mid a = 1) = \frac{\lambda_1 + \lambda_2[q\alpha_1 + (1 - q)\alpha_0]}{2\lambda_1 + \lambda_2[\alpha_1 + \alpha_0]} \\
p(x &= 1 \mid a = 0) = \frac{1 - q\alpha_1 - (1 - q)\alpha_0}{2 - \alpha_1 - \alpha_0}
\end{aligned}$$

Let us first guess

$$p(x = 1 \mid a = 1) - p(x = 1 \mid a = 0) < \frac{1}{2} - c$$

Then, $a = 1$ is optimal for data type 2 regardless of t. In this case, we need to consider perturbed strategies to ensure that $p(x = 1 \mid a = 0)$ is well-defined. Since α_0 and α_1 are arbitrarily close to 1, we obtain $p(x = 1 \mid a = 1) \approx \frac{1}{2}$, whereas we can set the perturbations such that $p(x = 1 \mid a = 0)$ can take any value in $(1 - q, q)$. It follows that it is always possible to sustain the guess in equilibrium, such that the DM will commit no error.

Let us now guess

$$p(x = 1 \mid a = 1) - p(x = 1 \mid a = 0) > \frac{1}{2} - c$$

Then, data type 2 will play $\alpha_t \equiv t$ in equilibrium. Plugging this into the expressions for $p(x = 1 \mid a)$, we obtain

$$p(x = 1 \mid a = 1) - p(x = 1 \mid a = 0) = \frac{\lambda_1 + \lambda_2 q}{2\lambda_1 + \lambda_2} - (1 - q)$$

It follows that if

$$c > \frac{1}{2} - \frac{2q - 1}{1 + \lambda_1}$$

the guess is consistent. In this case, there is an equilibrium in which type 2 follows his taste.

What sustains this equilibrium is the positive correlation between age and preferences. Young DMs like going to parties more than old DMs, and since the DM chooses according to his intrinsic taste with some probability (λ_2), there is positive correlation between attending parties and young age. In turn, this soften the negative correlation between a and y, to an extent that makes it optimal for type 2 DMs to follow their taste. The welfare loss in this equilibrium is

$$\frac{1}{2} \cdot \lambda_2 \cdot (\frac{1}{2} - c) < \frac{(2q - 1)(1 - \lambda_1)}{2(1 + \lambda_1)}$$

Note that the welfare loss decreases with the fraction of type 1. There are two forces behind this observation. First, lower λ_1 obviously means that there are more DMs in the population who are prone to error. Second, type 1 DMs do not vary their behavior with t (and hence with x), thus curbing the overall positive correlation between a and x that leads type 2 DMs to underestimate the causal effect of a on y.

There is potentially a third equilibrium in which $\alpha_1 = 1$ and $\alpha_0 \in (0, 1)$, such that

$$p(x = 1 \mid a = 1) - p(x = 1 \mid a = 0) = \frac{1}{2} - c$$

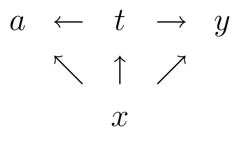
For brevity, I omit the full characterization of this equilibrium.

6 Relation to Other Solution Concepts

The model of behavioral causal inference presented in this paper poses a new question. However, it can be formulated in terms of adaptations of existing frameworks of equilibrium modeling with non-rational expectations.

Jehiel's (2005) concept of analogy-based expectations equilibrium (ABEE) 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, C_i defines type i's information set, whereas D_i defines type i's "analogy partition" (to use Jehiel's terminology). Two histories belong to the same cell in this partition 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. Thus, the model in this paper can be formulated as an application of ABEE.

The model can also be cast in the Bayesian-network language of Spiegler (2016). The objective distribution p in the baseline model (where a has no causal effect on y) is consistent with the following DAG:



In contrast, type i believes in the subjective causal model (organizing the variables on which he has data) given by the following DAG:

$$\begin{tikzcd}
x_{D_i \setminus C_i} \arrow[r] & y \\
x_{C_i} \arrow[u] \arrow[ur] \arrow[r] & a \arrow[u]
\end{tikzcd}$$

According to Spiegler (2016), the subjective belief that this model generates 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 conditional belief over y as a consequence of a given x_{C_i} is described by (2). Equilibrium in the present model is consistent with the notion of personal equilibrium in Spiegler (2016,2020) when the DM's subjective causal model is random.

The Bayesian-network framework in Spiegler (2016) can be subsumed into the more general concept of Berk-Nash equilibrium due to Esponda and Pouzo (2016). Adapting that concept to the present environment is feasible but less straightforward, partly because it calls for an adaptation of how Kullback-Leibler divergence features in the equilibrium definition.

References

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

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

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

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

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

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

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

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