Behavioral Causal Inference*

Ran Spiegler†

March 29, 2025

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 endogenous 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 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 at various journals, 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. 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 everyday decision-making. Motivated by the ubiquity of binary treatments in causal-inference studies (in both social and medical sciences), I consider a DM who chooses a binary action a and tries to assess its effect on a gross payoff outcome y. The DM incurs a cost of θ whenever $a \neq t$, where t is a random variable that indicates the DM's favorite action. The DM will take the unfavorite action only if he thinks this has a beneficial causal effect on y that offsets the cost. In the baseline model, I assume that the true causal effect is *null*, to sharpen the exposition.

The DM forms his subjective causal belief by applying an intuitive causal-inference procedure to long-run correlational data about actions, outcomes and a collection of exogenous variables — some of which are (or correlated with) the true causes of y. Specifically, he measures the correlation between

actions and outcomes, while *controlling* for some set of exogenous variables. This is a common procedure in scientific data analysis, but it is basic enough for lay people to practice it (at least in simple form).

For instance, when agents evaluate surgical masks' protective benefits, they may regard the raw correlation between mask wearing and infection rates as causal. Savvier agents may restrict attention to infection statistics for people in their *own* age group (if they has access to such fine-grained data). In this case, age is their control variable. Likewise, parents who consider whether to send their child to a private school may regard the raw correlation between school choice and college-admission outcomes as causal. Savvier parents may restrict attention to admission statistics for candidates with the same parental education (if they have access to such data). In this case, parental education is the control variable. As these examples suggest, DMs may differ in what they feel a need to control for, or in their access to data about potential controls.

In general, I represent long-run observational data by a joint probability distribution p over actions, outcomes, and exogenous variables $x_1, ..., x_K$. The DM's preference type t itself is not observable (but potentially correlated with x). Think of p as describing frequencies in a large database that records the historical behavior of many DMs of various types. Part of what defines a DM's type in my model is the set C of x variables he controls for (C is distributed independently of t). This reflects the idea that DMs may differ in their access to the historical database, or in the exogenous variables they deem relevant.

A DM of type C estimates the causal effect of actions on outcomes according to the empirical average of y given x_C and a. This DM can err in both directions: His set of controls C may omit a relevant exogenous variable or include an irrelevant one (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 that p represents is *endogenous*, reflecting individual DMs' subjective optimization with respect to the causal belief they extract from p. Specifically, the objective conditional distribution $p(a \mid t, x)$, which de-

scribes empirical action frequencies in the database, describes the aggregate behavior of the DM population arising from the strategies of all DM types. If agents commit causal-inference errors, this affects their behavior and therefore — unlike the case of empirical researchers — the data from which other agents extract causal lessons. This two-way relation between DMs' actions and causal beliefs calls for an *equilibrium* analytical approach. In equilibrium, each type's strategy (as given by p) prescribes best-replies to the causal belief he extracts from p using his set of controls. This equilibrium requirement turns out to have non-trivial implications for the decision costs of flawed causal inference in the form of bad controls, in a variety of economic contexts that involve health, education, lifestyle or business decisions.

Example 1.1: Chess and math

A parent wishes to know whether his children can improve their school math performance by playing chess. The gross benefit from high 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 like chess if and only if they are 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 our parent has long-run data resulting from previous parents who followed this strategy. Then, he will observe a perfect correlation between chess playing and math school performance. The correlation is due to confounding by children's latent preferences. However, if the parent falsely interprets the correlation as causal, he will administer chess to his children against their taste. The expected welfare loss from this misguided strategy is $\theta(1-\gamma)$. As successive generations of parents adhere to this strategy, the observed long-run correlation between chess playing and math performance gradually becomes weaker, since children who are bad at math now also play chess. Over time, this trend erodes parents' belief in the magical power of chess: More frequent chess playing dissipates its perceived causal effect.[1]

This dynamical process reaches an equilibrium when parents' stable-over-time choices constitute a best-reply to stable-over-time causal beliefs, which

[1] While this intuitive dynamic process motivates my equilibrium notion, it remains in the background in the formal exposition. It could be formalized to provide a rigorous learning foundation for equilibrium in this example. Generality of this foundation is an open problem.

are extracted from long-run data. In Section 4, I show formally that the unique equilibrium strategy administers chess to children who dislike it with some intermediate probability, such that the estimated causal benefit from playing chess equals the cost θ. The higher this cost, the lower the frequency of paying chess against one's liking. The equilibrium expected welfare loss is $\gamma(1-\theta)$, which is always below $\gamma(1-\gamma)$ (the same upper bound applies when $\gamma > \theta$). This is significantly below the non-equilibrium benchmark. In particular, when γ is small and $\theta \approx 1$ (i.e., when most kids really dislike chess) the non-equilibrium loss is approximately 1 while the equilibrium loss is close to 0.

Thus, equilibrium effects can drastically reduce the welfare loss due to naive causal interpretation of the raw action-consequence correlation. But now suppose that some parents in the population form causal estimates by controlling for various exogenous correlates of children's taste (which continues to be the sole true cause of math performance). For instance, 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 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 different sets of controls, parents create more elaborate confounding patterns in the data, giving fodder for more wrong causal inferences that can exacerbate the equilibrium welfare loss. It turns out that the maximal equilibrium welfare loss depends on the structure of the collection of parents' sets of controls. When these sets are nested via set inclusion, the upper bound on the welfare loss remains $\gamma(1-\gamma)$. In contrast, when the sets are not ordered by set inclusion, the tight upper bound is $\max\{\gamma, 1-\gamma\}$, which is significantly greater. $\square$

The main results in the paper characterize tight upper bounds on the DM's expected equilibrium welfare loss for various environments. Section 3 considers homogenous preferences (t is constant), imposing no restrictions 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, essentially formalizing 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 bound on the equilibrium welfare loss coincides with the non-equilibrium benchmark.

Thus, whether data types are "vertically" ordered by set inclusion is crucial for the worst-case analysis of the decision costs of bad controls. In applications, vertical ordering is plausible when all DMs are interested in the same variables, yet differ in their quality of data access: Some variables are available to everyone, while others are only available to agents with "premium access" — with possibly multiple degrees of exclusive access. In contrast, "horizontally" differentiated type spaces (with different types using different controls) are more appropriate for environments in which different DM types obtain their data from separate social-media echo chambers, or from experts in different fields who focus on different aspects of reality.

In addition to the upper-bound results, Sections 3 and 4 illustrate the model and its potential for economic applications, using examples that invoke various economic scenarios and involve natural parameterizations, rather than ones designed specifically for attaining the upper bounds. Later sections extend my analysis in two separate directions. Section 5 generalizes the model by relaxing the assumption that DMs can condition their action on every variable they control for. Section 6 presents a straightforward translation of all the results in the paper to settings in which a has an additively separable, non-null causal effect on y. I apply this extended model to a life-style decision problem, and use it to illustrate the non-trivial externality that sophisticated DMs exert on naifs.

The paper's main message is that the distinction between exogenous and endogenous datasets matters greatly for evaluating the cost of poor causal inference. Specifically, when we impose equilibrium discipline on the database from which DMs draw causal inferences, equilibrium forces can protect DMs from inference errors due to bad controls. However, from the point of view of worst-case analysis, the magnitude of this "equilibrium protection" is much greater when DMs' sets of controls are nested.

2 A Model

A decision-maker (DM) chooses an *action* $a \in A = \{0,1\}$. 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 be 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 $a = t$; this would be the DM's *rational* choice. The DM's *data type* $i \in N = \{1, ..., n\}$ is drawn independently from a distribution $\lambda \in \Delta(N)$.[2] 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 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 is unaware of them, lacks data about them, or dismisses them as being irrelevant. Note that t never belongs to the DM's set of control variables. The interpretation is that t is *private* information that never enters public datasets. However, note that the x variables, about which public data *is* available, can be highly (or even perfectly) correlated with 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 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 historical 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

[2] The independence assumption is immaterial for the results in Section 3 but plays a role in Section 4.

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 $p(t, x, a, y) = p(t, x)p(a \mid t, x)p(y \mid t, x)$, where $p(t, x)$ and $p(y \mid t, x)$ are exogenous, while $p(a \mid t, x)$ 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 a DM of any type relies on is *aggregate*, representing the endogenous choices of *all* 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. 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: $E_p(y \mid a, x_{C_i})$ is not invariant to the strategy profile σ_{-i}, which determines how a varies (in the aggregate data) with t and x_{-C_i}. Note that since C_i never includes the latent variable t, formula (1) does *not* include the rational benchmark as a special case (except when C_i happens to include a variable that is perfectly correlated with t).

Example 2.1: Illustrating formula (1)
Let $K = 1$ — i.e., there is a single potential control variable x. Suppose $x \in \{0, 1\}$ and $p(x = t \mid t) = q \in (\frac{1}{2}, 1)$ for every t. That is, x is a proxy of t whose precision is given by q. Suppose also that $p(y = t \mid t) = 1$ for every t. Let $n = 2$, such that $C_1 = \{1\}$ and $C_2 = \emptyset$. That is, the "sophisticated"

type 1 controls for x, whereas the "naive" type 2 either lacks access to data about x or finds it irrelevant. Suppose the DM's strategy σ prescribes $a = 1$ with certainty when $t = 1$. The probability that each DM type plays $a = 1$ at $t = 0$ (as a function of x, in the case of type 1) is denoted $\alpha_1(x) = \sigma_{t=0,i=1}(a = 1 \mid x)$ and $\alpha_2 = \sigma_{t=0,i=2}(a)$.

Let us now calculate $\Delta_1(x)$ and Δ_2. Recall that $p(y = 1 \mid a, x) \equiv p(t = 1 \mid a, x)$. Since all types play $a = 1$ whenever $t = 1$, $p(t = 1 \mid a = 0, x) = 0$ for every x. It follows that

$$\Delta_1(1) = p(t = 1 \mid a = 1, x = 1) = \frac{\gamma q}{\gamma q + (1-\gamma)(1-q)[\lambda_1\alpha_1(1) + \lambda_2\alpha_2)]}$$

$$\Delta_1(0) = p(t = 1 \mid a = 1, x = 0) = \frac{\gamma(1-q)}{\gamma(1-q) + (1-\gamma)q[\lambda_1\alpha_1(0) + \lambda_2\alpha_2]}$$

$$\Delta_2 = p(t = 1 \mid a = 1) = \frac{\gamma}{\gamma + (1-\gamma)[\lambda_1(q\alpha_1(0) + (1-q)\alpha_1(1)) + \lambda_2\alpha_2]}$$

Note that by definition, Δ_2 is a weighted average of $\Delta_1(1)$ and $\Delta_1(0)$. Therefore, if type 2 weakly prefers to play $a \neq t$, so must type 1 with some probability. We will revisit these formulas in Example 4.2. □

As the example demonstrates, the DM's estimate of the causal effect of a can be sensitive to the strategy profile σ. This observation suggests that to define subjective optimization with respect to the causal belief one extracts from an endogenous database, an equilibrium approach is called for.

Definition 1 (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 approaches the data like a frequentist statistician. In particular, he does not take into consideration the fact that the data was partly generated by DMs of other types. He 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. Equilibrium existence can be established by standard arguments (see Section 7).

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 t, x, a, y). 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 5.

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 a null objective 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*

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

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.

Comment: The rationality benchmark. The rational benchmark for this model is a DM who controls for t and x — i.e., he has data about all potential confounders. This DM always plays $a = t$. What would be a "rational" mode of behavior for a DM with limited data, who is aware that there may be confounders beyond those he has data on? From this paper's strict frequentist perspective, there is no clear prescription. Consider a competent econometrician who knows that his dataset excludes certain confounders. Rather than using whatever control variables he has at his disposal, the econometrician would simply refrain from making a causal-inference claim, knowing that he lacks a credible causal-identification strategy. Our DM cannot afford to do so, because he has to make an active choice. Consequently, he proceeds with his available controls as if they suffice. This is part of what makes his causal inference "behavioral".[3]

3 Analysis: Homogenous Preferences

This section characterizes the maximal equilibrium welfare loss 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 $\theta \cdot \Pr(a = 1)$. 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.

[3] A Bayesian-rational approach would assume that the DM has a subjective prior belief over the data-generating process, which he updates according to the data. If the DM correctly believes that the mapping from (t, x, a) to y is constant in a, he will always play $a = t$, regardless of what he learns from the data.

Example 3.1: Preventive healthcare
Let $K = 1$ and $n = 2$, such that $C_1 = \{1\}$ and $C_2 = \emptyset$, as in Example 2.1. An economic story behind this scenario is that x, y and a represent age, a health outcome, and a choice whether to adopt a costly, yet objectively useless preventive healthcare measure. Type 1 has access to data about how age is correlated with the other variables, and can therefore control for age when estimating the preventive measure's causal health effect. Type 2 lacks the data, and therefore fails to control for the potential confounder.

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, then 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. It is based on a notion of vertical ordering of data types.

Definition 3 (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. In this case, data types are naturally ranked in terms of how close they are to the ideal of controlling for all potential confounders (since the DM never controls for t, no type attains this ideal).

As mentioned in the Introduction, vertical ordering fits situations in which all DM types agree on what they consider to be relevant controls; however, they have limited degrees of data access. Variables are ordered in

terms of how easy it is to get access to data about them; DM types can be viewed as rungs on this ladder. More broadly, vertical ordering of DM types resonates with other areas of economic theory that classify 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 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, they are intuitively more "sophisticated" (though we will have opportunities to be reminded that adding controls is not necessarily beneficial). And as our analysis will now begin to show, vertically ordered DM spaces generate strong results about the equilibrium decision costs of bad controls.

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 which are reported in Section 5 and proved (like virtually all other results in this paper) 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 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 treats x_1 and x_2 symmetrically. Let $n = 2$, $\lambda_1 = \lambda_2 = \frac{1}{2}$, $C_i = \{i\}$. Denote $p_{x_1x_2} = p(x_1, x_2)$ and $\delta_{x_1x_2} = p(y = 1 \mid x_1, x_2)$.

For an economic story behind this specification, consider firms whose profitability (represented by y) is determined by financial and technical factors (represented by x_1 and x_2). A firm's business 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 technical 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. Since C_1 and C_2 are disjoint, this is an instance of "horizontal" differentiation between data types.

Let us guess the strategy $a = x_i$ for every x and i — i.e., the firm's action tracks the factor it controls for — and examine when this constitutes an equilibrium. Without loss of generality, we can focus entirely on type 1's reasoning. Begin by calculating his subjective estimate of actions' causal effect on profits, conditional on x_1. Since $y \perp a \mid x$,

$$p(y = 1 \mid a, x_1) = \sum_{x_2} p(x_2 \mid a, x_1)\delta_{x_1x_2}$$

for every a, x_1. For $a = x_1$ to be a best-reply for type 1, we must have

$$\begin{aligned} p(y &= 1 \mid a = 1, x_1 = 1) - p(y = 1 \mid a = 0, x_1 = 1) \geq \theta \\ p(y &= 1 \mid a = 1, x_1 = 0) - p(y = 1 \mid a = 0, x_1 = 0) \leq \theta \end{aligned}$$

Let us derive expressions for $p(x_2 = 1 \mid a, x_1)$ for all combinations of a and x_1, as induced by the firm's strategy:

$$p(x_2 = 1 \mid a = 1, x_1 = 1) = \frac{p_{11}}{p_{11} + \lambda_1 p_{10}}$$

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

$$p(x_2 = 1 \mid a = 0, x_1 = 0) = \frac{\lambda_1 p_{01}}{p_{00} + \lambda_1 p_{01}}$$

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 firm is of type 2.

Plugging the expressions for $p(x_2 \mid a, x_1)$ in the best-reply conditions, we obtain

$$\frac{2p_{11}}{2p_{11} + p_{10}}(\delta_{11} - \delta_{10}) \geq \theta \geq \frac{2p_{00}}{2p_{00} + p_{01}}(\delta_{01} - \delta_{00})$$

These inequalities ensure that the strategy we guessed is an equilibrium. The firm's expected loss in this equilibrium is

$$\theta \cdot (p_{11} + \lambda_1 p_{10} + \lambda_2 p_{01}) = \theta \cdot p(x_i = 1)$$

In particular, when the factors x_1 and x_2 are independently and uniformly distributed, the equilibrium condition simplifies into

$$\delta_{11} - \delta_{10} \geq \frac{3}{2}\theta \geq \delta_{01} - \delta_{00}$$

and the welfare loss is $\frac{1}{2}\theta \leq \frac{1}{3}$. A more extreme specification is $p_{11} \approx 1$ and $\delta_{x_1 x_2} = x_1 x_2$. In this case, the equilibrium condition holds for every $\theta < 1$, and the equilibrium expected welfare loss is approximately θ, which itself can be arbitrarily close to the non-equilibrium benchmark of 1.

The intuition behind this result is that since type i varies its 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. While correlation between x_1 and x_2 can exacerbate the DM's welfare loss, it is not necessary for it: As we saw, the loss can arise even when the two factors are independent. The reason is that although the two data types condition their actions on independent exogenous variables, their subjective causal estimates involve conditioning on a — a variable whose distribution records firms' *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 data 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 $y \in \{0,1\}$. Suppose $\delta_t = t$ — i.e., $y = 1$ (high math performance) if and only if $t = 1$ (the child likes playing chess). Let $K = 0$ and $n = 1$ — i.e., there is a unique data type, $C = \emptyset$. Denote $\alpha_t = \sigma_t(a = 1)$.

I 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

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

Since the DM's intrinsic payoff from playing $a = 1$ increases with t, we must have $\alpha_1 \geq \alpha_0$ in equilibrium. Now obtain explicit expressions for the terms that define Δ:

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

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

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. Therefore, the ε-equilibrium requirement is that $\alpha_0 > \varepsilon$, a contradiction. It follows that $\alpha_0 > 0$ in equilibrium. There are two cases to consider.

Case 1: $\alpha_0 \in (0, 1)$. This requires the DM to be indifferent between the two actions — i.e., $\Delta = \theta$. Therefore, $\gamma < \theta$ 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 $\gamma(1-\gamma)$.

Case 2: $\alpha_0 = 1$. This requires us to sustain equilibrium with 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 (and only if) $\gamma \gtrapprox \theta$, this is consistent with equilibrium. The DM's welfare loss in this equilibrium is $\theta \cdot (1 - \gamma) \cdot 1 \leq \gamma(1 - \gamma)$. 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). $\square$

As in Example 3.1, equilibrium forces in Example 4.1 "protect" the DM from causal errors, by pushing his welfare loss far below the non-equilibrium benchmark. 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 is associated with a lower equilibrium frequency of errors. This is why equilibrium effects limit the expected cost of failing to control for x.

I 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. An instance of this domain restriction is that a student's school performance is determined by whether he enjoys studying.

Lemma 1 *Suppose N is vertically ordered. If $y \perp x \mid t$, then in equilibrium, $\alpha_t = 1$ and $\Delta_i(x) \geq 0$ for every data type i.*

Thus, when data types are vertically ordered, the DM's estimated causal effect of a on y is always non-negative in equilibrium, regardless of his data

type and the realization of x. While data types may disagree on the magnitude of the causal effect of a on y, they all agree on its *sign*. Consequently, the DM must always play $a = 1$ when $t = 1$ in any equilibrium.

As with Proposition 1, the proof of Lemma 1 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 more intricate.

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.[4]

Example 4.2: Chess and math with a control variable
Enrich Example 4.1 by endowing it with the structure of Example 2.1: $K = 1$ and $n = 2$, such that $C_1 = \{1\}$ and $C_2 = \emptyset$; $x \in \{0, 1\}$ and $p(x = t \mid t) = q$ for every t. In the context of the chess-and-math story, x may represent the parent's scientific background.

By Lemma 1, we can take it for granted that in any equilibrium, every type plays $a = 1$ when $t = 1$. Therefore, the expressions for $\Delta_1(x)$ and Δ_2 that I derived in Example 2.1 apply here. Suppose $\alpha_1(1) = \alpha_2 = 1$ and $\alpha_1(0) = 0$ — i.e., the naive type always administers playing chess, while the sophisticated type administers playing chess against the child's liking if and only if $x = 1$. This strategy is consistent with equilibrium whenever $\Delta_1(1) \geq \Delta_2 \geq \theta \geq \Delta_1(0)$. To confirm that this condition is not vacuous, let $\gamma = \frac{1}{2}$, $q = \frac{2}{3}$, $\theta = \frac{3}{5}$, $\lambda_1 = \lambda_2 = \frac{1}{2}$. Plugging these parameter values in the expressions for $\Delta_1(1)$, $\Delta_1(0)$ and Δ_2, we can verify

[4]This limit case is not a special case of Section 3, because it implies $x \perp y$, which Section 3 obviously did not assume.

the equilibrium inequalities (it can also be shown that the equilibrium is unique). The equilibrium welfare loss is

$$\theta(1-\gamma)[\lambda_1(q\alpha_1(0)+(1-q)\alpha_1(1))+\lambda_2\alpha_2]=\frac{1}{5}$$

which is strictly below the upper bound for $\gamma=\frac{1}{2}$. $\square$

The following result provides a more general equilibrium characterization when there is a single potential control variable. In this case, the data types are vertically ordered because there the only possible types are $\{1\}$ and $\emptyset$. The result is easily extendible to any setting with $K>1$, $n=2$, and $C_2=\emptyset$. I impose minor restrictions on the primitives (the result does not rely on them; they merely shorten the proof).

Proposition 4 *Let $K=1$ and $n=2$, such that $C_1=\{1\}$ and $C_2=\emptyset$. Suppose that $p(t,x)$ has full support and that $\theta>\gamma$. Then, the equilibrium welfare loss is uniquely determined for any given $p(t,x),\lambda,\theta$. It is weakly below $\gamma(1-\theta)$, and weakly decreasing in λ_1.*

This result establishes that when there is a single potential control variable, the equilibrium welfare loss is pinned down by the primitives. It is always below the upper bound given by Proposition 3, and it decreases with the fraction of sophisticates. As we saw in Section 3, the latter property does not hold in general. Note that Proposition 4 does not rule out multiple equilibria. Specifically, when the likelihood ratio $p(x\mid t=0)/p(x\mid t=1)$ is sufficiently close to one for every x, there is a continuum of equilibria, in all of which the welfare loss is exactly $\gamma(1-\theta)$. (In any equilibrium, the "sophisticated" type 1 must make errors with positive probability.)

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 5 *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 the distinction between vertically ordered and unordered type spaces 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 5, 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.

Example 4.3: Chess and math with horizontally differentiated types
Enrich Example 4.1 by letting $K = 2$ and $n = 2$, such that $C_i = \{i\}$ for every $i = 1, 2$. Let $\lambda_1 = \lambda_2 = \frac{1}{2}$. The exogenous variables are $x_1, x_2 \in \{0, 1\}$. Denote $p_{x_1 x_2} = p(x_1, x_2)$. Suppose $p_{x_1, x_2} = \frac{1}{4}$ for every x_1, x_2; and $t = 1$ if and only if $x_1 = x_2$, such that $\gamma = \frac{1}{2}$. Finally, let $\theta < \frac{2}{3}$. One story behind this specification is that the two variables represent parents' country of origin and social class; in some countries, a taste for chess is common among the upper class, whereas in other countries, it is common among the working class. When parents estimate the effect of playing chess on math performance, some control for country of origin, while others control for social class.

We will now see that under this specification, there is an equilibrium in which each DM type i follows the strategy $a = x_i$. To see why, recall first that $p(y = 1 \mid a, x) = p(t = 1 \mid a, x)$. Furthermore, given the joint distribution over t, x and the DM's postulated strategy,

$$p(t = 1 \mid a = x_1 = 1) = \frac{p_{11}}{p_{11} + \frac{1}{2}p_{10}} = \frac{2}{3} \tag{3}$$

$$p(t = 1 \mid a = x_1 = 0) = \frac{p_{00}}{p_{00} + \frac{1}{2}p_{01}} = \frac{2}{3}$$

Since $p_{10} = p_{01}$, $p(t = 1 \mid a = x_2) = \frac{2}{3}$ for every x_2, too. Moreover, $p(t = 1 \mid a, x_i) = 0$ for all other realizations of (a, x_i). Therefore, since $\theta < \frac{2}{3}$, each type i's best-reply is $a = x_i$.

Note that in this equilibrium, the DM commits decision errors with positive probability at *both* realizations of t: Conditional on any t, the DM

plays $a \neq t$ with probability $\frac{1}{2}$. This happens because the sign of Δ_i is not constant, unlike the case of vertically ordered spaces, where according to Lemma 1 Δ_i is always non-negative. The equilibrium welfare loss is $\theta/2$, which can be arbitrarily close to $\frac{1}{3}$. Note that this loss lies *above* the upper bound of $\frac{1}{4}$ for vertically ordered type spaces under $\gamma = \frac{1}{2}$.

The latter observation, however, is not a robust feature of the horizontally differentiated type space. To see why, modify the example's primitives by setting $p_{11} = \gamma = \frac{1}{2}$ and $p_{00} = 0$. Guess the same strategy, $a = x_i$ for each type i. In this case, $p(t = 1 \mid a = x_1 = 1)$ continues to be given by (3), although this formula now takes the value $\frac{4}{5}$. However, now $p(t = 1 \mid a, x_1) = 0$ for *every* other combination of a and x_1. As long as $\theta < \frac{4}{5}$, the DM's strategy constitutes an equilibrium, though now it does not involve an error at $t = 0$. Moreover, the welfare loss it induces is $\theta(1-\gamma)/2 < \frac{1}{5}$, which lies *below* the upper bound for vertically ordered spaces. □

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 6 *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 three open problems. First, does the characterization in Proposition 4 extend to arbitrary vertically ordered type spaces? Second, does the upper bound $\gamma(1 - \gamma)$ obtained for vertically ordered types extend to distributions p for which $y \not\perp x \mid t$? Finally, 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

$$\Delta_i(x) = \sum_{x_{D \setminus C}} p(x_{D \setminus C} \mid x_C) \left[E_p(y \mid a = 1, x_D) - E_p(y \mid a = 0, x_D) \right] \quad (4)$$

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

Comment: Subjective state spaces

The perceived causal effect given by (4) — and by implication, the simpler formula (1) — 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.[5] The type's estimated causal effect is

$$\sum_{x_3} p(x_3)[E_p(y \mid a = 1, x_3) - E_p(y \mid a = 0, x_3)] \tag{5}$$

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

[5] 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.

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 (5), we obtain a non-null estimated causal effect. The intuition is as follows. Because x_3 is a common consequence of x_1 and x_2 (which are correlated with a and y, respectively), it is not necessarily true that $a \perp y \mid x_3$. Therefore, x_3 is a bad control that produces a biased causal estimate. □

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 3.

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 7 *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 8 *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 7 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 8 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 commonly assumed in empirical research. 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 basic model. The outcome y is purely caused by a and z (i.e., $y \perp (t, x) \mid (a, z)$), such that

$$E_p(y \mid a, z) = \beta a + (1 - \beta)z \tag{6}$$

where $\beta \in (0, 1)$ quantifies the true causal effect of a on y. The DM is unaware of the relation (6), nor does he know β. As before, 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

$$\gamma \cdot p(a = 0 \mid t = 1) \cdot (\theta + \beta) + (1 - \gamma) \cdot p(a = 1 \mid t = 0) \cdot (\theta - \beta) \tag{7}$$

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

Example 6.1: Extreme sports[6]
People sometimes evaluate the riskiness of life-style choices by consulting raw statistics about the prevalence of adverse health outcomes among people who engage in or avoid certain activities. These statistics can be confounded by demographic characteristics. In particular, young people may have a stronger taste for a dangerous activity as well as a lower background health risk. The following example explores the equilibrium effects of changes in the fraction of DMs who control for this confounder.

Suppose that $a = 1$ means that the DM refrains from a potentially pleasurable, yet dangerous activity. For the sake of the example, this activity is extreme sports. The outcome $y = 1$ represents good health — specifically, lack of injuries. Let x represent the DM's age ($x = 0$ indicates a young DM). Let t represent the DM's intrinsic taste for the activity: $t = 0$ means that the DM likes extreme sports. Let $\theta < \frac{1}{2}$.

The objective distribution p is described as follows. First, $p(x = 1) = \frac{1}{2}$. Second, $p(t = x \mid x) = q$ for all x, where $q \in (\frac{1}{2}, 1)$. That is, young (old) people tend to like (dislike) extreme sports. Finally, let $p(y = 1 \mid a, x) = \frac{1}{2}(a + 1 - x)$. That is, old age and extreme sports increase the propensity for injuries. The joint distribution p is consistent with the DAG

$$\begin{array}{ccc} t & \leftarrow & x \\ \downarrow & \swarrow & \downarrow \\ a & \rightarrow & y \end{array}$$

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 $\theta < \frac{1}{2}$, this DM data type will rationally play $a = 1$, independently of t and x. Data type 2 does not control for x.[7] This DM chooses a to maximize

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

[6] This example is dedicated to Kfir Eliaz.

[7] Recall that even if people tend to know their age, they may lack statistics about how age is correlated with a and y, in which case they cannot use the knowledge of their age.

Let us analyze equilibria in this example.

Claim 2 *The rational-choice benchmark can be sustained in equilibrium.*

Proof. 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

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

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. ■

However, equilibria that exhibit decision errors are also possible.

Claim 3 *Assume*

$$\theta > \frac{1}{2} - \frac{2q - 1}{1 + \lambda_1} \tag{8}$$

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

Proof. To verify this claim, let us guess

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

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

What sustains this equilibrium is the positive correlation between age and preferences. Young DMs like extreme sports more than old DMs, and since the DM chooses according to his intrinsic taste with probability $\lambda_2 > 0$, there is positive correlation between doing extreme sports and young age. In turn, this softens the negative correlation between a and y, to an extent that makes it subjectively optimal for type-2 DMs to follow their taste.

The expected welfare loss in this equilibrium is

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

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 his life-style choice. 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.[8]

7 Related Literature

This paper continues my line of research into the behavioral implications of flawed causal reasoning (Spiegler 2016,2020). More broadly, the paper is part of the program of developing equilibrium modeling frameworks with non-rational expectations. There is also a growing literature on (mostly Bayesian) learning foundations for equilibrium behavior under misspecified beliefs (e.g., Heidhues et al. (2018), Esponda et al. (2021), Fudenberg et al.

[8]There is potentially a third equilibrium in which $\alpha_1 = 1$ and $\alpha_0 \in (0,1)$. For brevity, I omit the characterization of this mixed-strategy equilibrium.

(2021), Bohren and Hauser (2021), Frick et al. (2023), Ba (2024)).

Against this background, this paper introduces several innovations. First, the problem it poses — quantifying the cost of flawed methods of causal inference from endogenous datasets — is novel. Second, 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 may fail to perceive that observed actions have direct causes that confound the observed action-consequence relation.[9] Moreover, DM types *differ* in their ability to account for these causes, via their different sets of controls. Since this heterogeneity potentially gives rise to more elaborate confounding patterns, it makes the problem of quantifying the decision costs of bad controls non-trivial. Thus, while the theme that misspecified beliefs can generate equilibrium effects in single-agent decision problems is central to the aforementioned literature, the systematic study of equilibrium effects that arise from DMs' diverse ability to deal with action-consequence confounding is without precedent.

I now demonstrate how existing modeling frameworks of equilibrium behavior under non-rational expectations 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 Jehiel relies on the sequential-equilibrium conceptual baggage.)

[9] 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).

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:

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

Data types can be described by their subjective causal model. Specifically, type i's causal model is given by the DAG

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

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 (4). 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. As a by-product, *equilibrium existence* in the present paper can be established using the same (conventional) arguments as in Spiegler (2016).[10]

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).

[10] 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.

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

DMs who form wrong beliefs because they ignore confounding effects (involving variables other than the DM's action) appear in several examples in Spiegler (2016,2020). Confounder neglect is related to the error of drawing inferences from selective datasets without internalizing 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 Esponda and Pouzo (2017), voters evaluate political candidates according to their historical record as elected policy makers, without taking into account the information conveyed by their election. In Jehiel (2018), entrepreneurs evaluate investments based on a dataset of implemented projects, without realizing they result from selectively positive signals.

The theme that sophisticated agent types can exert a "learning externality" on agent with misspecified models appears in the investment example of Jehiel (2018), as well as in a stylized Roy-model example in Spiegler (2020). Externalities in a similar spirit appear in Frick et al. (2020) and Bohren and Hauser's (2021) models of social learning by agents with heterogeneous prior beliefs or subjective models.

Empirical work has begun to examine how flawed causal inferences from endogenous data interact with individual choice behavior. Oster (2020) presents evidence that naive observational estimates of food supplements' health benefits may be partly driven by selection. The reason is that individuals who adopt a recommendation to consume supplements — based on potentially spurious correlation between action and consequence — are

also likely to engage in other health-improving behavior, thus augmenting the spurious correlation that gave rise to the recommendation in the first place. Angrisiani et al. (2024) present an empirical study of the role of false causal narratives regarding preventive measures in the context of Covid-19. Ambuehl and Thysen (2024) and Charles and Kendall (2024) study causal reasoning about observational data using an experimental methodology.

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 from endogenous datasets.

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

The analysis in this paper was entirely restricted to binary decision problems. As mentioned in the Introduction, this is a reasonable restriction, given the ubiquity of binary treatments in empirical causal-inference studies. Note that the results of Section 3 (and by extension, Proposition 7 in Section 5) are immediately extendible to arbitrary finite decision problems. In contrast, the analysis in Section 4 is tied to the binary specification.

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 that causal-inference errors are so widespread in real life?

References

[1] Ambuehl, S. and H. Thysen (2024). Choosing Between Causal Interpretations: An Experimental Study." Working paper.

[2] 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.

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

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

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

[6] Charles, C. and C. Kendall (2024). Causal Narratives. Working paper, SSRN 4669371.

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

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

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

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

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

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

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

[14] Esponda, I. and D. Pouzo (2017). Conditional Retrospective Voting in Large Elections. American Economic Journal: Microeconomics 9, 54-75.

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

[16] Frick, M., R. Iijima, and Y. Ishii (2020): Misinterpreting Others and the Fragility of Social Learning. Econometrica 88, 2281-2328.

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

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

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

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

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

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

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

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

[25] Oster, E. (2020). Health Recommendations and Selection in Health Behaviors. American Economic Review: Insights 2, 143-160.

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

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

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

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

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

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

Appendix: Proofs

The proofs are presented out of order, because Propositions 1 and 2 are special cases of Propositions 7 and 8.

Proposition 7

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 (4) 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 (4) then implies $\Delta_i(x) = 0$. It follows that in equilibrium, type i plays $a = 0$ for all x. ■

Proposition 8

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. ■

The two following proofs make use of the following notation: For every x and every $C \subseteq \{1, ..., K\}$, denote $\gamma(x) = p(t = 1 \mid x)$ and $\gamma(x_C) = p(t = 1 \mid x_C)$.

Lemma 1

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

$$\Delta_i(x) = [p(t = 1 \mid a = 1, x_{C_i}) - p(t = 1 \mid a = 0, x_{C_i})][\delta_1 - \delta_0] \tag{9}$$

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 (9) can be written as

$$\frac{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i})}{\gamma(x_{C_i})p(a \mid t = 1, x_{C_i}) + (1 - \gamma(x_{C_i}))p(a \mid t = 0, x_{C_i})} \tag{10}$$

Consider the terms $p(a \mid t, x_{C_1})$ in (10). Note that

$$p(a \mid t, x_{C_1}) = \sum_{x_{-C_1}} p(x_{-C_1} \mid t, x_{C_1})p(a \mid t, x_{C_1}, x_{_{-C_1}}) \tag{11}$$

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, (11) 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 (10) 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 (10) implies that $\Delta_i(x) \geq 0$ and $\sigma_{t=1,i}(a = 1 \mid x_{C_i}) = 1$. ■

Comment: Lemma 1 and Simpson's paradox
The key to the lemma's proof 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. A priori, 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}. Monotonicity of conditional probabilities is *not* always preserved under marginalization — a property known as *Simpson's paradox* (see Pearl (2009)). The challenge of the lemma's proof is to ensure that Simpson's paradox is moot in the present context. □

Proposition 3

The proof proceeds in two steps.

Step 1: 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 Lemma 1, $\sigma_{t=1,j}(a=1\mid x_{C_j})=1$ for all j, hence $p(a=1\mid t=1,x_{C_{i(x)}})=1$. Plugging this identity into (9)-(10) 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. □

Step 2: Deriving the upper bound on the DM's ex-ante expected equilibrium welfare loss

Proof: By Step 1, the ex-ante welfare loss is at most

$$\sum_{x}p(x)(1-\gamma(x))\cdot\sum_{x'}\beta(x',x)\gamma(x') \tag{12}$$

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 (12) 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 begin by introducing a few pieces of notation, First, denote

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

By assumption, $c \in (0, 1)$. By Lemma 1, the DM always plays $a = 1$ at $t = 1$. Adopting the notation of Example 4.2, let $\alpha_1(x)$ denote data type 1's probability of playing $a = 1$ at $t = 0$, and let α_2 denote type 2's probability of playing $a = 1$. Finally, denote $q_t(x) = p(x \mid t)$. The proof proceeds stepwise.

Step 1: Expressing equilibrium conditions in terms of α

Proof: Following the same arguments as in Examples 2.1 and 4.2, we can write the data types' estimated causal effects as follows:

$$\Delta_1(x) = \frac{\gamma q_1(x)}{\gamma q_1(x) + (1-\gamma) q_0(x)[\lambda_1 \alpha_1(x) + \lambda_2 \alpha_2]}$$

$$\Delta_2 = \frac{\gamma}{\gamma + (1-\gamma)[\lambda_1 \sum_x q_0(x)\alpha_1(x) + \lambda_2 \alpha_2]}$$

Recall that playing $a = 1$ at $t = 0$ and x is optimal only if $\Delta_i(x) \geq \theta$. Therefore, $\alpha_1(x) > 0$ only if

$$p(a = 1 \mid t = 0, x) = q_0(x)[\lambda_1 \alpha_1(x) + \lambda_2 \alpha_2] \leq c q_1(x) \tag{13}$$

(where $\alpha_1(x) = 1$ if the inequality is strict), and $\alpha_2 > 0$ only if

$$p(a = 1 \mid t = 0) = \lambda_1 \sum_x q_0(x)\alpha_1(x) + \lambda_2 \alpha_2 \leq c \tag{14}$$

(where $\alpha_2 = 1$ if the inequality is strict). □

Step 2: In equilibrium, $p(a = 1 \mid t = 0) \leq c$.

Proof: Assume there is an equilibrium such that $p(a = 1 \mid t = 0) > c$. Then, $\Delta_2 < \theta$ and $\alpha_2 = 0$. Since Δ_2 is a weighted average of all $\Delta_1(x)$'s, there is some x for which $\Delta_1(x) < \theta$ and therefore, $\alpha_1(x) = 0$. But this means that for this value of x, the L.H.S of (13) is zero, hence the inequality is strict, implying $\alpha_1(x) = 1$, a contradiction. □

Step 3: In equilibrium, $p(a = 1 \mid t = 0)$ is uniquely determined.
Proof: Assume the contrary — i.e., there exist two equilibria, represented by α and α', respectively, that induce different conditional action probabilities, $p(a = 1 \mid t = 0) < p'(a = 1 \mid t = 0)$. By Step 2, $p(a = 1 \mid t = 0) < c$. Therefore, $\alpha_2 = 1 \geq \alpha'_2$. Since (14) is the sum of all (13), there must be some x for which

$$\lambda_1\alpha_1(x) + \lambda_2\alpha_2 < \lambda_1\alpha'_1(x) + \lambda_2\alpha'_2$$

and $\alpha_1(x) < \alpha'_1(x)$. In particular, this means that $\alpha'_1(x) > 0$, hence (13) holds for α'. But then it follows that (13) holds strictly for α, such that $\alpha_1(x) = 1$, a contradiction. □

Step 4: The equilibrium level of $p(a = 1 \mid t = 0)$ weakly increases with λ_2.
Proof: Assume the contrary. Then, there exist $\lambda_2 \in (0, 1)$ and $\lambda'_2 = \lambda_2 + \varepsilon$, where $\varepsilon > 0$ is arbitrarily small, such that the conditional action probabilities that these two values induce satisfy $p'(a = 1 \mid t = 0) < p(a = 1 \mid t = 0)$. Therefore, (14) holds strictly under λ'_2, which means that $\alpha'_2 = 1$. In addition, there is x for which

$$\lambda'_1\alpha'_1(x) + \lambda'_2\alpha'_2 < \lambda_1\alpha_1(x) + \lambda_2\alpha_2$$

Suppose $\alpha'_1(x) \geq \alpha_1(x)$. Then,

$$(\lambda_1 - \varepsilon)\alpha_1(x) + (\lambda_2 + \varepsilon) \cdot 1 < \lambda_1\alpha_1(x) + \lambda_2\alpha_2$$

Since ε is arbitrarily small, this leads to a contradiction. Therefore, it must be the case that $\alpha'_1(x) < \alpha_1(x)$. In particular, this means that $\alpha_1(x) > 0$, hence (13) holds for α under λ_2. But then it follows that (13) holds strictly for α' under λ'_2 such that $\alpha'_1(x) = 1$, a contradiction. ■

Proposition 5

(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

$$B_0(x,i) \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} B_1(x,i) \tag{15}$$

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

$$\bar{B}_t(i) = \sum_{x \in X_i^*} p(x \mid t) p(a = 1 \mid t, x) \tag{16}$$

Performing this summation over x_{C_i} on both sides of (15) 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

$$\max_i \bar{B}_0(i) \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} \max_i \bar{B}_1(i) \tag{17}$$

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 (16), we can reformulate (17) 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, (17) can be rewritten as

$$\max_i \sum_{M \mid i \in M} \mu(M) \sum_{j \in M} \lambda_j^* \leq \frac{\gamma(1-\theta)}{\theta(1-\gamma)} \max_i \bar{B}_1(i) \tag{18}$$

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

Step 2: The following inequality holds:

$$\max_i \sum_{M \mid i \in M} \mu(M) \sum_{j \in M} \lambda_j^* \geq \left(\sum_M \mu(M) \sum_{j \in M} \lambda_j^* \right)^2 \tag{19}$$

Proof:[11] If we prove that

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

then this will immediately imply (19) 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

[11] This proof is due to Omer Tamuz.

more demanding inequality:

$$\sum_i \lambda_i^* \sum_{M|i\in M} \mu(M) \sum_{j\in M} \lambda_j^* \geq \left(\sum_M \mu(M) \sum_{j\in M} \lambda_j^* \right)^2 \tag{20}$$

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 (20), 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 (16), 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

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

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 $x_i \in \{0, 1\}$.[12] 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 $\Delta_1(x_1 = 1)$:

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

[12] 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 $x_i = 0, 1$.

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 6

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 $\theta < 1$, 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. ■