Behavioral Causal Inference*

Ran Spiegler†

August 8, 2024

Abstract

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

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

†Tel Aviv University and University College London

1 Introduction

Learning causal effects from observational data is an important economic activity. Indeed, applied economists do it for a living. However, even lay decision-makers (DMs) 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. Will a college degree improve one's long-run economic prospects? Will wearing surgical masks on airplanes lower one's chances of catching a virus? Is coffee drinking good for one's health?

The way professional researchers and lay DMs practice causal inference differs in two major respects. First, researchers employ sophisticated inference methods that are subjected to stringent peer review. In contrast, lay DMs use intuitive, elementary methods, and face no pushback for doing so inappropriately. Second, while researchers are typically outside observers, lay DMs interact with the economic system in question; the aggregate behavior resulting from their causal inferences affects the correlations that inform these very inferences. For example, the inferences that parents make about the value of a college degree affect their children's educational choices, which in turn shape the correlational patterns that future parents rely on to evaluate college degrees. In both respects, it is apt to refer to the causal inferences that lay DMs engage in as "*behavioral*".

This paper is an attempt to model "behavioral causal inference", and analyze its welfare implications for DMs "in the field". I consider a DM who chooses a binary action a and tries to assess its effect on a payoff outcome y. The DM's payoff is $y - \theta \cdot \mathbf{1}[a \neq t]$, where t is a binary preference parameter that indicates the DM's favorite action, and θ is the cost the DM incurs when he does not take it. The DM will only do so if he thinks this has a beneficial causal effect on y. In the baseline model (presented in Section 2), I assume that the true effect is *null*. I do so to sharpen the exposition; later in the paper (in Section 6), I present a straightforward translation of my results to settings in which a has an additively separable, non-null causal effect on y.[1]

The DM forms his subjective causal belief by applying an intuitive causal-

[1] For a general discussion of this "placebo" methodology, see Spiegler (2024, Ch. 9).

inference procedure to long-run correlational data about actions, outcomes and a collection of exogenous variables (some of which are, or correlates of, the true causes of y). The data is generated by the behavior of other DMs in similar situations. The intuitive procedure is simple: Measuring the correlation between actions and outcomes, while *controlling* for some set of exogenous variables. This is a ubiquitous procedure in scientific data analysis, but it is basic enough for lay people to practice it (at least in simple form).

For example, when agents evaluate the protective benefit of wearing a surgical mask, some of them will only consider the correlation between mask wearing and infection rates, without controlling for any variable, and naively regard this correlation as causal. A somewhat savvier agent might restrict attention to infection statistics for people in his *own* age group (assuming he has access to such fine-grained data). In this case, age is the agent's single control variable. As the example suggests, agents may differ in what they feel a need to control for, as well as in their access to data about potential controls.

In general, suppose that long-run observational data is represented by a joint probability distribution p over actions, outcomes, and a collection of exogenous variables $x_1, ..., x_K$. The preference type t itself is not observable (but it may have arbitrary good proxies among the observable x variables). Think of p as describing frequencies in a large aggregate database that reflects the historical behavior of many DMs of various types. Part of what defines a DM's type is the set C of x variables he controls for (I assume that C is distributed independently of t). The DM estimates the causal effect of actions on outcomes according to the empirical average outcome given x_C and the action. This DM can err in both directions: His set of controls C may omit a relevant variable or include an irrelevant variable (Angrist and Pischke (2009), Cinelli et al. (2022)). Both errors can produce biased causal estimates. The bias can be large, if the exogenous variables x are strongly correlated with both a and y. This is why using bad controls is a grave error for empirical researchers.

However, in our setting, the correlation between x and a in the aggregate database given by p is *endogenous*, reflecting individual DMs' subjective optimization with respect to the causal belief they extract from p. Specifically,

the objective conditional distribution $p(a \mid t, x)$, given by the empirical frequencies in the database, describes the aggregate behavior of the DM population arising from the strategies of all DM types. In *equilibrium*, each type's strategy prescribes best-replies to his causal belief.

This description raises several questions. How large is the decision cost of flawed causal inference due to bad controls, when p is required to be consistent with equilibrium behavior? Does heterogeneity in DMs' sets of controls exacerbate or ameliorate this damage? Answering these questions will set theoretical benchmarks for the welfare implications of flawed causal inferences by agents "on the ground" (as opposed to researchers in their office), in real-life domains such as educational choice or preventive healthcare.

Example 1.1: Chess and math

To illustrate how equilibrium effects shape the welfare cost of flawed causal inference, consider the following example. A parent wishes to know whether his children can improve their school math performance by playing chess. The gross benefit from good math performance is 1, and the cost of playing chess against one's liking is $\theta < 1$. The fraction of children who like playing chess is $\gamma < \theta$. In reality, children who like chess are also good at math; but *ceteris paribus*, playing chess has *no* causal effect on math performance. The optimal strategy based on rational expectations is thus to let children play chess if and only if they like it.

Suppose a parent has access to long-run data resulting from this strategy. Then, he will observe a perfect correlation between chess playing and math school performance. The correlation is due to the confounding role of children's latent preferences. However, if the parent falsely regards it as causal, he will start pressuring his children to play chess against their preference. The expected welfare loss from this strategy can be arbitrarily close to 1 (if γ is small and θ is large — i.e., when most kids hate chess).

Over time, this parental behavior will weaken the observed long-run correlation between chess playing and math performance, thus eroding parents' belief in the magical effects of chess. In other words, more frequent chess playing dissipates its perceived causal effect. The system will reach a state of equilibrium when parents' choices constitute a best-reply to their causal beliefs. The equilibrium strategy administers reluctant chess playing with some intermediate probability, such that the estimated causal benefit from

playing chess is equal to its cost θ. As I will demonstrate formally in Section 4, the equilibrium expected welfare loss can never exceed $\gamma(1-\gamma)$. This bound, which is significantly below the non-equilibrium benchmark, reflects the negative relation between the frequency of chess playing and the strength of its estimated causal effect. The bound is tight; it can be approximated if θ is close to γ.

Thus, equilibrium effects can drastically reduce the welfare loss due to flawed causal inference. Yet, what happens when some parents in the population form causal estimates by controlling for various exogenous variables, which are proxies of the child's preferences (and objectively independent of math performance conditional on these preferences)? E.g., one can control for parents' scientific background or country of origin. How will heterogeneity among parents in terms of their sets of controls affect the equilibrium welfare loss?

There are two conflicting intuitions. On one hand, controlling for proxies of children's latent preferences brings parents closer to the ideal of shutting down preferences' confounding effect; this should curb causal-inference errors and shrink the welfare loss. On the other hand, by varying their behavior with these controls, parents create more elaborate confounding patterns in the data, giving fodder for more wrong causal inferences which exacerbate the equilibrium welfare loss.

It turns out that the maximal equilibrium welfare loss depends on the structure of the collection of sets of controls that the various types of parents employ. When these sets are ordered via set inclusion — such that parents are intuitively ranked in terms of their distance from the ideal of controlling for all potential confounders — the upper bound on the equilibrium welfare loss remains $\gamma(1-\gamma)$. In contrast, when the sets are not ordered via set inclusion, the upper bound is $\max\{\gamma, 1-\gamma\}$. This bound, too, is tight (as long as control variables can take at least three values). Thus, the welfare implications of flawed inferences about the causal effects of chess playing on math performance depend on whether DM types are "vertically" ordered. $\square$

The main results in the paper characterize tight upper bounds on the DM's expected equilibrium welfare loss, for various families of objective joint distributions over t, x, y. In Section 3 I consider environments with no preference heterogeneity (i.e., no variation in t), and impose no restriction on

the joint distribution of x and y. The characterization has a "bang-bang" flavor. As in Example 1.1, the bound depends on whether DM types' sets of controls contain one another. When they do, the equilibrium welfare loss is *zero* — i.e., equilibrium forces fully "protect" DMs from causal-inference errors. Otherwise, the tight upper bound coincides with the non-equilibrium benchmark. Section 4 examines environments with preference heterogeneity, mainly generalizing Example 1.1. It also shows that when DM types are not vertically ordered and there are no restrictions on the joint distribution over t, x, y, the tight upper bound on the DM's welfare loss coincides with the non-equilibrium benchmark.

The vertical ordering of DM types resonates with other areas of economic theory, which often make use of typologies that rank agents according to some linear ordering (ranking preference types by a willingness-to-pay scalar, ranking information types by the quality of their signal, etc.). Such typologies are attractive to economic theorists because they are interpretable and because they generate sharp results. The same holds in this paper. DMs with larger sets of control variables are intuitively closer to the ideal of correct causal inference. Therefore, in a naive sense, they are more "sophisticated" (though we will have opportunities to be reminded that adding controls is not necessarily beneficial). And as the analysis in Sections 3 and 4 shows, vertically ordered DM spaces generate strong results about the equilibrium decision costs of bad controls.

Section 5 extends the model by relaxing the assumption that DMs can condition their action on every variable they control for. For instance, in the surgical-mask example mentioned in the opening paragraph, a DM may have access to long-run statistical data about the prevalence of certain genes and their correlation with viral infection, without knowing his own genome. The DM can then use the data and to control for genetic variables without conditioning on them, by adjusting for their correlation with the variables he *can* condition on. A DM type in this extended environment consists of the set of variables he controls for and the subset of variables he conditions on. I generalize the analysis of the homogenous-preference case to this environment, via an appropriate extension of the notion of vertically ordered DM types.

The paper's main substantive message is that imposing equilibrium dis-

cipline on the database from which DMs draw flawed causal inferences can have significant welfare implications. Unlike bad empirical researchers in their office, DMs "in the field" are protected to some extent from causal-inference errors by equilibrium forces. However, this guarantee applies only when DM types can be ordered "vertically" in terms of how close they are to the ideal of neutralizing all confounding effects.

2 A Model

Let $a \in A = \{0, 1\}$ be an *action* that a decision maker (DM) chooses. Let $t \in \{0, 1\}$ be the DM's *preference type*. Let $y \in Y \subset [0, 1]$ be an *outcome*. Let $x = (x_1, ..., x_K)$ be a collection of *exogenous variables* that are realized jointly with t, prior to the realization of a and y. Let X_k denote the finite set of values that x_k can take. For every $M \subseteq \{1, ..., K\}$, denote $x_M = (x_k)_{k \in M}$ and $X_M = \times_{k \in M} X_k$. I assume that x and t are the only potential causes of y — i.e., a has *no causal effect* on y (Section 6 relaxes this assumption).

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

The DM's *data type* $i \in N = \{1, ..., n\}$ is drawn independently from a distribution $\lambda \in \Delta(N)$ (the independence assumption is immaterial for the results in Section 3 but plays a role in Section 4). Each type $i \in N$ is associated with a *distinct* subset $C_i \subseteq \{1, ..., K\}$. I refer to C_i as type i's set of *control variables*. A strategy for type (t, i) is a function $\sigma_{t,i} : X_{C_i} \to \Delta(A)$.

The interpretation of data types is as follows. Type i observes the realization of x_{C_i} prior to making his decision. He also has access to "public data" about the long-run joint distribution of x_{C_i}, a, y (I will introduce this distribution below). The DM believes that in order to learn the causal effect of a on y, he should control for these variables. As far as variables outside C_i are concerned, data type i lacks data on them or finds them irrelevant.

Note that t never belongs to the DM's set of control variables. The interpretation is that t is a *private* piece of information that never enters publicly available datasets. However, note that we can always allow one of

the variables x_i to be perfectly correlated with t; in this sense, the assumption does not rule out the possibility of effectively controlling for t. Also, the model does not admit variables that are caused by a or y as possible controls — it only focuses on exogenous, "pre-treatment" controls.

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

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

where $p(t, x)$ and $p(y \mid t, x)$ are exogenous, whereas $p(a \mid t, x)$ is endogenous, representing the DM's average behavior across data types:

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

Thus, the public data that the DM of any data type relies on is *aggregate*, representing the behavior of all data types. In what follows, I take it for granted that σ is implicit in p, without notating this explicitly.

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

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

If the DM had long-run data about all exogenous variables (including t), he could control for all of them, and thus correctly infer the action's null causal effect. This is the *rational-expectations benchmark* for this model. In contrast, our DM may end up believing that a has a non-null causal effect on y because he fails to control for some exogenous variables. In this case, he misinterprets part of the correlation between a and y as a causal effect, whereas in reality this correlation is entirely due to confounding by t, x.

What makes the model non-trivial is that these confounding patterns are *endogenous*, as they are affected by the strategy profile. Specifically, $E_p(y \mid a, x_{C_i})$ is not invariant to the profiles $\sigma_{t=0,-i}$ and $\sigma_{t=1,-i}$, which determine how a varies (in the aggregate data) with t and x_{-C_i}.

Expression (1) has the appearance of an expected-utility calculation by a standard Savage DM who receives a signal x_{C_i}. There is a fundamental difference, however, arising from the endogeneity of p and from its interpretation as an empirical frequency table from which the DM draws causal inferences. The DM makes a one-shot decision and approaches the data like a frequentist statistician; he is unaware that the data was partly generated by DMs of other types. Because this frequentist perspective departs from the Bayesian approach of the Savage model, I refrain from referring to x_{C_i} as a signal, and further discuss the connection to the Savage framework in Section 5.

Since controlling for all exogenous variables is an ideal of correct causal inference, it is natural to examine type spaces that rank DMs in terms of how far they are from this ideal.

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

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

The observation that $\Delta_i(x)$ is not invariant to σ suggests that to formalize subjective optimization by the DM, we need an equilibrium approach.

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

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

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

This definition captures a steady state in an underlying dynamic process. At every period, a new DM makes a one-shot decision after observing a large

sample of past realizations of t, x, a, y. The DM extracts a causal belief from this sample and best-replies to it, thus contributing a new data point to future DMs' samples. Equilibrium means that the statistical patterns of DMs' choices remain stable over time. Formalizing a dynamic convergence claim along these lines is outside the scope of this paper.

The trembling-hand aspect of the equilibrium concept means that the database the DM relies on involves a small element of blind experimentation by some data types. Technically, it ensures that all the conditional probabilities that are implicit in $E_p(y \mid a, x_{C_i})$ are well-defined (recall that these conditional probabilities are derived from the objective *joint* distribution p over all variables). We can therefore avoid a discussion of "off-path" beliefs, which would be alien to the strictly frequentist perspective of this paper. At any rate, trembles play a minor role in this paper. Their exact form is irrelevant for the upper-bound characterizations, with the single exception of Proposition 4.

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

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

$$\theta \sum_{t,x} p(t,x) \sum_{i \in N} \lambda_i \sigma_{t,i}(a \neq t \mid x) \tag{2}$$

My main analytical task in the next sections will be to derive *upper bounds* on this quantity when σ is required to be *an equilibrium*. Without this equilibrium condition, the upper bound is 1. To illustrate why, suppose that $t = 0$ with certainty, and that $x \in \{0, 1\}$. Assume $y = x$ with certainty for every x, and consider the strategy σ that prescribes $a = x$ with probability one. By definition, the probability of error is $p(x = 1)$. If $p(x = 1) \approx 1$ and $\theta \approx 1$, the welfare loss is approximately 1. However, the strategy σ is inconsistent with equilibrium. If a data type i varies his action with x, then he controls for it and correctly estimates the null causal effect of a. As a result, he will always play $a = 0$, contradicting the assumption that a varies with x in the aggregate data. It follows that the requirement that σ is an

equilibrium strategy can have bite.

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

3 Analysis: Homogenous Preferences

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

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

Example 3.1: Preventive healthcare
Let $K = 1$ and $n = 2$, such that $C_1 = \{1\}$ and $C_2 = \emptyset$. An economic story behind this scenario is that x, y and a represent age, a health outcome, and a costly (yet objectively useless) preventive healthcare decision. Type 1 is "sophisticated" in the sense of controlling for age when trying to infer the

preventive measure's causal health effect. Type 2 is "naive" in the sense of failing to control for any potential confounder.[2]

Since type 1 controls for x, he correctly estimates a null causal effect of a on y. This type plays $a = 0$ regardless of x — i.e., he ends up not varying his action with x. Type 2 potentially commits an error of causal inference because he fails to control for x, and interprets any empirical correlation between a and y as a causal effect. However, by definition, this type, too, does not vary his action with x. It follows that *none* of the two types vary their actions with x. If p is consistent with equilibrium, a and x must be statistically independent, thus destroying any possibility of x acting as a confounder of the relation between a and y. Yet, in the absence of confounding, failure to control for x is harmless. It follows that under the equilibrium restriction that $p(a \mid x)$ reflects data types' subjective optimization with respect to their causal beliefs, the DM incurs *no* welfare loss due to bad controls. □

The first result generalizes the example.

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

Thus, when $\gamma = 0$ and data types are vertically ordered, the equilibrium requirement fully "protects" the DM from choice errors due to bad controls. It does so by shutting down the channels through which the choice behavior of some data types could confound the statistical relation between actions and outcomes.

The results in this section are special cases of results reported in Section 5. Like all the results in this paper, they are proved in the Appendix. Here I make do with an informal sketch of the proof of Proposition 1, which is elementary in the special case covered by this section. It proceeds by induction on the set of data types. Type 1 effectively controls for all sources of correlation between a and y. Even when he fails to control for some exogenous variables, this does not matter because no other type conditions on them, hence they generate no confounding effect. As a result, type 1's

[2] Angrisiani et al. (2024) present an empirical study of false causal narratives regarding preventive measures in the context of Covid-19.

subjective best-reply is always correct — i.e., $a = 0$. Since type 1's strategy generates no variation in behavior, type 2 effectively controls for all potential confounders — which would not be the case if we did not impose the equilibrium condition on type 1's behavior. This equilibrium effect spreads down the set of data types, via the inductive argument.

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

Example 3.2: Analysts with diverse expertise
Let $K = 2$. All variables take values in $\{0, 1\}$, and their joint distribution satisfies:[3]

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

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

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

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

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

[3] These marginal and conditional distributions suffice for our purposes; there is no need for a full specification of p.

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

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

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

Second,

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

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

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

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

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

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

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

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

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

The following result generalizes this example.

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

Thus, when types are not vertically ordered, equilibrium forces do not curb the maximal welfare loss due to faulty causal inference. The reason is that the equilibrium behavior of different types can create confounding patterns that feed each other's inference errors.

4 Analysis: Heterogeneous Preferences

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

Example 4.1: Chess and math revisited
This is a formalization of the first part of Example 1.1. Let $y \in \{0, 1\}$, where $y = 1$ indicates high math school performance. Suppose $\delta_t = t$ — i.e., $y = 1$ if and only if $t = 1$ (which indicates that the child likes playing chess). Let $K = 0$ and $n = 1$ — i.e., there is a unique data type, $C = \emptyset$.

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

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

Denote $\alpha_t = \sigma_t(a = 1)$. By the DM's preferences, $\alpha_1 \geq \alpha_0$ in equilibrium. Now obtain explicit expressions for the terms that define Δ:

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

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

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

If $\alpha_0 \leq \varepsilon$ in ε-equilibrium, then $\Delta \to 1$ in the $\varepsilon \to 0$ limit. But then $\Delta > \theta$, hence playing $a = 1$ at $t = 0$ is the unique subjective best-reply, in contradiction to $\alpha_0 \leq \varepsilon$. It follows that $\alpha_0 > 0$ in equilibrium. There are two cases to consider. First, suppose $\alpha_0 \in (0, 1)$. This requires the DM to be indifferent between the two actions — i.e., $\Delta = \theta$. Therefore, and $\gamma < \theta$ and

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

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

equilibrium welfare loss is

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

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

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

Thus, for any configuration of θ and γ, there is a unique equilibrium in this setting. The DM's equilibrium welfare loss in this equilibrium is always weakly below $\gamma(1-\gamma)$. This bound can be approximated arbitrarily well by setting $\theta \approx \gamma$ (and if we set θ above γ, the equilibrium that approximates the upper bound does not rely on trembles).

As in Example 3.1, equilibrium forces in this example "protect" the DM from causal errors, by pushing his welfare loss far below the non-equilibrium benchmark. Let us elaborate on the intuition provided in the Introduction. The DM mistakes the correlation between a and y for a causal effect. This correlation is large when a varies strongly with t; it hits the maximal level when a always coincides with t. However, that extreme case is precisely when the DM commits *no* error. At the other extreme, if the DM almost always plays $a = 1$ because his estimated causal effect of a on y is above θ, the frequency of the DM's error is maximal. However, since in this case a varies little with y, the estimated causal effect is small. In general, a larger estimated causal effect goes hand in hand with a lower equilibrium frequency of making a decision error. This is why equilibrium effects limit the expected cost of failing to control for x. □

Let us now turn to a characterization of the upper bound on the DM's equilibrium welfare loss for any value of K, for a restricted domain of data-generating processes. Specifically, I assume that $p(y \mid t, x) \equiv p(y \mid t)$ — i.e., $y \perp x \mid t$. This fits situations in which the DM's preference type is a sufficient statistic for determining the outcome; the x variables are merely observable correlates of this statistic. For instance, whether a student enjoys studying determines his school performance. This latent attitude may be correlated

with observable characteristics, but these are only indirect causes, or proxies for the true cause.

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

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

As with Proposition 1, the proof of Proposition 3 proceeds by induction on the set of data types, starting with type 1, whose set of controls is the largest. Although this type controls for every x variable the other data types condition on, this does not mean he is immune to neglecting confounders, because he cannot control for t. Furthermore, since this type varies his behavior with t, he exerts a "confounding externality" (of the kind we encountered in Example 3.2) on the other data types. This makes the inductive proof considerably more intricate. A key argument in the proof is that while data types may disagree on the magnitude of the causal effect of a on y, they all agree on its *sign*, which is positive (since $\delta_1 > \delta_0$). The proof establishes that this is a feature of any equilibrium when the type space is vertically ordered.

Example 4.2: Chess and math with controls
Enrich Example 4.1 by letting $K = 1$ and $n = 2$, such that $C_1 = \{1\}$ and $C_2 = \emptyset$. The exogenous variable is $x \in \{0, 1\}$. This is an observable proxy for t; its conditional distribution is $p(x = t \mid t) = q \in (\frac{1}{2}, 1)$ for every t. In the context of the chess-and-math story, x may represent the parent's scientific background. The "sophisticated" type 1 controls for x when estimating the causal effect of chess playing on math performance. The "naive" type 2 either lacks access to data about x or finds it irrelevant.

I make two observations about this specification.

Observation 1: Both data types commit errors in equilibrium.
It is natural to expect that the sophisticated type's ability to control for x might protect him from the mistake of playing $a \neq t$. However, this cannot

be the case in equilibrium. To see why, recall that all types always estimate a non-negative causal effect of a on y. Therefore, the DM always plays $a = 1$ when $t = 1$ in equilibrium. It follows that $p(t = 1 \mid a = 0, x) = 0$ for every x. Denote

$$\alpha(x) = \sum_i \lambda_i \sigma_{t=0,i}(a = 1 \mid x)$$

This is the (aggregate) probability that the DM plays $a = 1$ at $t = 0$ and x. Since $y \equiv t$ by assumption, the sophisticated type's estimated causal effect of a on y given x is

$$\Delta_1(x) = p(t = 1 \mid a = 1, x) = \frac{\gamma p(x \mid t = 1)}{\gamma p(x \mid t = 1) + (1 - \gamma) p(x \mid t = 0)\alpha(x)}$$

while the naive type's estimated causal effect is

$$\Delta_2 = p(t = 1 \mid a = 1) = \frac{\gamma}{\gamma + (1 - \gamma) \sum_{x'} p(x \mid t = 0)\alpha(x')}$$

Suppose both types always play $a = t$. Then, since the sophisticated type does not vary his action with x, the distinction between the two types disappears. This sends us back to Example 4.1, where we saw that $a \neq t$ with positive probability in equilibrium, a contradiction. It follows that in equilibrium, the DM must play $a = 1$ with positive probability when $t = 0$. The DM will do so when his estimated causal effect of a on y is weakly above θ. Crucially, Δ_2 is a *weighted average* of the $\Delta_1(x)$'s. Moreover, it can be checked that $\Delta_1(x)$ is not constant in x because $q > \frac{1}{2}$. Therefore, $\max_x \Delta_1(x) > \Delta_2 \geq \theta$, such that the sophisticated type must mistakenly play $a = 1$ when $t = 0$ for some x. No matter how accurate x is as a proxy of t, it cannot fully protect the sophisticated type from his failure to control for t, when equilibrium effects are taken into account.

Is it possible that the naive type never commits an error? If so, then we must have $\Delta_2 \leq \theta$. But since Δ_2 is a non-trivial weighted average of the $\Delta_1(x)$'s, there is x for which $\Delta_1(x) < \Delta_2 \leq \theta$, such that the sophisticated DM plays $a = 0$ when $t = 0$ at x. But then $\alpha(x) = 0$, which implies $\Delta_1(x) = 1 > \theta$, a contradiction.

For the following observation, fix $\gamma = \lambda = \frac{1}{2}$, to simplify calculations.

Observation 2: Welfare comparison with Example 4.1
Suppose that

$$\frac{1}{2} < \theta < \frac{2}{4-q}$$

Then under the conditions of Example 4.1, where the DM is naive, he plays $a = 1$ at $t = 0$ with probability strictly below 1 (since $\theta > \gamma$). In contrast, in the present example, the following profile is an equilibrium if q is not too close to $\frac{1}{2}$: The naive type always plays $a = 1$, while the sophisticated type plays $a = 1$ if and only if $t = 1$ or $x = 1$. Thus, the presence of the sophisticated type increases the equilibrium probability that the naive type commits an error. The reason is that by varying his action with t, the sophisticate amplifies the confounding effect of t and therefore exacerbates the naif's causal error.

Nevertheless, the DM's ex-ante welfare loss in this equilibrium is below its level in Example 4.1. To see why, recall that since $\theta > \gamma = \frac{1}{2}$, the DM's equilibrium welfare loss in Example 4.1 is $\frac{1}{2}(1-\theta)$. In contrast, the above-constructed equilibrium in the present example generates a welfare loss of

$$\theta \cdot [\lambda_1 \cdot (1-\gamma) \cdot (1-q) + \lambda_1 \cdot (1-\gamma)] = \frac{\theta(2-q)}{4}$$

which is below $\frac{1}{2}(1-\theta)$, by the range restriction on θ. Thus, while the sophisticate exerts a negative externality on the naif, his presence makes the DM better off on average. Also observe that by the range restriction, the welfare loss is below

$$\frac{2(2-q)}{4(4-q)} < \frac{1}{4}$$

where the R.H.S is the upper bound obtained in Proposition 3 for $\gamma = \frac{1}{2}$. □

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

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

This result carries the relevance of vertical ordering of data types to the heterogeneous-preferences setting. The gap between the upper bounds in the two cases — $\gamma(1-\gamma)$ vs. $\max(\gamma, 1-\gamma)$ — is significant, and gets wider as the preference type distribution becomes more unbalanced. To attain the upper bound given by Proposition 4, I use suitable 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. Unlike the case of vertically ordered types, different data types may disagree on the causal effect's *sign*; indeed, this feature plays a key role in my implementation of the upper bound.

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

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

The results in this section leave two open problems. First, does the upper bound $\gamma(1-\gamma)$ obtained for vertically ordered types extend to distributions p for which $y \not\perp x \mid t$? Second, how do results change when the distribution over data types is allowed to be correlated with t and x?

5 Controlling without Conditioning

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

However, in some cases it makes sense to assume that the DM has access to statistical data about variables, without knowing their realization at the moment of choice. For example, a firm may know how its performance is correlated with macroeconomic indicators, yet it need not know their current

value when making its business decisions because the indicators are published with delay. In such cases, the DM can still control for such variables, even when he cannot condition on their realization. I refer to this mode of controlling as *adjustment* as opposed to conditioning.

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

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

$$\Delta_i(x) = \sum_{x_{D\setminus C}} p(x_{D\setminus C} \mid x_C) \left[E_p(y \mid a = 1, x_D) - E_p(y \mid a = 0, x_D)\right] \quad (3)$$

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

Subjective state spaces

The perceived causal effect given by (3) can be interpreted traditionally in terms of the Savage framework, where the state space itself is *subjective*. According to this interpretation, X_{D_i} is type i's subjective state space and X_{C_i} is his set of signals. The novelty here is that while the state space is subjective, the DM's belief is a projection of the *objective* distribution p on his subjective state space. Moreover, unlike the standard Savage model, the stochastic mapping from the DM's subjective states to outcomes is affected by the behavior of other DM types, hence it is an endogenous object. In my opinion, these deviations from the Savage framework are so drastic that

they justify my decision to avoid the Savage terminology altogether in the paper's formal exposition.

Example 5.1: Adjusting for an irrelevant variable
This example illustrates the danger of excessive controlling for "pre-treatment" variables, independently of equilibrium considerations. It is adapted from Cinelli et al. (2022), a guide to "good and bad controls" that, following Pearl (2009), makes use of the formalism of directed acyclic graphs (DAGs). Let $t = 0$ with certainty, and suppose that the true causal structure underlying p is given by the DAG

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

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

Suppose, however, that one of the DM types has $C = \emptyset$ and $D = \{3\}$ — i.e., he does not condition on any variable, while adjusting for x_3.[4] The type's estimated causal effect is

$$\sum_{x_3} p(x_3)[E_p(y \mid a = 1, x_3) - E_p(y \mid a = 0, x_3)] \tag{4}$$

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

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

Plugging these values in (4), we obtain a non-null estimated causal effect. The intuition is as follows. Because x_3 is a common consequence of x_1 and x_2 (which are correlated with a and y, respectively), it is not necessarily true

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

that $a \perp y \mid x_3$. Therefore, x_3 is a bad control that produces a biased causal estimate. $\square$

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

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

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

I now extend the notion of vertically ordered types. Define a binary relation P over data types: iPj if $D_i \supseteq C_j$. The meaning of iPj is that data type i controls for every variable that type j conditions on. Since $D_i \supseteq C_i$ for every $i \in N$, P is reflexive. Let P^* be the asymmetric (strict) part of P — i.e., iP^*j if iPj and $j \not P i$. Following Sen (1969), P is *quasitransitive* if P^* is transitive.

Definition 5 *The set N is vertically ordered if the binary relation P is complete and quasitransitive.*

When $C_i = D_i$ for every $i \in N$, this definition collapses to Definition 1.

The following observation is standard. We say that type i is P^*-undominated in a set of types M, if there is no $j \in M$ such that jP^*i.

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

The partition induced by a complete and quasitransitive P is the extended model's analogue of vertical ordering of types.

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

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

Proposition 7 *Let $\gamma = 0$. Suppose N is not vertically ordered. Then, for any $\theta, \beta \in (0, 1)$, there exist λ and $(p(x, y))$ such that $\Pr(a = 1) > \beta$ in some equilibrium. In particular, when $\theta \approx 1$, the equilibrium welfare loss can be arbitrarily close to 1.*

The proof of Proposition 6 is by induction on the partition induced by P. The reasoning is essentially the same as provided by the informal sketch for Proposition 1.

Proposition 7 shows the other side of the "bang-bang" characterization. When N is vertically unordered, the equilibrium requirement does not constrain the maximal possible welfare loss due to bad controls. The proof is constructive, involving more elaborate versions of Example 3.2. In particular, when P is complete but not quasitransitive, the construction involves *three* data types.

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

6 Consequential Actions

So far, I have focused on the extreme case in which the DM's action has no objective causal effect on the outcome. This facilitated the definition of the DM's equilibrium welfare loss due to poor controls, relative to the rational-expectations benchmark. This section extends the analysis to situations in which actions have an additively separable causal effect on outcomes. This kind of separability is the bread and butter of observational research, with its common reliance on linear-regression models. Therefore, it is sensible to assume it in our context as well.

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

$$E_p(y \mid a,z) = \beta a + (1-\beta)z \tag{5}$$

where $\beta \in (0,1)$ quantifies the true causal effect of a on y. The DM is unaware of the relation (5), nor does he know β. As in the basic model, he forms beliefs by examining the observed joint distribution of a, y, and his set of control x variables. Nevertheless, we can express type i's estimated causal effect of switching from $a=0$ to $a=1$ on y given x as

$$\beta + (1-\beta)\Delta_i^z(x)$$

where

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

This expression makes use of the extended notion of DM types presented in Section 5, which allows D_i to be a strict superset of 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 previous sections.

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

$$\gamma \cdot p(a=0 \mid t=1) \cdot (\theta+\beta) + (1-\gamma) \cdot p(a=1 \mid t=0) \cdot (\theta-\beta) \tag{6}$$

The DM chooses $a=0$ at $(t=1,x)$ only if $\theta+\beta \leq -(1-\beta)\Delta_i^z(x)$. Likewise, he chooses $a=1$ at $(t=0,x)$ only if $\theta-\beta \leq (1-\beta)\Delta_i^z(x)$. Consequently, by (6), the upper bounds on the DM's equilibrium welfare loss are the same as in Sections 3-4, multiplied by $1-\beta$.

7 Related Literature

This paper continues my line of research into the behavioral implications of flawed causal reasoning (Spiegler 2016,2020). The problem it poses — quantifying the decision costs of faulty causal inference — is novel and lacks precedents in the literature. The paper introduces another novelty. Prior work has focused on DMs who misperceive the causal mapping from actions to consequences, but fully account for the determinants of actions. In contrast, the DM in this paper fails to perceive that actions have direct causes that confound the observed relation between actions and consequences.[5] Moreover, DM types *differ* in their understanding of these causes, via their different sets of controls. This heterogeneity is what makes the problem of quantifying the decision costs of bad controls non-trivial.

More broadly, the paper is part of the program of developing equilibrium modeling frameworks in which agents have non-rational expectations that systematically distort empirical regularities.[6] I will now show how existing frameworks can be adapted to incorporate the novel features in this paper. To make the comparison complete, I make use of the extended formalism of Section 5.

Analogy-based expectations

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

[5] Clyde (2023) effectively shares this feature, by assuming that the DM forms equilibrium beliefs on the basis of data about *proxies* of relevant variables (including actions).

[6] A growing literature focuses not on equilibrium behavior but on learning dynamics under misspecfied beliefs — e.g., Heidhues et al. (2018), Esponda et al. (2021), Fudenberg et al. (2021), Bohren and Hauser (2021), Frick et al. (2023), Ba (2024).

Jehiel relies on the sequential-equilibrium conceptual baggage.)

Bayesian networks

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

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

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

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

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

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

The DM's perceived causal effect of a on y given x_{C_i} is thus given by (3). Equilibrium in the present model is consistent with the notion of personal equilibrium in Spiegler (2016), with the modification that the DM's subjective causal model itself is random.[7]

Berk-Nash equilibrium

The Bayesian-network framework can be subsumed into the more general concept of Berk-Nash equilibrium (Esponda and Pouzo (2016)). According to this concept, a misspecified subjective model is represented by a set of conditional distributions (mapping from signals and actions to outcomes). The DM best-replies to a belief in this set that minimizes a weighted version

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

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, it is relatively simple and self-contained. Second, it does not require readers to absorb modified versions of previous (and possibly unfamiliar) frameworks. Third, by drawing a connection with the familiar and intuitive notion of "bad controls" and the work habits of empirical researchers, this paper will hopefully help inspiring new research about how everyday DMs perform causal inference.

DMs who form wrong beliefs because they ignore confounding effects (involving variables other than the DM's action) appear in various examples in Spiegler (2016,2020). Confounder neglect is related to the error of drawing inferences from selective datasets that fail to internalize their selectiveness. In Esponda (2008), buyers infer product quality from a sample of observed trades, without realizing that it results from sellers' adversely selective response to market prices. In Jehiel (2018), entrepreneurs evaluate investments based on a dataset of implemented projects, without realizing they result from selectively positive signals.

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

8 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 common error, when DMs rely on endogenous datasets and may differ in their sets of control variables. Since DMs' causal inferences determine how they vary their actions with exogenous variables, and since this response in turn shapes the very correlations from which DMs draw their inferences, equilibrium analysis is required to evaluate the decision cost of erroneous causal inference due to bad controls.

The general insight that emerged from this analysis was that when DM types are "vertically" differentiated in terms of the sets of controls, the equilibrium cost of bad controls falls significantly below the non-equilibrium benchmark; sometimes it completely vanishes. I substantiated the role of vertical differentiation by showing that the upper bound on the welfare loss is significantly higher when types are not vertically ordered; sometimes it coincides with the non-equilibrium benchmark. Of course, worst-case analyses have a built-in limitation: The worse the worst case gets, the less useful it is. From this point of view, the results on vertically ordered types are more meaningful, and the role of the other results is to put them in perspective.

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

References

[1] Angrisani, M., A. Samek, and R. Serrano-Padial (2024). Competing Narratives in Action: An Empirical Analysis of Model Adoption Dynamics. NBER working paper no. w32242.

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

[3] Ba, C. (2024). Robust Misspecified Models and Paradigm Shifts. Working paper.

[4] Bohren, A. and D. Hauser (2021). Learning with Heterogeneous Misspecified Models: Characterization and Robustness. Econometrica 89, 3025–3077.

[5] Clyde, A. (2023). Proxy Variables and Feedback Effects in Decision Making. Working paper.

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

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

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

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

[10] Esponda, I. (2008). Behavioral Equilibrium in Economies with Adverse Selection. American Economic Review 98, 1269-1291.

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

[12] Esponda, I., D. Pouzo, and Y. Yamamoto (2021). Asymptotic Behavior of Bayesian Learners with Misspecified Models. Journal of Economic Theory 195, 105260.

[13] Frick, M., R. Iijima, and Y. Ishii (2023): Belief Convergence under Misspecified Learning: A Martingale Approach. Review of Economic Studies 90, 781–814.

[14] Fudenberg, D., G. Lanzani, and P. Strack (2021). Limit Points of Endogenous Misspecified Learning. Econometrica 89, 1065–1098.

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

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

[17] Heidhues, P., B. Koszegi, and P. Strack (2018). Unrealistic Expectations and Misguided Learning. Econometrica 86, 1159–1214.

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

[19] Jehiel, P. (2018). Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism. American Economic Review 108, 1582-1597.

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

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

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

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

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

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

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

[27] Spiegler, R. (2024). The Curious Culture of Economic Theory. Cambridge: MIT Press.

Appendix: Proofs

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

Proposition 6

I will show that $a = 0$ with probability one in equilibrium. The proof is by induction with respect to the partition induced by P. Consider an arbitrary type i in the top layer N_1. This type satisfies $D_i \supseteq C_j$ for all $j \in N$. Hence, there is no x variable outside D_i that *any* DM type conditions his action on. Since t is constant, this means that $y \perp a \mid x_{D_i}$ — i.e., $p(y \mid a, x_{D_i}) = p(y \mid x_{D_i})$. Formula (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 7

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

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

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

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

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

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

and

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

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

Proposition 3

For every x and every $C \subseteq \{1, ..., K\}$, denote $\gamma(x) = p(t = 1 \mid x)$ and $\gamma(x_C) = p(t = 1 \mid x_C)$. By assumption, $C_1 \supset \cdots \supset C_n$. The proof proceeds stepwise.

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 (1), we obtain

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

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

Step 2: For every x, $\Delta_1(x) \geq 0$ and $\sigma_{t=1,1}(a = 1 \mid x_{C_1}) = 1$.
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

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

By the DM's preferences, $\sigma_{t=1,i}(a = 1 \mid x_{C_i}) \geq \sigma_{t=0,i}(a = 1 \mid x_{C_i})$ 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 $\delta_1 - \delta_0 \geq 0$, $\Delta_1(x) \geq 0$. $\square$

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

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

By the inductive step,

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

for every $j \leq m$. By definition, $C_j \subseteq C_i$ for every $j > m$, hence $\sigma_{t,j}(a \mid x_{C_j})$

is constant in x_{-C_i}. Therefore,

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

We already observed that $\sigma_{t=1,j}(a = 1 \mid x_{C_j}) \geq \sigma_{t=0,j}(a = 1 \mid x_{C_j})$ for every x_{C_j}. It follows that

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

As in the proof of Step 2, applying this inequality to (8) implies that $\Delta_i(x) \geq 0$ and $\sigma_{t=1,i}(a = 1 \mid x_{C_i}) = 1$. 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 $p(a = 1 \mid t, x) = \sum_i \lambda_i \sigma_{t=0,i}(a = 1 \mid x_{C_i})$ *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 $\sigma_{t=0,j}(a = 1 \mid x_{C_j}) > 0$.

Then, the DM's expected welfare loss given x is

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

In order for type $i(x)$ to play $a=1$ given x and $t=0$, it must be the case that $\theta\leq\Delta_{i(x)}(x)$. By Step 3, $\sigma_{t=1,j}(a=1\mid x_{C_j})=1$ 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 $\sigma_{t=0,j}(a=1\mid x_{C_j})>0$, it follows that none of these types j condition on $x_{-C_{i(x)}}$. Therefore,

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

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

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

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

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

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

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

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

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

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 $\sigma_{t=0,i}(a = 1 \mid x) > 0$. Denote

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

Define

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

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

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

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

which can be written equivalently as

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

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

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

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

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

Note that

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

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

$$\lambda_j^* = \lambda_j \sum_x p(x \mid t = 0, x \in X_j^*) \sigma_{t=0,j}(a = 1 \mid x)$$

Then, (13) can be rewritten as

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

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

Step 2: The following inequality holds:

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

Proof:[8] If we prove that

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

then this will immediately imply (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,

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

[8] This proof is due to Omer Tamuz.

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

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

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

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

which by assumption exceeds γ. Rewriting this inequality as

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

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

(ii) Implementing the upper bound

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

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

where β is arbitrarily small. 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 $x_i \in \{0, 1\}$.[9] In addition, suppose each type i plays $a = 0$ with probability $1 - \varepsilon$ when $x_i = \#$, where ε is arbitrarily small. Let us calculate the terms in $\Delta_1(x_1 = 1)$:

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

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

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

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

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

Proposition 5

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

[9] 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 $x_i = 0, 1$.

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

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

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

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

Second,

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

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

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