Behavioral Causal Inference[*]

Ran Spiegler[†]

June 27, 2023

Abstract

When inferring the causal effect of one variable on another from correlational data, a common practice by professional researchers, and to a lesser extent by lay decision makers, is to control for some set of exogenous confounding variables. An inappropriate set of control variables can lead to erroneous causal inferences. This paper presents a model of 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 "vertically differentiated" 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 and the Foerder Institute is gratefully acknowledged. I thank Alex Clyde, Nathan Hancart, Heidi Thysen, and numerous seminar participants, 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. 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 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 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 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 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* rather than conditioning. 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. Even if he does not know his own relevant genetic background, he can nevertheless adjust his beliefs according to the available data about the correlation between this variable and others.

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 $x_1, ..., x_K$. The DM is able to control for the variables indexed by $D \subseteq \{1, ..., K\}$; he conditions on a subset $C \subseteq D$, and adjusts for the variables in $D \setminus C$. The DM's estimated causal effect of a on y is given by the formula

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

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

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 an 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 null 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 because there is no causal path from a to y. However, in this case the quantity $p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0)$ 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 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?[1]

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 beneficial causal effect on y. In the baseline model, 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 independently from some given set N, which is defined by a distinct pair (D,C) as described above, 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 variables. 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 how this solution concept can be recast in earlier frameworks of equilibrium modeling with non-rational expectations, due to 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. Agents' response to their beliefs change these very patterns, and hence the causal effects they deduce from them.

Example 1.1

The previous pair of examples of "bad controls" offer an extreme illustration of this insight. Suppose that $t = 0$ with certainty — i.e., there are no preference shocks. In the first example, the single exogenous variable x which causes y is also the sole direct cause of a. For the latter causal relation to be

[1] Eliaz et al. (2021a) perform worst-case analysis of estimated correlations due to causal misperceptions. Spiegler (2022) presents an example of how equilibrium forces can restrict the cost of committing a reverse-causality misperception.

non-null, however, it must be the case that some DM data types condition their action on x. Since these same types control for x, they 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. This means that in equilibrium, the DM will not incur *any* welfare loss due to poor causal inference.

The same reasoning applies to the second example, which involves three exogenous x variables. 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. Since no DM type varies his behavior with x_1, the link $a \leftarrow x_1$ that makes x_2 a "bad control" is effectively severed. □

The main results in this paper — presented in Sections 3 and 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 expected payoff from the rational-expectations strategy $a \equiv t$. When a has no causal effect on y, the welfare loss is simply $c \cdot \Pr(a \neq t)$.

It turns out that a simple binary relation over the set of 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 controls for every variable the latter type conditions on. When t is constant, the upper bound is 0 when the domination relation over N is complete and quasitransitive, and c when it is not.[2] Thus, when data types are ordered in a particular 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

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

action distribution. The former case fits situations in which DM types are "vertically differentiated", (roughly) in the sense that some types control for more variables than others. The latter case fits "horizontal" differentiation, in the sense that different types control for different variables.

I obtain partial characterization results when there is variation in t. Specifically, I assume that $D = C$ for all data types, and that t is the sole direct cause of y. When the domination relation is complete (and therefore transitive), 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 is $\max\{\Pr(t = 1), \Pr(t = 0)\}$. When we relax all restrictions on the set of data types or the data-generating process, the upper bound is 1. Once again, whether data types are "vertically" or "horizontally" differentiated plays a key role in how equilibrium forces constrain the cost of flawed causal inference.

2 A Model

Let a, t and y be three 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 that are realized jointly with t, 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 $a = t$.

There is a set $N = \{1, ..., n\}$ of DM *data types*. Each type $i \in N$ is associated with a distinct pair (C_i, D_i), where $C_i \subseteq D_i \subseteq \{1, ..., K\}$. The interpretation is that C_i defines the set of x variables that type i can condition on because he observes their realization before taking an action; and D_i is the set of exogenous variables about which he has long-run data (note that t is never among these variables). 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, which is *independent* of all other variables (this independence 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 $\sigma_{t,i} : X \to \Delta(A)$. By definition, this strategy is measurable with respect to X_{C_i}.

Let p be a joint probability distribution over t, x, a, y. I interpret p as a steady-state or long-run distribution. 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:

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

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}:

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

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

The $\tilde{p}$ notation indicates that this is a subjective belief. 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. As to the other variables (represented by $D_i \backslash C_i$), the DM 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.[4]

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

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

Plugging (2) into this definition, we obtain:

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

This formula will serve us throughout this paper.

If the DM had long-run data about all exogenous variables (including t), he could control for all of them, and thus correctly infer the action's null causal effect. 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

[4] Formula (2) can also be interpreted in terms of standard subjective expected utility, where the state space itself is *subjective*: X_{D_i} is type i's subjective state space and X_{C_i} is his set of signals.

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. 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 1 *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*

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

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

The trembling-hand aspect of the equilibrium concept is required to ensure that all conditional probabilities it involves are well-defined. The exact 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

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

Since a has no causal effect on y, playing $a \neq t$ yields a welfare loss.

Definition 2 (Expected welfare loss) *Given a strategy profile σ, the DM's expected welfare loss is*

$$c \sum_{t,x} p(t, x) \sum_{i \in N} \lambda_i \sigma_i(a \neq t \mid t, x) \tag{4}$$

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 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, by definition, the probability of error is one. And if $c \approx 1$, this means that the welfare loss is approximately 1.

However, the strategy σ is inconsistent with equilibrium, for essentially the same reason as in Example 1.1. 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 $a \neq t$ for some realization of x. It follows that the requirement that σ is an equilibrium strategy can have bite.

Comment: Why does $C \subseteq D$? 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 easily 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. 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.

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 two features of the receiver's environment. The conventional feature is a distribution over the receiver's signals. The less conventional feature (but one that is closer in spirit to Eliaz et al. (2021b)) involves the long-run statistical data to which the receiver has access, according to which he forms his beliefs. Worst-case analysis can thus be viewed as finding the sender's optimal data provision strategy.

3 Analysis: Homogenous Preferences

In this section I characterize the maximal welfare loss that is consistent with equilibrium behavior, when there is no variation in the DM's preferences. Specifically, assume that $t = 0$ with probability one, such that the DM's expected welfare loss is simply c times the ex-ante probability that he plays $a = 1$. I show that the upper bound on this probability depends on a simple property of the set of data types.

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. Therefore, 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 following sequence of perturbations around this strategy: for every $\varepsilon \in (0, \frac{1}{2})$, every data type i plays $a = 1$ with probability ε, 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, and how large this probability can get. The following example serves to illustrate this problem.

Example 3.1
Let $K = 2$. The two exogenous variables x_1 and x_2 take values in $\{0, 1\}$, and their joint distribution satisfies:

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

Let $n = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, where $C_i = D_i = \{i\}$. That is, each type conditions his action on a distinct aspect of x and fails to adjust for the other.

The following is an interpretation of this specification. A business executive chooses a strategy for a company whose environment is defined by financial and technological factors (represented by x_1 and x_2). The company is profitable if both factors are favorable. The executive's decision is informed by an analyst's report. There are two types of analysts, who specialize in (and therefore monitor) the technological and financial environments, respectively.

Suppose that each type $i = 1, 2$ always plays $a = x_i$. Let us examine whether this strategy profile is an equilibrium. Begin by calculating type 1's subjective estimate of actions' causal effect on profits, given his information. First, observe that since $y = x_1 x_2$ independently of a,

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

for every a. (Note that these quantities never involve conditioning on a zero-probability event. For example, the combination $a = 0, x_1 = 1$ occurs when $x_2 = 0$ and the DM is of type 2.) Therefore, we only need to calculate the following conditional probabilities, which also make use of the DM's postulated strategy:

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

It follows that

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

Therefore, if $2q/(1+q) > c$, 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 $q > c/(2-c)$, 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. Thus, unlike Example 1.1, here equilibrium forces do not "protect" DMs from their errors of causal inference.

The intuition behind this result is that since type i conditions his action on 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 behavior — as a causal effect. Note that the result does *not* necessitate correlation between x_1 and x_2. Indeed, even when $q = \frac{1}{2}$, the above equilibrium can be sustained as long as $c < \frac{2}{3}$. 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 that 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 that type $C_1 = \{1\}$ and $C_2 = D_2 = \emptyset$ — i.e., type 2 now does not control for any variable. By definition, he does not vary his action with x, and therefore x_2 is not a confounding variable. This means that type 1 effectively controls for any potential confounder, and therefore he will not commit any error in equilibrium. □

Examples 1.1 and 3.1 demonstrate that for some sets of data types, the equilibrium welfare loss is zero, while for others, it can be as large as when we do not impose any equilibrium restriction. The results in this section generalize this lesson. They will make use of the following binary relation P over data types.

Definition 3 *For data types $i, j \in N$, 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.

Lemma 1 *Suppose a binary relation P over N is complete and quasitransitive. Then, N can be partitioned into L classes, $N_1, ..., N_L$, such that 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 $j \in \cup_{h \geq \ell} N_h$.

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 types, the second layer consists of all P^*-undominated types outside the first layer, and so forth.

When all data types are *simple* (i.e., $C_i = D_i$ for all $i \in N$), the structure of P is simplified: iPj means $C_i \supset C_j$, hence P is automatically asymmetric and transitive.[5] The relevant distinction in this case is thus between complete and incomplete P. Furthermore, if P is complete, it is a *linear ordering* over N.

The following results fully characterize the maximal equilibrium welfare loss, as a function of P. The first result generalizes Example 1.1, whereas the second result generalizes Example 3.1.

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

[5] Strict containment follows from our assumption that all data types are distinct. Thus, when all types are simple, $C_i \neq C_j$ whenever $i \neq j$.

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. 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 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 spreads through all layers of the partition.

Proposition 2 *Let $\gamma = 0$. Suppose P violates completeness or quasitransitivity. Then, for any $c, \beta \in (0, 1)$, there exist λ and $(p(x, y))$ such that* $\Pr(a = 1) > \beta$ *in some equilibrium. In particular, when $c \approx 1$, the equilibrium welfare loss can be arbitrarily close to* 1.

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, the equilibrium behavior of some data types cannot generate a variation that produces confounding patterns that other data types 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. The proof is constructive, involving a more elaborate version of Example 3.1.

The distinction between the two cases can be described as a distinction between "vertical" and "horizontal" differentiation among data types. This is especially palpable in the case of simple types, where the results hinge on whether P^* is a linear ordering. When it is, the types' sets of control variables are ordered by set inclusion, and in this case the equilibrium welfare loss is zero. When it is not, the difference between types is that they control for different variables, and this "horizontal" differentiation enables them to create mutually reinforcing confounding patterns.

4 Analysis: Heterogeneous Preferences

In this section I reintroduce preference heterogeneity, by assuming $\gamma \in (0,1)$. Unlike the homoegenous-preference case, here I lack a complete characterization of the maximal equilibrium welfare loss, and present a number of partial results. In particular, I restrict attention to *simple data types*, as defined in Section 2 — that is, $C_i = D_i$ for every data type i. Recall that in this case, P is complete if and only if it is a linear ordering. Denote $\delta_t = p(y = 1 \mid t)$. Without loss of generality, assume $\delta_1 \geq \delta_0$.

Example 4.1
Suppose $\delta_t \equiv t$. Let $K = 0$ and $n = 1$ — i.e., there is a unique data type, with $C = \emptyset$. One interpretation for this setting 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 therefore 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 show establish uniqueness of equilibrium in this setting, and characterize the DM's expected 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 $\alpha_t = \sigma(a = 1 \mid t)$. When the DM's strategy is fully mixed, $\alpha_t \in (0,1)$ for every 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 $\delta_1 = 1 > 0 = \delta_0$ 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 $t = 1$ and the estimated effect of his choice on y. Plugging the known expressions for α_1 and δ_t and taking the $\varepsilon \rightarrow 0$ limit, we obtain

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

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

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

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

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

By setting $c \approx \gamma$, 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 $\varepsilon \rightarrow 0$, we obtain $p(y = 1 \mid a = 1) \approx \gamma$ and $p(y = 1 \mid a = 1) \approx 0$. If $\gamma > c$, this is consistent with equilibrium. The DM's welfare loss in this equilibrium is

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

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

Thus, for any configuration of c 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 $c \approx \gamma$. 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 against 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 c, the frequency of the DM's error is maximal. However, since in this case a varies little with y, the estimated causal effect is smaller.

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 welfare loss due to failure to control for x. □

Let us now turn to characterizations of the upper bound on the DM's equilibrium welfare loss, under certain restrictions on the data-generating process. 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; 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 cause of her school performance. This attitude (which is not observable to others) may be correlated with observable socioeconomic indicators, but these are only indirect causes or mere proxies for the true cause.

Proposition 3 *Suppose all data types in N are simple and P is complete. 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. 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 can restrict the decision cost of faulty causal inferences. When $\gamma \to 0$, this loss converges to zero.

As in the case of Proposition 1, the proof of Proposition 3 proceeds by induction on the set of data types, starting with the type having the largest set of control variables. 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, who do not control for every x variable he conditions on. This makes the inductive proof considerably more intricate than the proof of Proposition 1. A key argument in the proof is that while the different data types may disagree on the magnitude of the causal effect of a on y, they all agree on its *sign*, which is always $sign(\delta_1 - \delta_0)$. This feature holds in any equilibrium when P is complete.

When completeness of P is relaxed, the tight upper bound on the DM's expected welfare loss when $y \perp x \mid t$ is significantly higher.

Proposition 4 *Suppose all data types are simple and P is incomplete. If $y \perp x \mid t$, then the DM's expected welfare loss in equilibrium is at most* $\max(\gamma, 1-\gamma)$. *Moreover, this upper bound can be approximated arbitrarily well, by appropriately selecting c, λ and $(p(x, y \mid t))$, if we allow each x_i to get at least three values.*

This result carries the relevance of the distinction between complete and incomplete P to the setting with preference heterogeneity. The gap between the upper bounds in the two cases ($\gamma(1-\gamma)$ vs. $\max(\gamma, 1-\gamma)$) is significant. 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. In addition, unlike the case of complete P, the sign of the DM's estimated causal effect need not be constant; indeed, this feature plays an important role in the implementation of the upper bound.

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 all data types are simple and P is incomplete. Then, for every $\gamma, c \in (0,1)$, 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 three open problems. First, does the upper bound of $\gamma(1-\gamma)$ obtained for complete P in Proposition 3 extend to distributions p that violate $y \perp x \mid t$? Second, do the results extend to general (non-simple) data types? Finally, how do results change when the distribution over data types is allowed to be correlated with t and x?

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:

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

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

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

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

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

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

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

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

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

$$\gamma \cdot p(a=0 \mid t=1) \cdot (c+\beta) + (1-\gamma) \cdot p(a=1 \mid t=0) \cdot (c-\beta) \qquad (5)$$

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, by (5), the upper bounds on the DM's equilibrium welfare loss are the same as in Sections 3-4, multiplied by $1 - \beta$.

An example: Partying during an epidemic

This paper performed worst-case analysis of the equilibrium welfare implications of using bad controls for causal inference. Nevertheless, in economic applications we wish to restrict the objective process so that it can capture an underlying economic reality, and typically this process will not implement the worst case. I now present a simple example of such an application.

Suppose that $a = 1$ means that the DM chooses to socially distance himself during an epidemic — 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 dislikes 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 health effect of switching from $a = 0$ to $a = 1$ 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 the knowledge of his age). This DM chooses a to maximize

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

Let us analyze equilibria in this example.

Claim 1 *The rational-choice benchmark can be sustained in equilibrium.*

To prove this claim, recall that data type 1's strategy is $\sigma_1(a = 1 \mid t, x) = 1$ for all t, x. Denote $\sigma_2(a = 1 \mid t) = \alpha_t$. Then,

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

First, let us 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 $p(x = 1 \mid a = 1) \approx \frac{1}{2}$. We can also set the perturbations such that $p(x = 1 \mid a = 0) = \frac{1}{2}$. It follows that it is always possible to sustain the guess in equilibrium, such that the DM will commit no error.

Claim 2 *Assume*

$$c > \frac{1}{2} - \frac{2q - 1}{1 + \lambda_1} \tag{6}$$

Then, there is an equilibrium in which type 2 *always plays* $a = t$*.*

To verify this claim, let us 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

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

Condition (6) means that this expression exceeds $\frac{1}{2} - c$, thus confirming that the guess is consistent with equilibrium.

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 probability $\lambda_2 > 0$, there is positive correlation between attending parties and young age. In turn, this softens the negative correlation between a and y, to an extent that makes it optimal for type 2 DMs to follow their taste. The expected welfare loss in this equilibrium is

$$\frac{1}{2} \cdot \lambda_2 \cdot (\frac{1}{2} - c) < \left(q - \frac{1}{2}\right) \frac{\lambda_2}{2 - \lambda_2}$$

The R.H.S of this inequality represents the maximal welfare loss in this setting. It increases with the fraction of type 2. There are two forces behind this observation. First, higher λ_2 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 or x, thus curbing the overall positive correlation between a and y that leads type 2 DMs to underestimate the health consequences of social distancing. The latter force is a beneficial "*equilibrium externality*" that the sophisticated DM type exerts on the naive type: A larger share of sophisticates implies that naifs commit a smaller error. Put differently, if public health authorities could somehow "educate" part of the population to reason better about causality, this would have a "*multiplier effect*" thanks to this equilibrium externality.

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 by adapting existing frameworks of equilibrium modeling with non-rational 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, C_i defines type i's information set, whereas D_i defines type i's "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.

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:

$$\begin{tikzcd}
a & t \arrow[l] \arrow[r] & y \\
& x \arrow[ul] \arrow[u] \arrow[ur] &
\end{tikzcd}$$

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 believes in a causal model that involving the variables on which he has data, and is given by the following DAG:

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

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). 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.

The reason I chose to present the model in a new language is twofold. First, this mode of exposition is relatively simple and self-contained, hence easier to follow for readers who may not know the previous frameworks. Second, by drawing a connection with the familiar notion of "bad controls", this paper will hopefully help inspiring new research about how everyday decision makers perform causal inference.

7 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 DM types differ in their sets of control variables. The main theme of the paper was that 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 poor controls.

The main general insight that emerged from this analysis was that the upper bound on this decision cost depends on whether DM types are differentiated "vertically" or "horizontally". In the former case, types can be partially ordered in some sense according to the size of their control variables. The equilibrium cost of bad controls is significantly lower than the non-equilibrium benchmark, and sometimes it completely vanishes. In the latter case, types control for different variables, which can give rise to mutually reinforcing confounding patterns, such that the maximal equilibrium decision cost is significantly higher than in the former case; sometimes it coincides with the non-equilibrium benchmark.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix: Proofs

Lemma 1

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 are empty, the procedure terminates after at most n steps. ■

Proposition 1

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. Since t is constant, this means that $y \perp a \mid x_{D_i}$ — i.e., $p(y = 1 \mid a, x_{D_i}) = p(y = 1 \mid x_{D_i})$. Formula (3) 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 (3) then implies $\Delta_i(x) = 0$. It follows that in equilibrium, type i plays $a = 0$ for all x. ■

Proposition 2

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

$x_1, x_2 \in \{0,1\}$, 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 $1 \in C_1 \backslash D_2$, $2 \in C_2 \backslash D_3$ and $3 \in C_3 \backslash D_1$. Suppose $\lambda_1, \lambda_2, \lambda_3 > 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Let $x_1, x_2, x_3 \in \{0,1\}$. 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 $\Delta_i(x_i = 1) \rightarrow 1$ as $\varepsilon \rightarrow 0$. 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 (3), we obtain

$$\Delta_i(x) = [p(t=1 \mid a=1, x_{C_i}) - p(t=1 \mid a=0, x_{C_i})][\delta_1 - \delta_0] \tag{7}$$

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

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 (7) can be written as

$$\frac{\gamma(x_{C_i})p(a \mid t=1, x_{C_i})}{\gamma(x_{C_i})p(a \mid t=1, x_{C_i}) + (1-\gamma(x_{C_i}))p(a \mid t=0, x_{C_i})} \tag{8}$$

Consider the terms $p(a \mid t, x_{C_1})$ in (8). Note that

$$p(a \mid t, x_{C_1}) = \sum_{x_{-C_1}} p(x_{-C_1} \mid t, x_{C_1})p(a \mid t, x_{C_1}, x_{-C_1}) \tag{9}$$

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, (9) 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 (8) is weakly increasing in a for $i = 1$. Since $\delta_1 - \delta_0 \geq 0$, $\Delta_1(x) \geq 0$. □

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 (8) 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

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

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 (7)-(8) 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. □

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

$$\sum_{x} p(x)(1-\gamma(x)) \cdot \sum_{x'} \beta(x',x)\gamma(x') \tag{10}$$

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 (10) 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

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

which can be written equivalently as

$$B_0(x,i) \leq \frac{\gamma(1-c)}{c(1-\gamma)} B_1(x,i) \tag{11}$$

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

$$\bar{B}_t(i) = \sum_{x \in X_i^*} p(x \mid t)p(a=1 \mid t,x) \tag{12}$$

Performing this summation over x_{C_i} on both sides of (11) implies

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

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

$$\max_i \bar{B}_0(i) \leq \frac{\gamma(1-c)}{c(1-\gamma)} \max_i \bar{B}_1(i) \tag{13}$$

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 (12), we can reformulate (13) 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, (13) can be rewritten as

$$\max_i \sum_{M|i\in M} \mu(M) \sum_{j\in M} \lambda_j^* \leq \frac{\gamma(1-c)}{c(1-\gamma)} \max_i \bar{B}_1(i) \tag{14}$$

This inequality is a necessary condition for the equilibrium welfare loss to exceed γ. □

Step 2: The following inequality holds:

$$\max_i \sum_{M|i\in M} \mu(M) \sum_{j\in M} \lambda_j^* \geq \left(\sum_M \mu(M) \sum_{j\in M} \lambda_j^* \right)^2 \tag{15}$$

Proof:[6] If we prove that

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

[6] This proof is due to Omer Tamuz.

then this will immediately imply (15) 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:

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

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 (16), 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 (12), β is a lower bound on $\Pr(a = 1 \mid t = 1)$. Therefore,

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

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

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

which by assumption exceeds γ. Rewriting this inequality as

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

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 $x_1, x_2 \in \{0, 1, \#\}$, and 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
$1-\gamma$	0	#	#
$\gamma - \beta - 2\beta^2$	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 $x_i \in \{0, 1\}$.[7]

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

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 $\Delta_1(x_1 = 0)$:

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

such that $\Delta_1(x_1 = 0) \approx -1$. It follows that $\Delta_1(x_1 = 1) > c$ and $\Delta_1(x_1 = 0) < -c$, 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 c 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 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 $c < 1$, 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. ■