# A CAUSAL HIERARCHY FOR AGENTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Agents are defined by the preferences they induce over the different actions available to them. Understanding an agent involves predicting their preferences in any scenario they might face at deployment, as well as in counterfactual circumstances "had things been different". We might ask whether, even in principle, we could determine an agent's preferences (i.e., ranking) over candidate actions empirically, from observations of behaviour? In this paper, we show that this is not possible in general. We show a *preference-level* counterpart to the Causal Hierarchy Theorem (Bareinboim et al., 2022). Specifically, while the CHT states that probabilities at a higher layer (e.g., counterfactual, interventional regimes) are under-determined from lower-layer probabilities (e.g., interventional, observational regimes, respectively), we show that also the *ranking* of actions (a strictly coarser object than a distribution) at a higher layer is generically under-determined: at almost every agent some higher-layer ranking is not determined by lower-layer probabilities. For a fixed counterfactual query such as harm, under-determination is stable under perturbations and becomes typical as the number of actions grows. For example, the question of whether one action produces more harm than another is a statement that is not typically deducible from behavioural data. We conclude with a discussion of the implications of this result for the evaluation of agents, AI Safety, and related disciplines.

## 1 INTRODUCTION

Humans make sense of one another by inferring each other's *preferences*. For example, the fact that a physician consistently prescribes one drug over another might lead us to anticipate that she believes the drug plays a causal role in improving health or that she would also exhibit the same preferences in related circumstances. This habit is frequent and useful and also increasingly extended to AI agents.

The question we ask in this paper is what kinds of inferences about an agent's preferences can be justified from empirical observations? The answer depends on how consistent we expect the agent's decision-making process to be; very little can be predicted of an irrational decision-maker. We might therefore constrain ourselves to agents whose behaviour can be *rationalised* by an (implicit) utility function. In the context of healthcare, biomarkers, tests, and indicators of disease might serve as utility variables, that is, the choice of action is given by the one that optimises the physicians expected value of the biomarker. In other words, we consider agents whose preferences, say between actions $x$ and $x'$, satisfy the following relationship,

$$x \prec_M x' \quad \textit{if and only if} \quad \mathbb{E}_{P_M}[\,\mathcal{U}(Y_x)\,] < \mathbb{E}_{P_M}[\,\mathcal{U}(Y_{x'})\,].$$

Here, the agent is assumed to optimise an (implicit) utility $\mathcal{U}$ derived from an outcome variable $Y$ in the agent's world model $M$. By changing the conditioning set and regime (observational, interventional, counterfactual) we can define a very diverse set of preferences in different cirumstances. In particular, questions about ethics of decision-making, such as fairness (Kusner et al., 2017) and harm (Richens et al., 2022), can be formulated as invariances in what we will define as "counterfactual" preferences.

Preference elicitation has its roots in decision theory (Afriat, 1967; Jeffrey, 1990; Savage, 1972). It is also prominent in inverse reinforcement learning (Abbeel & Ng, 2004; Ng & Russell, 2000; Amin & Singh, 2016), AI Safety (Skalse et al., 2023; Richens et al., 2022; Richens & Everitt, 2024; Bellot et al., 2025), and causality (Pearl, 2009; Mueller & Pearl, 2023). In these fields, a range of impossibility results constrain what inferences are possible, even in principle. In particular,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Bareinboim et al., 2022) showed that the set of structural causal models (a general parameterization of a world model) for which the set of counterfactual probabilities is uniquely determined by (lower layer) interventional facts is measure zero. A general version of this result is known as Pearl's Causal Hierarchy Theorem, CHT for short (Bareinboim et al., 2022; Ibeling & Icard, 2021). Consequently, it holds that two agents that agree on the probabilities of all interventional events might still disagree on the probabilities of counterfactuals.

To understand an agent, in the sense of the intentional stance (Dennett, 1987; Davidson, 1963), is to predict what it would choose and how it would judge particular cases. For example, whether it would favour one action over another. For this purpose, it is sufficient to identify the *ranking* of different expectations, a strictly coarser object than the probabilities themselves. An agent's *preferences* can be invariant to changes in the underlying probabilities that define them as long as *rankings* $x \prec_M x', \forall x, x'$ are preserved. Under-determination of probabilities does not necessarily imply under-determination of preferences. We hypothesize that the choices and intentions of (rational) agents might be predictable from purely empirical facts.

The main result of this paper is to show that,

*An agent's preferences over interventions and counterfactuals are generically underdetermined from lower-layer information.*

This is a counterpart to the CHT for the inference of preference relationships defined with respect to a structural causal model. It says that even with access to all of the agent's observational and interventional probability assignments, preferences over interventions and counterfactuals, respectively, cannot be uniquely inferred in general. The emphasis is on the word *generic*, in the same sense as the CHT: at all but a negligible set of agents, some higher-layer ranking is not determined by lower-layer probabilities, for any number of actions. For a fixed query the picture is finer. Whether an action is harmful, for instance, can be determined from experiments for some agents, but it fails to be determined on an open set of agents, and failure becomes typical as the number of actions grows. As a corollary, this result implies also that preferences can also not be inferred given the agent's policy or the agent's lower-layer preferences (as both are function of the agent's probability assingments).

In the context of the wider AI Safety community and society (Amodei et al., 2016; Dalrymple et al., 2024), this emphasizes the theoretical limits to how much we can infer about an agent's cognition and decision-making process from observations, even in an idealised setting in which the utility function is known and can be described by a structural causal model. On the other hand, it also encourages a positive program of searching for structural assumptions that would enable preference identification (in a similar way that structural constraints allow for the identification of probabilities across the causal hierarchy, see e.g. Tian & Pearl (2002)). Fields of study such as mechanistic interpretability (Olah et al., 2018; Geiger et al., 2024) and causality (Pearl, 2009) are important pieces towards this goal.

## 2 Causality, Agents, and Preferences

In this section we outline some basic principles that we use to reason about preferences and how they might be (implicitly) encoded within its world model.

We use capital letters to denote variables ($X$), small letters for their values ($x$), bold letters for sets of variables ($\boldsymbol{X}$) and their values ($\boldsymbol{x}$), and use $\mathrm{supp}$ to denote their domains of definition ($x \in \mathrm{supp}_X$). To denote $P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{X} = \boldsymbol{x})$, we use the shorthand $P(\boldsymbol{y} \mid \boldsymbol{x})$. We use $\mathbb{1}_{\{\cdot\}}$ for the indicator function equal to 1 if the statement in $\{\cdot\}$ evaluates to true, and equal to 0 otherwise.

### 2.1 Causality

Actions, plans, and hypothetical outcomes can be evaluated by symbolic operations on a model that represents the functional relationships in the world, known as a Structural Causal Model (Pearl, 2009, Definition 7.1.1), or SCM for short, that formalize the notion of an agent's world model.

**Definition 1** (Structural Causal Model)**.** *An SCM $M$ is a tuple $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ where each observed variable $V \in \boldsymbol{V}$ is a deterministic function of a subset of variables $\boldsymbol{Pa}_V \subset \boldsymbol{V}$ and latent variables $\boldsymbol{U}_V \subset \boldsymbol{U}$, i.e., $v := f_V(\boldsymbol{pa}_V, \boldsymbol{u}_V)$, $f_V \in \mathcal{F}$. Each latent variable $U \in \boldsymbol{U}$ is distributed*

*according to a probability measure $P(u)$. We assume the model to be recursive, i.e., that there are no cyclic dependencies among the variables.*

In an SCM $M$, each draw $\boldsymbol{u} \sim P(\boldsymbol{u})$ evaluates to a potential response $\boldsymbol{Y}(\boldsymbol{u}) = \boldsymbol{y}$ and entails a distribution over the possible outcomes $P(\boldsymbol{y})$. The power of SCMs is that they specify not only the joint distribution $P(\boldsymbol{v})$ but also the distribution of variables under all interventions, including incompatible interventions (counterfactuals). Formally, an intervention $do(\boldsymbol{x})$ is modelled as a symbolic operation where values of a set of variables $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, replacing the functions $\{f_X : X \in \boldsymbol{X}\}$ that would normally determine their values. This effectively induces a sub-model of $M$, denoted $M_{\boldsymbol{x}}$. The variables obtained in $M_{\boldsymbol{x}}$ are denoted $\boldsymbol{Y}_{\boldsymbol{x}}$. Counterfactuals are events defined over multiple different sub-models. For example the probability of outcome $\boldsymbol{y}$ under interventions $\boldsymbol{x}$ and simultaneously $\boldsymbol{w}$ under interventions $\boldsymbol{z}$ say, is given by,

$$P(\boldsymbol{y}_{\boldsymbol{x}}, \boldsymbol{w}_{\boldsymbol{z}}) = \int_{\boldsymbol{u}:\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u})=\boldsymbol{y}_{\boldsymbol{x}}, \boldsymbol{W}_{\boldsymbol{z}}(\boldsymbol{u})=\boldsymbol{w}_{\boldsymbol{z}}} dP(\boldsymbol{u}). \tag{1}$$

Implicit in these definitions is a hierarchy of causal expressivity. Following the metaphor offered in (Pearl, 2009), the language of counterfactuals expresses events that are stricly more general than interventional ones, which are in turn more general than observational ones. This result is formalized as Pearl's Causal Hierarchy Theorem (Bareinboim et al., 2022, Thm. 1).

### 2.2 Agents and their preferences

An agent is defined by an (internal) world model and a utility function (possibly specified outside of the agent).

**Definition 2** (Utility)**.** A utility is a non-constant function $\mathcal{U} : \text{supp}_Y \to \{0, 1\}$ of an outcome $Y \in \boldsymbol{V}$.

**Definition 3** (Agent)**.** An agent is a tuple $\mathcal{A} = (M, \mathcal{U})$ where $M$ is an SCM, its internal world model, and $\mathcal{U}$ is the utility that drives its preferences.

Decision-making and preferences depend on a specific expected-utility proposition, such as $\mathbb{E}_{P_M}[\,\mathcal{U}(Y_x)]$ or $\mathbb{E}_{P_M}[\,\mathcal{U}(Y_{x,z}) \mid \boldsymbol{c}]$, depending on the context and question at hand. A proposition like $\mathbb{E}_{P_M}[\,\mathcal{U}(Y_x)]$ imagines that the agent unfolds the effect of action $x$ in its world model $M$ to produce a sub-model $M_x$ and then takes the expectation of the utility over the distribution of the variables in $M_x$. The utility $\mathcal{U}$ is assumed known exactly to the observer.

We call a preference between two actions "interventional" if the expected-utility proposition that rationalizes that choice is defined over an interventional probability.

**Definition 4** (Interventional preference)**.** For any agent $\mathcal{A} = (M, \mathcal{U})$, a preference is interventional if the expected-utility proposition of an action is defined over an interventional probability.

For example, if we instruct an agent $\mathcal{A}$ to maximize a factual utility function $\mathcal{U}$ given its observations so far, its preferences will be of interventional nature, i.e.,

$$x \prec_M x' \quad \textit{if and only if} \quad \mathbb{E}_{P_M}[\,\mathcal{U}(Y_x) \mid \boldsymbol{c}_x] < \mathbb{E}_{P_M}[\,\mathcal{U}(Y_{x'}) \mid \boldsymbol{c}_x], \tag{2}$$

where $\boldsymbol{c}_x$ is some observed context before taking action $x$ (here in fact $\boldsymbol{c}_x = \boldsymbol{c}$ as the context is observed before the action is taken). The key observation is that the expectation is taken over the distribution of the variables in the world model $M_x$, no other submodel or counterfactual reasoning is involved.

We call a preference between two actions "counterfactual" if the expected utility that rationalizes that choice is defined over a counterfactual probability.

**Definition 5** (Counterfactual preference)**.** For any agent $\mathcal{A} = (M, \mathcal{U})$, a preference is counterfactual if the expected-utility proposition of an action is defined over a counterfactual probability.

For example, we might be interested in the agent's beliefs about the causal role of an attribute $Z$, say gender, in its hiring decision $x$. This is a question about a counterfactual preference: "would you hire this person, with recorded features $\boldsymbol{w}$, had the gender been different?", formalized as,

$$x \prec_M x' \quad \textit{if and only if} \quad \mathbb{E}_{P_M}[\,\mathcal{U}(Y_{x,z}) \mid z', \boldsymbol{w}] < \mathbb{E}_{P_M}[\,\mathcal{U}(Y_{x',z}) \mid z', \boldsymbol{w}]. \tag{3}$$

The key observation is that the expectation is taken over the joint distribution of the variables in two different sub-models of the agent's world model, namely $M$ and $M_{x,z}$, i.e., a counterfactual distribution. Here, we may say that the agent is "fair" if its preference order is invariant to swapping $z$ for $z'$ in $\mathbb{E}_{P_M}[\,\mathcal{U}(\boldsymbol{V}_{x,z}) \mid z', \boldsymbol{w}]$, see also (Kusner et al., 2017, Def. 5).

As another example, suppose that a system showed an aggressive version of a job ad (action $x$) to a candidate that ended up not applying for the position (utility $\mathcal{U}(Y)$). We might ask whether, given this information, the AI system still believes that showing the aggressive job ad was the right decision, relating to the a milder version i.e., that aggressive $\prec_M$ mild, which is a question about whether,

$$\mathbb{E}_{P_M}[\,\mathcal{U}(Y_{\text{mild}}) \mid X = \text{agg.}, Y = \text{mild}] < \mathbb{E}_{P_M}[\,\mathcal{U}(Y_{\text{agg.}}) \mid X = \text{agg.}, Y = \text{mild}]. \tag{4}$$

This is again a counterfactual proposition as we are simultaneously considering outcomes under $M$ and $M_x$. This ordering implies a notion of "no harm". That is, the agent, even if the decision on the ad ended up not being successful, expects that changing its decision would not have improved outcomes for this individual, see also Richens et al. (2022). Related notions of intentionality, reponsibility, etc., can also be formalized as statements about counterfactual preferences.

We generically denote an expected-utility proposition by $\varphi$, and the preference order it induces for agent $\mathcal{A}$ by $\preceq^{\varphi}_{\mathcal{A}}$. We write $\preceq^{\text{int}}_{\mathcal{A}}$ and $\preceq^{\text{cf}}_{\mathcal{A}}$ (or even $\preceq^{\text{int}}$ and $\preceq^{\text{cf}}$) for the collection of all interventional and all counterfactual preferences of $\mathcal{A}$, respectively.

### 2.3 Preference inference and collapse

Many problems in AI safety, alignment, and the evaluation of AI agents, can be formalized as a problem of *inferring an agent's preferences from data*. Similarly to causal inference problems, they typically involve ascending levels of the causal hierarchy. Some examples are:

- Action choice elicitation: given an agent $\mathcal{A}$'s beliefs about the probability of events $P(\boldsymbol{V})$, infer the agent's interventional preferences $\preceq^{\text{int}}_{\mathcal{A}}$ for a set of actions in a particular context.
- Causal attribution: given the agent's observed behaviour in some environemnt $P(\boldsymbol{V}_x)$, infer its beliefs on the causal role of an attribute $Z$ in an agent's preferences. Notions of harm and fairness are statements about causal attribution.
- Cross-environment inference: given beliefs about the probability of events under different actions $P(\boldsymbol{V}_x)$ from logs, infer interventional preferences $\preceq^{\text{int}}$ under an additional intervention $do(z)$ that changes the agent's environment.
- Relating competence and intentionality: given that two agents share interventional preferences $\preceq^{\text{int}}$, infer whether they share counterfactual preferences $\preceq^{\text{cf}}$.

In this section we will outline the problem of preference collapse, i.e., when lower-layer probabilities determine higher-layer preferences.

For this it willl be useful to define the notion of agent equivalence. We say that two agents $\mathcal{A}$ and $\mathcal{A}'$ with world models $M$ and $M'$ agree on the probability of observational events, denoted $M \sim_{\text{obs}} M'$ if $P_M(\boldsymbol{v}) = P_{M'}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \text{supp}_{\boldsymbol{V}}$. This will often be the case for competent AI systems trained on the same corpus of facts, whose loss function is defined with respect to an observational distirbution $P(\boldsymbol{v})$. Similarly, we say that they agree on the probability of interventional events, denoted $M \sim_{\text{int}} M'$ if $P_{M_{\boldsymbol{x}}}(\boldsymbol{v}) = P_{M'_{\boldsymbol{x}}}(\boldsymbol{v})$ for all $\boldsymbol{v} \in \text{supp}_{\boldsymbol{V}}$ and $\boldsymbol{x} \in \text{supp}_{\boldsymbol{X}}, \boldsymbol{X} \subset \boldsymbol{V}$. We can expect this kind of equivalence for agents trained with reinforcement learning in similar environments and with similar reward functions.

Two agents are equivalent at a given layer of expressivity (i.e., observational or interventional) if they agree on the probability of events and have the same utility function.

**Definition 6** (Agent equivalence). $\mathcal{A} \equiv_i \mathcal{A}' \iff M \sim_i M' \wedge \mathcal{U} = \mathcal{U}'$, for $i \in \{\text{obs}, \text{int}, \text{cf}\}$.

The relation $\equiv_{\text{int}}$ (that is agreement on interventional world models and utilities) is strictly finer than agreement on interventional rankings. Many interventional distributions induce the same order. For example, an agent that predicts $\mathbb{E}[\,\mathcal{U}(Y_x)\,] = 0.3, \mathbb{E}[\,\mathcal{U}(Y_{x'})\,] = 0.6$ for two actions $x$ and $x'$ will have the same interventional preferences, as an agent that predicts $\mathbb{E}[\,\mathcal{U}(Y_x)\,] = 0.1, \mathbb{E}[\,\mathcal{U}(Y_{x'})\,] = 0.9$ for the same actions – both have the same ranking $\preceq^{\text{int}}$ but different interventional world models.

We say that an agent's preferences collapse if lower-layer probabilities determine higher-layer preferences.

**Definition 7** (Preference collapse)**.** Fix an agent $\mathcal{A}^*$. Interventional preferences *collapse* at $\mathcal{A}^*$ if $\mathcal{A} \equiv_{\text{obs}} \mathcal{A}^*$ implies $\preceq^{\text{int}}_{\mathcal{A}} = \preceq^{\text{int}}_{\mathcal{A}^*}$. Counterfactual preferences *collapse* at $\mathcal{A}^*$ if $\mathcal{A} \equiv_{\text{int}} \mathcal{A}^*$ implies $\preceq^{\text{cf}}_{\mathcal{A}} = \preceq^{\text{cf}}_{\mathcal{A}^*}$.

This mirrors the notion of layer collapse of Bareinboim et al. (2022, Def. 10), with preference rankings in place of probabilities: collapse fails as soon as a single higher-layer ranking is left undetermined. The finer question for a single proposition, such as harm, is treated in Appendix B.

The possibility of collapse is an important proposition. To understand an agent is, in large part, to understand what it would choose and what it judges the causes behind its choices to be (Dennett, 1987). If interventional preferences collapsed at $\mathcal{A}^*$, then every agent that matches $\mathcal{A}^*$ on observational beliefs and utility would share $\mathcal{A}^*$'s ranking over interventions: "seeing would be enough for action", borrowing Pearl (2009) metaphor. If counterfactual preferences collapsed at $\mathcal{A}^*$, then agreement on what experiments reveal would settle how $\mathcal{A}^*$ judges particular counterfactual cases. For example, it should be feasible to simply construct a artificial agent with the right observational or interventional facts and read off its higher-level preferences from its world model. In this case it would not matter that the artificial agent does not match in all respect the reference agent $\mathcal{A}^*$, as any differences would not be visible in their counterfactual preferences. A whole rung of the intentional vocabulary—advice, credit, responsibility, recourse—would reduce to a lower epistemic base.

Collapse is not a feature of "agents in general," but a property of a particular agent $\mathcal{A}^*$. For some agents, and some expected utility propositions such as $\mathbb{E}_P[\,\mathcal{U}(Y_x) \mid y']$, the reduction might hold; for others the layers of preference come apart, as the following example illustrates.

**Example 1.** Let $X \in \{\text{drug}, \text{placebo}\}$ and $Y \in \{\text{recover}, \text{not recover}\}$ encode the action an agent takes and the outcome of that action, respectively. We are interested in inferring the agent's preference in a counterfactual scenario: specifically, understanding its beliefs about the relative chance that a "non-recovered" patient *would* have recovered with the drug versus with the placebo. This is formalized as a difference in expectations $\mathbb{E}_P[\,\mathcal{U}(Y_x) \mid y'] - \mathbb{E}_P[\,\mathcal{U}(Y_{x'}) \mid y']$, where $\mathcal{U}(y) = y$ is the utility function. The question is whether we can determine the agent's preferences (drug vs placebo) in this counterfactual scenario from its beliefs over interventional probabilities $\{P_M(\boldsymbol{v}_x) : x \in \text{supp}_X\}$?

In this example, due to the specific choice of utility function and discreteness of the variables, the tight bounds on counterfactuals of Tian & Pearl (2000) apply, i.e.,

$$\max\left\{0,\ \frac{P(y_x) - P(y)}{P(y')}\right\} \le \mathbb{E}_P[\,\mathcal{U}(Y_x) \le \min\left\{P(x', y'),\ \frac{P(y_x) - P(x, y)}{P(y')}\right\},$$

and for every pair of values $x, x' \in \text{supp}_X$ any value in this interval is attained by some agent with the given interventional probabilities (Lem. 2). Therefore $x' \prec_M x$ holds for *every* such agent if and only if the lower end of the first interval exceeds the upper end of the second,

$$\max\{0,\ P(y_x) - P(y)\} \ >\ \min\{P(x, y'),\ P(y_{x'}) - P(x', y)\},$$

a condition that can be checked from interventional probabilities alone. For instance, with $P(x, y) = \frac{1}{4}$ for all cells, $P(y_x) = 0.7$ and $P(y_{x'}) = 0.3$ give $0.2 > 0.05$: every agent with these beliefs prefers the drug in the counterfactual scenario, so this particular ranking is determined. With $P(y_x) = 0.55$ and $P(y_{x'}) = 0.45$ instead, the contrast ranges over $[-0.3, 0.5]$ and both strict orders are realized by agents with the same interventional beliefs. □

There exist Layer 2 theories at which the counterfactual probabilities entering $\varphi$ are not point-identified, and yet the counterfactual preference $\preceq^{\varphi}$ is identified, as in the first case above. And moreover, it is also not hard to show that there exist $M \not\sim_{\text{int}} M'$ (hence disagreeing on Layer 2 probabilities) that induce the same $\preceq^{\text{cf}}$. Thus preference collapse is *neither necessary nor sufficient* for probabilistic collapse.

## 3 Agent Causal Hierarchy

This section makes precise the senses in which preference determination from data fails. We state the results informally; the setting (an action $X$ with $k$ possible values, a utility that records the

success or failure of an outcome $Y$, and pre-action contexts $\boldsymbol{C}$) and the precise statements are given in Appendix B.

**Assumption 1** (Product measure)**.** Let $\mu$ be a product (Lebesgue-continuous) measure on a suitable encoding of the observational —or, respectively, interventional—distributions of $\boldsymbol{V}$.

Collapse is defined relative to a fixed agent, hence relative to a fixed lower-layer distribution. The product measure defines a notion of typicality on the space of those distributions. For observational distributions, it places a smooth measure on the simplex, i.e., the space of all discrete probability distributions on the variables $\boldsymbol{V}$. For interventional distributions, it places a smooth law on the interventional probabilities $\{P(\boldsymbol{V}_x)\}_{x\in\text{supp}_X}$, drawn independently across actions within the range allowed by the observational distribution.

**Theorem 1** (Agent Causal Hierarchy)**.** *With respect to $\mu$, the set of distributions at which interventional (resp. counterfactual) preferences collapse has probability zero, for any number of actions. Moreover, if no action dominates the observational distribution, the probability that a given counterfactual proposition, such as harm, determines its ranking vanishes as the number of actions grows.*

In words, the theorem states that if you pick an agent's lower-layer beliefs "at random" (observational, or experimental), those beliefs almost never pin down its higher-layer preferences. Like the Causal Hierarchy Theorem of Bareinboim et al. (2022, Thm. 1), this is a statement about a whole layer. For a single counterfactual question, such as harm, the ranking can be determined at some agents (Ex. 1), but the chance that it is goes to zero as the agent faces more possible actions. We illustrate this idea in Fig. 1.

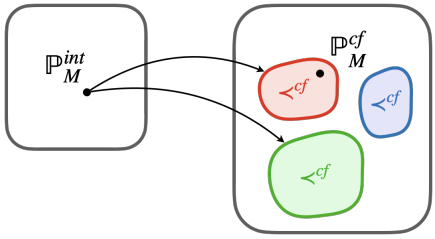


Figure 1: Illustration of Thm. 1: the set of interventional distributions induced by an SCM $M$ (right, written $\mathbb{P}^{int}_{M}$) is generally compatible with multiple counterfactual preference orders (left, with $\mathbb{P}^{cf}_{M}$ the set of counterfactual distributions induced by $M$).

There are essentially two parts to the argument, one showing that interventional preferences do not collapse given observational beliefs, the second showing that counterfactual preferences do not collapse given interventional beliefs. The first part is stronger: observational data alone *never* determines any interventional ranking, for any number of actions. For any given agent $\mathcal{A}$ with full-support observational beliefs $P_M(\boldsymbol{V})$ we can *always* find a different agent $\mathcal{A}'$ that matches $\mathcal{A}$ on observational beliefs and utility but disagrees on interventional preferences, even when only 2 actions are considered.

**Example 2.** Fix an agent $\mathcal{A}$ with a world model $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P\rangle$ and utility $\mathcal{U}(y) = y$. Here $\boldsymbol{V} = \{X, Y\}$ and $\boldsymbol{U} = \{U\}$. Continuing with the medical use case, we might imagine that $X \in \{0, 1\}$ encodes drug assignment, $Y \in \{0, 1\}$ patient recovery, and $U$ a latent patient type, uniform on $\{1, 2, 3, 4\}$. $\mathcal{A}$'s world model is given by

$$\mathcal{F} \;=\; \begin{cases} X \leftarrow \mathbb{1}_{\{U\in\{1,2\}\}} \\ Y \leftarrow \mathbb{1}_{\{U\in\{1,3\}\}} \;\vee\; \big(X \wedge \mathbb{1}_{\{U=4\}}\big). \end{cases}$$

We can read off from the model that $P_M(x, y) = \frac{1}{4}$ for all $x, y$, and that $\mathcal{A}$ prefers the drug over the placebo, i.e., placebo $\prec^{\text{int}}_{\mathcal{A}}$ drug since,

$$\mathbb{E}_{P_{\mathcal{A}}}[\,\mathcal{U}(Y) \mid \text{do}(X{=}1)] \;=\; \tfrac{3}{4} \geq \tfrac{1}{2} \;=\; \mathbb{E}_{P_M}[\,\mathcal{U}(Y) \mid \text{do}(X{=}0)].$$

Now, consider a different agent $\mathcal{A}'$ that internalizes a world model $M'$ with the following set of mechanisms:

$$\mathcal{F}' \;=\; \begin{cases} X \leftarrow \mathbb{1}_{\{U\in\{1,2\}\}} \\ Y \leftarrow \mathbb{1}_{\{U\in\{1,3\}\}} \;\vee\; \big(\neg X \wedge \mathbb{1}_{\{U=2\}}\big). \end{cases}$$

Observational beliefs are the same: $P_{\mathcal{A}'}(x, y) = P_M(x, y) = \frac{1}{4}$ and $M' \sim_{\text{obs}} M$. But $\mathcal{A}'$ prefers the placebo over the drug, i.e., drug $\prec^{\text{int}}_{\mathcal{A}'}$ placebo since,

$$\mathbb{E}_{P_{\mathcal{A}'}}[\,\mathcal{U}(Y) \mid \text{do}(X{=}1)] \;=\; \tfrac{1}{2} \leq \tfrac{3}{4} \;=\; \mathbb{E}_{P_{M'}}[\,\mathcal{U}(Y) \mid \text{do}(X{=}0)].$$

Two agents that share the uniform observational beliefs and the utility $\mathcal{U}(y) = y$ can therefore disagree on which treatment to prefer. □

The situation for counterfactual preferences is different. Ex. 1 showed that a counterfactual ranking *can* be determined: when interventional effects are sufficiently separated, bounds on the difference in counterfactual expectations can be shown to determine the ranking. In contrast, when those effects are close enough, the same bounds are wider, and SCMs that induce different sets of counterfactual probabilities will realize different counterfactual rankings while preserving every observational and interventional probability. The next example shows such a case.

**Example 3.** Consider a system with $\boldsymbol{V} = \{X, Y\}$ with treatments $X \in \{0, 1, 2\}$ and recovery $Y \in \{0, 1\}$, and exogenous $\boldsymbol{U} = \{U_X, U_Y\}$ with $U_X, U_Y$ uniform on $\{0, 1, 2\}$ and $\{0, 1\}$, respectively. The agent $\mathcal{A}$ is placed in a counterfactual scenario, assuming the patient did not recover from treatment 0, and asked about preferences between treatments given that information, i.e.,

$$\{X = 1\} \preceq^{\text{cf}}_{M} \{X = 2\} \quad \iff \quad \mathbb{E}_M[\,\mathcal{U}(Y_1) \mid X{=}0, Y{=}0] \;\leq\; \mathbb{E}_M[\,\mathcal{U}(Y_2) \mid X{=}0, Y{=}0].$$

Its world model is given by

$$\mathcal{F} \;=\; \begin{cases} X \leftarrow U_X \\ Y \leftarrow U_Y \mathbb{1}_{\{X \in \{0,1\}\}} + (1 - U_Y)\mathbb{1}_{\{X=2\}} \end{cases}$$

Then $P_M(x, y) = \frac{1}{6}$ for every $x, y$ and $P_M(Y{=}1 \mid \text{do}(x)) = \frac{1}{2}$ for every $x$. Evaluating the counterfactual expectation reveals a preference $\{X = 1\} \prec^{\text{cf}}_{\mathcal{A}} \{X = 2\}$ since $P_M(Y_1{=}1 \mid X{=}0, Y{=}0) = 0$ and $P_M(Y_2{=}1 \mid X{=}0, Y{=}0) = 1$.

Now consider a second agent $\mathcal{A}'$ with the following world model:

$$\mathcal{F}' \;=\; \begin{cases} X \leftarrow U_X \\ Y \leftarrow U_Y \mathbb{1}_{\{X \in \{0,2\}\}} + (1 - U_Y)\mathbb{1}_{\{X=1\}} \end{cases}$$

We can verify that this assignment matches observational and interventional probabilities exactly but reverses the counterfactual preference since now: $P_{M'}(Y_1{=}1 \mid X{=}0, Y{=}0) = 1$ and $P_{M'}(Y_2{=}1 \mid X{=}0, Y{=}0) = 0$, that is $\{X = 2\} \prec^{\text{cf}}_{\mathcal{A}'} \{X = 1\}$. □

Thm. 1 shows that this situation becomes typical as the number of actions grows. The following theorem shows that the property of non-collapse is stable in the sense that it persists under small perturbations of the observed (resp. experimental) distribution.

**Theorem 2** (Stability)**.** *With respect to the Euclidean topology on a suitable encoding of the observational —or, respectively, interventional—distributions of $\boldsymbol{V}$, the set at which interventional (resp. counterfactual) preferences fail to collapse contains an open and dense set.*

In words, outside a thin set, preference determination fails at every distribution and at all nearby ones. For example, an observer who only approximately knows the agent's lower-layer beliefs cannot treat non-collapse as a measure-zero pathology: if the true agent beliefs sit in that open set, small estimation error still leaves the higher-layer ranking underdetermined. For a single counterfactual proposition such as harm the same holds on a nonempty open set of experimental distributions (Thm. 4).

As a corollary, we can show that if instead of probabilities we are given only a ranking of interventional preferences, or only a policy, then counterfactual rankings are *never* determined.

**Corollary 1** (Preferences do not determine preferences)**.** *For every interventional ranking $\preceq^{\text{int}}$, there exist agents with the same ranking that disagree on $\preceq^{\text{cf}}$, for example on harm.*

For example, as an observer, if the only insight into the agent's world model is its preferences over different actions, then we are never justified in making conclusions about its counterfactual preferences. Similarly, in the context of training artificial agents to match human preferences, even perfect imitation of interventional preferences is not sufficient to determine counterfactual ones.

Further, consider an agent's policy, possibly in environments perturbed by an intervention $\text{do}(z)$ on the context, i.e.,

$$\pi_{\mathcal{A}}(c; z) \in \arg\max_x \mathbb{E}_{P_M}\big[\,\mathcal{U}(\boldsymbol{V}_x) \mid \text{do}(z),\, c\,\big].$$

Then, it holds that two agents may have the same policy in every such environment, but still disagree on counterfactual rankings.

**Corollary 2** (Matched policy does not determine $\preceq^{\text{cf}}$)**.** *For every policy $\pi$ there exist agents $\mathcal{A} = (M, \mathcal{U})$ and $\mathcal{A}' = (M', \mathcal{U})$ such that $\pi_{\mathcal{A}}(c; z) = \pi_{\mathcal{A}'}(c; z) = \pi(c; z)$ for every context $c$ and every intervention $\text{do}(z)$, and $\preceq^{\text{cf}}_{\mathcal{A}} \neq \preceq^{\text{cf}}_{\mathcal{A}'}$.*

Ex. 4 in Appendix B illustrates both corollaries with two agents that share a ranking and a policy but disagree on whether a second drug would have helped a patient for whom the first failed.

We have assumed in this section that the agent's utility function $\mathcal{U}$ as well as the agent's implicit probability assignments, e.g., $P_M(\boldsymbol{V})$, are known exactly to the observer. Relaxations of these would further emphasize the non-collapse of preferences across the causal hierarchy.

## 4 Discussion

We discuss in this section the consequences, limitations, and opportunities of the agent causal hierarchy theorems in Sec. 3.

The theorems are limits on what an observer can certify, in the case most favorable to the observer: the utility is known, the agent is an expected-utility maximizer, and the probabilities of lower-layer events are known exactly. Evaluations of agents encountered in practice (human or artificial) are harder than that, so that these limits are a "floor" on what can be certified.

**Evaluations.** Evaluations give insight into an agent's probability beliefs and preferences at a particular layer of expressivity, i.e., observational or interventional. For example, a set of offline observations, without clear knowledge of the underlying policy for assigning actions, gives us information on an observational distribution. By Thm. 1 (see Thm. 6), every full-support distribution is compatible with opposite interventional rankings, as in Ex. 2: two agents can share the same medical records and the same utility and still disagree on how they judge the influence of the drug. In contrast, if we can query the agent repeatedly in environments that can be arbitrarily perturbed, make observations, and learn about it's policy in each one of them, we may approximate an interventional layer of distributions $\{P(\boldsymbol{V}_{\boldsymbol{z}}) : \boldsymbol{z} \in \text{supp}_{\boldsymbol{Z}}, \boldsymbol{Z} \subset \boldsymbol{V}\}$. This, however, does not determine a genuine counterfactual ranking, in general. Matching the agent's choices is weaker still. Cor. 1 says that every family of interventional rankings is realized by agents that disagree on harm. Cor. 2 adds that every policy, including the policy after interventions on the context, is realized by agents that disagree on harm. Approaches that involve imitation or reinforcement learning therefore carry the same limitations. Training a system to match observed choices, or to match an interventional ranking, fixes $\pi$ or $\preceq^{\text{int}}$, which is not sufficient to determine $\preceq^{\text{cf}}$.

**AI Safety and counterfactuals.** A possible inference that is tempting to make in alignment work is that if an agent behaved well on the benchmark, it also "believes" the safety-relevant counterfactuals. Important safety properties are counterfactual notions: harm, fairness, responsibility, and recourse are counterfactual preferences (Amodei et al., 2016; Shah et al., 2022; Dalrymple et al., 2024). The results in this paper show that this kind of cross-layer inference is unavailable in general from empirical knowledge (i.e., data) alone. At almost every agent some counterfactual ranking is left undetermined by its experimental beliefs, and for a fixed notion such as harm, failure becomes typical as the number of actions grows (Thm. 1); non-collapse is moreover stable under perturbations (Thm. 2). An evaluator who only approximately knows the experimental table therefore cannot treat the failure as a property of specific idealised distributions only. It is therefore not the case that an agent which behaves well on a benchmark thereby judges the safety-relevant counterfactuals the same way. Any attempt to recover the agent's "causal world model" (to the extent that this is done from empirical data), e.g., Richens & Everitt (2024); Richens et al. (2025), does not bridge this gap either: counterfactual preferences remain under-determined.

**Impossibility results**. In the literature, several impossibility results carry a similar signature to the one considered in this paper; in the sense that they attempt to make inference on a feature of the underlying SCM from data sampled from that SCM. (Bareinboim et al., 2022) study a notion of probabilistic collapse, showing that we are exceedingly unlikely to identify the an entire layer of probability statements, e.g., all counterfactual probabilities, from lower-layer information, e.g., all interventional probabilities implied by an SCM. Learning "causal world models" has a similar

signature: is the problem of characterizing the functional dependencies between relevant variables of the SCM from data. It is well known that only a coarse description of it, like equivalence classes, is identifiable from data. An interesting variation of it in the context of agents was proposed by (Richens & Everitt, 2024; Richens et al., 2025); where the authors approximate the agent's world model from the agent's policy in different environemnts or under different reward functions. In the decision theory and inverse reinforcement learning literatures, the free object is often the reward or utility function (Ng & Russell, 2000; Abbeel & Ng, 2004; Amin & Singh, 2016; Rothkopf & Dimitrakakis, 2011).

The impossibility results in this paper, however, also point to a positive program to overcome the lack of identifiability from data. The theorems assume an observer who has only lower-layer data or insist on recovering the set of preferences exactly. Partial identification, interpretability, elicitation, and counterfactual language are ways of leaving that assumption.

**Partial Identification and further assumptions.** The preference ordering is coarser than the probability assignments, which makes it identifiable in pieces. In Ex. 1 the probability of benefit is an interval, and the preference is determined as soon as the interval lies entirely above or below zero. Monotonicity, a graph, an exclusion restriction, or a restricted function class shrink the interval further, in the same way that monotonicity collapses the probability of necessity and sufficiency to the average treatment effect, and in the way structural constraints restore cross-layer identification of probabilities more generally (Tian & Pearl, 2000; 2002). Further advances in partial causal identification (Balke & Pearl, 1994; 1997; Zhang et al., 2022; Bellot & Chiappa, 2024; Jalaldoust et al., 2024; Bellot, 2024), with the right assumptions and inductive biases can improve our predictions of (rational) agent behaviour. One example of inductive bias in the context of agents might arise from the mechanistic interpretability literature (Olah et al., 2020; Geiger et al., 2021; 2024; Nanda et al., 2023). A recovered circuit or causal abstraction is informative for $\preceq^{\text{cf}}$ when it constrains that joint response. Interpretability could be relevant because it is access to the mechanism, which is the kind of structure the bounds can use.

## 5 Conclusion

A rational agent's preferences are a ranking of actions by expected utility. Whether this preference ranking can be inferred from observations or data or not is an important question for the development of AI safety and alignment. For example, we might be reluctant to deploy an agent that we cannot certify is aligned with our safety and ethical preferences; as might be expected in domains such as law or medicine. This paper studies this question from a causal perspective. We show that preference rankings (for a range of different types of preferences one might define) are generically underdetermined from lower-layer information, even in an idealised setting in which the utility function is known and the agent's internal world model can be described by a structural causal model. The prospects for predicting agent preferences are not necessarily purely negative, however. A causal hierarchy theorem, as noted by (Bareinboim et al., 2022; Ibeling & Icard, 2021), is a reason to study the full range of causal-inductive assumptions: how they relate to one another, and what each of them makes it possible to infer.

### AI use statement

In this work, we used generative AI tools to help investigate and surface related work (suggesting relevant comparisons and interpretations of our results in the context of related work), check the correctness of the proofs, and draft the examples used throughout the paper. We have not used generative AI tools to develop the theoretical model, formulate the mathematical claims, or supply major ingredients of the proofs, or to generate figures. We have reviewed all AI-assisted work and take responsibility for the final content of this work. The other tasks for which disclosure is required are not applicable to this work: the paper reports no experiments, datasets, synthetic data, implementation, translation, or qualitative analysis.

## References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 1–8, 2004.

Sydney N Afriat. The construction of utility functions from expenditure data. *International economic review*, 8:67–77, 1967.

Kareem Amin and Satinder Singh. Towards resolving unidentifiability in inverse reinforcement learning. *arXiv preprint arXiv:1601.06569*, 2016.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 46–54, 1994.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, 2022.

Alexis Bellot. Towards bounding causal effects under Markov equivalence. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 308–332. PMLR, 2024.

Alexis Bellot and Silvia Chiappa. Towards estimating bounds on the effect of policies under unobserved confounding. *Advances in Neural Information Processing Systems*, 37:104556–104594, 2024.

Alexis Bellot, Jonathan Richens, and Tom Everitt. The limits of predicting agents from behaviour. *arXiv preprint arXiv:2506.02923*, 2025.

David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

Donald Davidson. Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700, 1963.

Daniel C Dennett. *The Intentional Stance*. MIT Press, 1987.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Duligur Ibeling and Thomas Icard. A topological perspective on causal inference. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Kasra Jalaldoust, Alexis Bellot, and Elias Bareinboim. Partial transportability for domain generalization. *Advances in Neural Information Processing Systems*, 37:137768–137805, 2024.

Richard C Jeffrey. *The logic of decision*. University of Chicago press, 1990.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80 (2):319–323, 1990.

Scott Mueller and Judea Pearl. Personalized decision making–a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 663–670, 2000.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *arXiv preprint arXiv:2402.10877*, 2024.

Jonathan Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687. PMLR, 2025.

Constantin A Rothkopf and Christos Dimitrakakis. Preference elicitation and inverse reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pp. 34–48. Springer, 2011.

Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pp. 32033–32058. PMLR, 2023.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 18, pp. 567–573, 2002.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26548–26558. PMLR, 2022.

# A RELATIONSHIP TO CHT

Our notions of preference collapse and the Agent Causal Hierarchy Theorem are deliberate specializations of the framework of Bareinboim et al. (2022). In Bareinboim et al. (2022, Def. 10), Layer $j$ *collapses* to Layer $i$ *relative to an SCM* $M^*$ if $M^* \sim_i M$ implies $M^* \sim_j M$ for every SCM $M$ over the same endogenous variables: agreement on the entire Layer-$i$ probabilistic theory forces agreement on the *entire* Layer-$j$ theory—all interventional probabilities, or all counterfactual probabilities. Their Causal Hierarchy Theorem then states that, with respect to Lebesgue measure on a suitable encoding of SCMs, the set of $M^*$ at which *any* such layer-collapse occurs has measure zero (Bareinboim et al., 2022, Thm. 1): Layer-2 never collapses to Layer-1, and Layer-3 almost never collapses to Layer-2.

We retain the same logical shape—collapse is always relative to a fixed reference—but change both the reference object and what must be determined. The reference is an *agent* $\mathcal{A}^* = (M^*, \mathcal{U}^*)$, not a bare SCM: equivalence $\mathcal{A} \equiv_i \mathcal{A}^*$ requires not only $M \sim_i M^*$ but also $\mathcal{U} = \mathcal{U}^*$ (Def. 7). What must match is not the full higher layer of probabilities, but the agent's *preference rankings* induced by that fixed utility. We use this at two granularities. Layer-level collapse (Def. 7) asks that every higher-layer ranking be determined, the direct analogue of Bareinboim et al. (2022, Def. 10). Query-level collapse (Def. 8) asks this of a single proposition $\varphi$, such as harm. Both are coarser than classical layer-collapse: they ask whether lower-layer data pin expected-utility orderings, not every Layer-$(i{+}1)$ probability. The definitions are relative to the utility of $\mathcal{A}^*$, not to every conceivable utility. (A stronger "collapse for all $\mathcal{U}$" notion would sit closer to determining the underlying interventional or counterfactual probabilities; that is not our primary object.)

The layer-level theorem (Thm. 1, formally Thm. 3) has the same form as the classical CHT. At every agent with full-support observational beliefs, and whose experimental beliefs are not on the boundary of the Manski bounds, some higher-layer ranking is undetermined; the exceptional agents form a closed, nowhere dense, null set for every number of actions. The observational half is stronger than "almost never": *every* interventional proposition fails to collapse at every full-support observational law. For a fixed counterfactual proposition, an almost-everywhere claim is not available, because some Layer 2 theories do determine $\preceq^{\varphi}$ (Exs. 1 and 5). There we prove instead a dichotomy with stability (Thm. 4): either $\varphi$ collapses everywhere, or it fails to collapse on a nonempty open set. We also prove a typicality statement as the number of actions grows (Thm. 5). These results are proved in the setting of Assump. 2, which is narrower than the class of all SCMs over $\boldsymbol{V}$ treated by Bareinboim et al. (2022).

We may argue that the coarser, query-relative object is better suited to understanding agents because agents are often used as optimizers under beliefs: what we need for prediction, advice, harm, and fairness is whether they would *choose* or *judge* alike, not whether they agree on all interventional or counterfactual probability statements. As mentioned, two agents may in principle disagree on probability assignments but still agree on preference rankings.

# B FORMAL SETTING AND STATEMENTS

This appendix states the setting in which our results are proved, and the precise versions of Thms. 1 and 2 and Cors. 1 and 2. Proofs are given in Appendices C and D.

Every negative result below is proved the same way: we build two world models that agree on everything the lower layer records, but rank two actions differently. The room to do so comes from one gap in the data. Observational and experimental data both describe what happens to units under the action they actually receive, or the distribution of outcomes under each action separately. Neither records how the *same* unit would have fared under an action it did not take. Two models may fill this gap differently, and when the ranking depends on how it is filled, the ranking is not determined[1].

- Appendix C (observational data). We fill the gap in the most favourable way for one action and the least favourable way for another, and then swap. This reverses the interventional ranking at every full-support distribution.
- Appendix D (experimental data). Experiments fix how often each action succeeds, but not *which* units succeed (Lem. 1). We compute the exact range of values a counterfactual score can take given the experiments (Lem. 2), and show that the ranking is determined exactly when the ranges of any two actions never cross (Lem. 4). The remaining proofs check when they do cross.

## B.1 SETTING

**Assumption 2** (Setting)**.** All world models considered, that of the reference agent and those of its competitors, are SCMs with finite domains over variables $\boldsymbol{V} = \boldsymbol{C} \cup \{X\} \cup \mathbf{D}$, where

(i) $X$ is the action, with $k := |\mathrm{supp}_X| \geq 2$ values;

(ii) the *contexts* $\boldsymbol{C}$ (possibly empty) are determined before the action: $X$ is not a cause of any variable in $\boldsymbol{C}$;

(iii) $\mathbf{D}$ contains the outcome $Y$, and possibly other variables determined after the action, such as mediators: no variable in $\mathbf{D}$ is a cause of $X$ or of any variable in $\boldsymbol{C}$.

Nothing else is assumed about the causal graph or the exogenous distributions: contexts may cause one another, and any variables may be confounded, including the action with the outcome and the contexts with both. In other words, the observer knows the temporal order of the variables and nothing more about how they are caused. The examples use the simplest case $\boldsymbol{V} = \boldsymbol{C} \cup \{X, Y\}$. A context $\boldsymbol{c}$ is a value of all of $\boldsymbol{C}$.

The utility is that of Def. 2, a non-constant $\mathcal{U} : \mathrm{supp}_Y \to \{0, 1\}$; any utility of the outcome that takes two values reduces to this case by a positive affine transformation, which changes no ranking. We write $G := \{y : \mathcal{U}(y) = 1\}$ for the success outcomes, a nonempty proper subset of $\mathrm{supp}_Y$, so that expected utilities are probabilities of success, $\mathbb{E}[\mathcal{U}(Y_x) \mid \cdot\,] = P(Y_x \in G \mid \cdot\,)$. Consistency ($Y_x = Y$ whenever $X = x$) yields the bounds of Manski (1990) on the interventional probability of success in context $\boldsymbol{c}$,

$$P(X{=}x,\, Y{\in}G \mid \boldsymbol{c}) \;\leq\; P(Y_x{\in}G \mid \boldsymbol{c}) \;\leq\; P(X{=}x,\, Y{\in}G \mid \boldsymbol{c}) + P(X{\neq}x \mid \boldsymbol{c}). \tag{5}$$

**Propositions.** An *interventional proposition* has the form $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}]$, conditioning on context variables observed before the action (possibly none); we write $\Phi_{\mathrm{int}}$ for the set of these. A *counterfactual proposition* has the form $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_{x,\alpha}) \mid \beta]$, with $\alpha$ an additional intervention (possibly empty) and $\beta$ an event of positive probability in the unintervened world, such that $\beta$ is not determined by the context alone or $\alpha$ sets a context variable to a value that contradicts $\beta$, as in (3) and (4); we write $\Phi_{\mathrm{cf}}$ for the set of these. If, in addition, the contexts are roots whose exogenous variables are independent of all others, then propositions in which $\beta$ is determined by the context, such as (3), reduce to interventional probabilities, e.g. $\mathbb{E}[\mathcal{U}(Y_{x,z}) \mid z', \boldsymbol{w}] = P(Y_x \in G \mid z, \boldsymbol{w})$, and therefore collapse; without this condition they need not. The propositions we use to show non-collapse have $\beta$ involving $X$ or $Y$, and need no such condition. In Def. 7, $\preceq^{\mathrm{int}}$ and $\preceq^{\mathrm{cf}}$ are the

[1] Readers interested only in the main ideas can read the *Strategy* paragraph at the start of each proof and skip the details.

collections of rankings induced by all propositions in $\Phi_{\text{int}}$ and $\Phi_{\text{cf}}$, respectively. We also use collapse of a single proposition.

**Definition 8** (Query-level preference collapse)**.** Fix an agent $\mathcal{A}^*$ and a proposition $\varphi$. If $\varphi \in \Phi_{\text{int}}$, then $\varphi$ *collapses* at $\mathcal{A}^*$ when $\mathcal{A} \equiv_{\text{obs}} \mathcal{A}^*$ implies $\preceq^{\varphi}_{\mathcal{A}} = \preceq^{\varphi}_{\mathcal{A}^*}$. If $\varphi \in \Phi_{\text{cf}}$, then $\varphi$ *collapses* at $\mathcal{A}^*$ when $\mathcal{A} \equiv_{\text{int}} \mathcal{A}^*$ implies $\preceq^{\varphi}_{\mathcal{A}} = \preceq^{\varphi}_{\mathcal{A}^*}$.

Interventional (resp. counterfactual) preferences collapse at $\mathcal{A}^*$ in the sense of Def. 7 if and only if every $\varphi \in \Phi_{\text{int}}$ (resp. $\varphi \in \Phi_{\text{cf}}$) collapses at $\mathcal{A}^*$.

**Encoding.** We encode an observational distribution as a point of the probability simplex on $\text{supp}_{\boldsymbol{V}}$, and a Layer-2 theory as the vector of all probabilities $P(\boldsymbol{V}_{\boldsymbol{z}} = \mathbf{v})$. The Layer-2 theories form a polytope $T$ (Appendix D), which we equip with the Euclidean topology and with Lebesgue measure on its affine hull.

### B.2 Layer-level statements

**Theorem 3** (Agent Causal Hierarchy, formal)**.** *Under Assump. 2, let $\mathcal{A}^* = (M^*, \mathcal{U})$ be an agent whose observational distribution $P_{M^*}(\boldsymbol{V})$ has full support.*

*(a) No interventional proposition $\varphi \in \Phi_{\text{int}}$ collapses at $\mathcal{A}^*$. In particular, interventional preferences do not collapse at $\mathcal{A}^*$.*

*(b) If, for some context $\mathbf{c}$ and action $x$, the interventional probability $P_{M^*}(Y_x \in G \mid \mathbf{c})$ lies strictly inside the bounds (5), then counterfactual preferences do not collapse at $\mathcal{A}^*$.*

*Consequently, for every number of actions $k \geq 2$, the observational distributions (resp. Layer-2 theories) at which interventional (resp. counterfactual) preferences collapse are contained in a closed, nowhere dense set of measure zero in the simplex (resp. in $T$).*

None of the results in this appendix requires the contexts to be unconfounded roots or the absence of mediators. The simpler structure only makes Layer 2 explicit (Lem. 3), which we use for examples.

Part (b) is existential in the proposition: the witnesses in its proof are harm-type propositions, conditioned on a factual failure or success of some action. For a fixed counterfactual proposition an almost-everywhere statement at fixed $k$ is not available, because some Layer-2 theories do determine the ranking (Ex. 1). The next two results describe that case.

### B.3 Statements for a fixed counterfactual proposition

**Theorem 4** (Dichotomy)**.** *Under Assump. 2, fix a context $\mathbf{c}$ and a proposition $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \beta]$, where $\beta$ is a nonempty union of cells $\{\boldsymbol{C} = \mathbf{c}, X = x', Y = y'\}$, considered at agents with $P(\beta) > 0$. Exactly one of the following holds.*

*(i) $\varphi$ collapses at every such agent.*

*(ii) The set of Layer-2 theories at which $\varphi$ fails to collapse contains a nonempty open subset of $T$.*

*Case (ii) holds if and only if there is an action $a$ such that $\beta$ contains some, but not all, of the cells $\{\boldsymbol{C} = \mathbf{c}, X = x', Y = y'\}$ with $x' \neq a$.*

Harm and the probability of sufficiency (e.g., $\beta = \{X = x_0, Y \notin G\}$) and the job-ad proposition (4) fall under (ii). The effect of treatment on the treated with two actions ($\beta = \{X = x_0\}$, $k = 2$) falls under (i). The set at which $\varphi$ fails to collapse need not itself be open: for harm, there are points of non-collapse arbitrarily close to points of collapse (Rem. 3).

**Assumption 3** (Product measure, formal)**.** For each $k$, fix a full-support observational distribution $P^{(k)}(\boldsymbol{V})$ with $k$ actions and a context $\mathbf{c}$, and let $\mu_k$ draw the interventional probabilities $\{P(Y_x \in G \mid \mathbf{c})\}_{x \in \text{supp}_X}$ independently across actions, each with a density on its interval (5). Assume there are $\varepsilon, \rho > 0$, independent of $k$ and $x$, such that each of these probabilities lies in $[\varepsilon, 1 - \varepsilon]$ with $\mu_k$-probability at least $\rho$.

For the propositions below, whether $\varphi$ collapses depends on the Layer-2 theory only through $P^{(k)}$ and these probabilities (Lems. 2 and 4), and every value allowed by (5) is realized by some model (Con. 2), so $\mu_k(\varphi \text{ collapses})$ is well defined. The independence across actions is the substantive modelling choice. It is what makes a large menu likely to contain a near-tie, and a near-tie is what leaves a counterfactual ranking free.

**Theorem 5** (Many actions)**.** *Under Assumps. 2 and 3, let* $\varphi^{(k)}(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \beta^{(k)}]$ *with* $\beta^{(k)}$ *a nonempty union of cells* $\{\boldsymbol{C} = \boldsymbol{c}, X = x', Y = y'\}$*, and suppose that no action dominates the observational distribution,* $\max_x P^{(k)}(X{=}x \mid \boldsymbol{c}) \to 0$*. If either*

*(a)* $\beta^{(k)} \subseteq \{X = x_0\}$ *for a fixed action* $x_0$*, as for harm, or*

*(b)* $P^{(k)}(\beta^{(k)} \mid \boldsymbol{c}) \in [p_0, 1 - p_0]$ *for some* $p_0 > 0$ *independent of* $k$*,*

*then* $\mu_k\big(\varphi^{(k)} \text{ collapses}\big) \to 0$ *as* $k \to \infty$*.*

### B.4 Main-text theorems

*Proof of Thm. 1. Strategy.* Thm. 3 shows that collapse can only happen on a thin set of distributions; a measure with a density gives such a set probability zero.

For interventional preferences, encode observational distributions as points of the simplex on $\mathrm{supp}_{\boldsymbol{V}}$ and let $\mu$ have a density there. By Thm. 3(a), collapse can only occur on the relative boundary of the simplex, which is $\mu$-null for every $k$. For counterfactual preferences, with $\mu = \mu_k$ as in Assump. 3, Thm. 3(b) shows that collapse requires every $P(Y_x{\in}G \mid \boldsymbol{c})$ to equal one of its two bounds (5). Each of these probabilities has a density, so this event has $\mu_k$-probability zero for every $k$. The statement for a given proposition is Thm. 5. □

*Proof of Thm. 2. Strategy.* The complement of a closed, nowhere dense set is open and dense.

By the consequence in Thm. 3, the collapse set is contained in a closed, nowhere dense set. Its complement is open and dense and consists of points of non-collapse. □

### B.5 Corollaries

**Corollary 3** (Preferences do not determine preferences, formal)**.** *Under Assump. 2, let* $(R_{\boldsymbol{c}})_{\boldsymbol{c}}$ *be any family of complete preorders on* $\mathrm{supp}_X$*, one for each context, and let* $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}, X{=}x_0, Y{\notin}G]$ *be a harm proposition. There exist agents* $\mathcal{A} \equiv_{\mathrm{int}} \mathcal{A}'$ *whose interventional ranking in every context* $\boldsymbol{c}$ *is* $R_{\boldsymbol{c}}$*, and with* $\preceq^{\varphi}_{\mathcal{A}} \neq \preceq^{\varphi}_{\mathcal{A}'}$*.*

**Corollary 4** (Matched policy does not determine $\preceq^{\varphi}$, formal)**.** *Under Assump. 2, let* $\varphi$ *be a harm proposition as in Cor. 3. For every policy* $\pi : \mathrm{supp}_{\boldsymbol{C}} \to \mathrm{supp}_X$ *there exist agents* $\mathcal{A} = (M, \mathcal{U})$ *and* $\mathcal{A}' = (M', \mathcal{U})$ *with* $\mathcal{A} \equiv_{\mathrm{int}} \mathcal{A}'$ *such that, in every context* $\boldsymbol{c}$*,* $\pi(\boldsymbol{c})$ *is the unique maximizer of* $\mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}]$ *for both agents, and* $\preceq^{\varphi}_{\mathcal{A}} \neq \preceq^{\varphi}_{\mathcal{A}'}$*.*

Cors. 3 and 4 imply Cors. 1 and 2: a disagreement on the harm proposition is a disagreement on $\preceq^{\mathrm{cf}}$. Moreover, the agents constructed in the proofs share the Layer-2 theory of a fill-in model (Con. 2), in which the contexts are roots with their own independent exogenous variables. For such a model, intervening on a context is the same as conditioning on it, so the policy in an environment perturbed by $\mathrm{do}(\boldsymbol{z})$ with $\boldsymbol{Z} \subseteq \boldsymbol{C}$ is the context policy evaluated at the perturbed context; since policies are determined by Layer 2, this holds for both agents.

The following example illustrates Cors. 1 and 2. Here the two agents do not even share interventional probabilities; they share only the ranking and the policy.

**Example 4.** Fix an agent $\mathcal{A}$ with a world model $M = \langle \boldsymbol{V}, \boldsymbol{U}, \mathcal{F}, P \rangle$ and utility $\mathcal{U}(y) = y$. Here $\boldsymbol{V} = \{X, Y\}$ and $\boldsymbol{U} = \{U_X, U\}$, with $U_X$ uniform on $\{0, 1, 2\}$ and $U$ uniform on $\{1, 2, 3, 4\}$, independent. As in Ex. 3, $X \in \{0, 1, 2\}$ encodes a treatment (supportive care, drug A, drug B) and $Y \in \{0, 1\}$ recovery. The counterfactual question is whether, given that A was given and the patient did not recover, would B have worked?

$$\{X = 1\} \preceq^{\mathrm{cf}}_{M} \{X = 2\} \quad \Longleftrightarrow \quad \mathbb{E}_M[\,\mathcal{U}(Y_1) \mid X{=}1, Y{=}0] \;\leq\; \mathbb{E}_M[\,\mathcal{U}(Y_2) \mid X{=}1, Y{=}0].$$

Agent $\mathcal{A}$'s world model here assumes that recovery under B implies recovery under A as follows:

$$\mathcal{F} \;=\; \begin{cases} X \leftarrow U_X \\ Y \leftarrow \mathbb{1}_{\{U=1\}} \;\vee\; \big(\mathbb{1}_{\{X\in\{1,2\}\}} \wedge \mathbb{1}_{\{U=2\}}\big) \;\vee\; \big(\mathbb{1}_{\{X=1\}} \wedge \mathbb{1}_{\{U=3\}}\big). \end{cases}$$

Then $P_M(Y{=}1 \mid \mathrm{do}(0)) = \frac{1}{4}$, $P_M(Y{=}1 \mid \mathrm{do}(2)) = \frac{1}{2}$ and $P_M(Y{=}1 \mid \mathrm{do}(1)) = \frac{3}{4}$, so $\{X = 0\} \prec^{\mathrm{int}}_M \{X = 2\} \prec^{\mathrm{int}}_M \{X = 1\}$ and $\pi_{\mathcal{A}} = 1$. Moreover, a patient with $X = 1, Y = 0$ has $U = 4$ and would not have recovered under B either, so $P_M(Y_2{=}1 \mid X{=}1, Y{=}0) = 0 = P_M(Y_1{=}1 \mid X{=}1, Y{=}0)$ and $\{X = 1\} \sim^{\mathrm{cf}}_M \{X = 2\}$.

A second agent $\mathcal{A}'$'s world model, however, assumes that A and B cover complementary strains, with an additional never-recoverer. Let $U'$ be uniform on $\{1, 2, 3, 4, 5\}$, independent of $U_X$ and

$$\mathcal{F}' \;=\; \begin{cases} X \leftarrow U_X \\ Y \leftarrow \mathbb{1}_{\{U'=1\}} \;\vee\; \big(\mathbb{1}_{\{X=1\}} \wedge \mathbb{1}_{\{U'\in\{2,3\}\}}\big) \;\vee\; \big(\mathbb{1}_{\{X=2\}} \wedge \mathbb{1}_{\{U'=4\}}\big). \end{cases}$$

The interventional probabilities are different, $\frac{1}{5}, \frac{3}{5}, \frac{2}{5}$ for $X = 0, 1, 2$, but the ranking and policy match: $\{X = 0\} \prec^{\mathrm{int}}_{\mathcal{A}'} \{X = 2\} \prec^{\mathrm{int}}_{\mathcal{A}'} \{X = 1\}$ and $\pi_{\mathcal{A}'} = 1$. The two agents, however, have different counterfactual preferences: a patient with $X = 1, Y = 0$ has $U' \in \{4, 5\}$, so $P_{M'}(Y_2{=}1 \mid X{=}1, Y{=}0) = \frac{1}{2} > 0 = P_{M'}(Y_1{=}1 \mid X{=}1, Y{=}0)$ and $\{X = 1\} \prec^{\mathrm{cf}}_{\mathcal{A}'} \{X = 2\}$. □

The results above are proved under Assump. 2, which fixes only the temporal order of the variables. Knowledge of the causal graph, non-binary utilities, and propositions in which $\alpha$ changes the context while $\beta$ involves $X$ or $Y$ are left open (Rem. 4).

## C Observational data and interventional preferences

This appendix proves part (a) of Thm. 3. Throughout, Assump. 2 holds. Fix a success outcome $y^+ \in G$ and a failure outcome $y^- \notin G$.

The observational distribution records each unit's outcome under the action it took, and nothing about what would have happened under another action. We build two models that reproduce the observational distribution exactly and fill in these missing outcomes in opposite ways.

**Construction 1** (Two endpoint SCMs)**.** Let $P$ be an observational distribution of $\boldsymbol{V}$, let $\boldsymbol{c}$ be a context, and let $x_0 \neq x_1$ be two actions. Let $U' \sim \mathrm{Unif}[0,1]$, and choose a function $h = (h_{\boldsymbol{C}}, h_X, h_{\mathbf{D}})$ on $[0,1]$ such that $h(U')$ has distribution $P$. Define

$$M_+ : \quad \begin{cases} \boldsymbol{C} \leftarrow h_{\boldsymbol{C}}(U') \\ X \leftarrow h_X(U') \\ D \leftarrow h_D(U') \quad \text{for each } D \in \mathbf{D} \setminus \{Y\} \\ Y \leftarrow \begin{cases} h_Y(U') & \text{if } X = h_X(U'), \\ y^+ & \text{if } X = x_1 \neq h_X(U') \text{ and } \boldsymbol{C} = \boldsymbol{c}, \\ y^- & \text{otherwise,} \end{cases} \end{cases}$$

and let $M_-$ be the same model with $x_0$ in place of $x_1$ in the second line of the assignment of $Y$. Both models satisfy Assump. 2: $\boldsymbol{C}$, $X$ and $\mathbf{D} \setminus \{Y\}$ are functions of $U'$ alone, and $Y$ is a function of $(\boldsymbol{C}, X, U')$.

In words, both models copy the observed behaviour. They differ only in what a unit would experience under an action it did not take: in $M_+$, action $x_1$ always succeeds on such units, in $M_-$, action $x_0$ does.

**Theorem 6** (Observational probabilities do not determine interventional preferences)**.** *Let $P$ have full support, and let $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}]$. The models $M_+$ and $M_-$ of Con. 1 both have observational distribution $P$, and $x_0 \prec^{\varphi}_{M_+} x_1$ while $x_1 \prec^{\varphi}_{M_-} x_0$.*

*Proof. Strategy.* In $M_+$, every unit that did not take $x_1$ would succeed under it, which makes $x_1$ as good as the data allow; $x_0$ is made as bad as the data allow. $M_-$ does the opposite. We check that both models reproduce $P$, and that the resulting gap between the two actions is strictly positive under full support.

*Same observational distribution.* With no intervention, $X = h_X(U')$, so the first branch of the assignment of $Y$ is always taken and $\boldsymbol{V} = h(U')$. This has distribution $P$.

*Rankings.* All probabilities below are conditional on $\boldsymbol{C} = \boldsymbol{c}$. Under $\mathrm{do}(x)$ the contexts keep their values, a unit with $h_X(U') = x$ keeps its observed outcome, and every other unit receives the outcome of the second or third branch. Hence

$$\mathbb{E}_{M_+}[\mathcal{U}(Y_{x_1})] = P(X{=}x_1, Y{\in}G) + P(X{\neq}x_1), \qquad \mathbb{E}_{M_+}[\mathcal{U}(Y_{x_0})] = P(X{=}x_0, Y{\in}G),$$
$$\mathbb{E}_{M_-}[\mathcal{U}(Y_{x_0})] = P(X{=}x_0, Y{\in}G) + P(X{\neq}x_0), \qquad \mathbb{E}_{M_-}[\mathcal{U}(Y_{x_1})] = P(X{=}x_1, Y{\in}G).$$

Since $P(X{=}x_0, Y{\in}G) \leq P(X{=}x_0) \leq P(X{\neq}x_1)$,

$$\mathbb{E}_{M_+}[\mathcal{U}(Y_{x_1})] - \mathbb{E}_{M_+}[\mathcal{U}(Y_{x_0})] \;\geq\; P(X{=}x_1, Y{\in}G) \;>\; 0,$$

where the last inequality uses full support. In the same way, $\mathbb{E}_{M_-}[\mathcal{U}(Y_{x_0})] - \mathbb{E}_{M_-}[\mathcal{U}(Y_{x_1})] \geq P(X{=}x_0, Y{\in}G) > 0$. So $x_0 \prec^{\varphi}_{M_+} x_1$ and $x_1 \prec^{\varphi}_{M_-} x_0$. □

*Proof of Thm. 3(a). Strategy.* Thm. 6 gives two agents with the same observational distribution as $\mathcal{A}^*$ and opposite rankings; at least one of them disagrees with $\mathcal{A}^*$.

Let $\mathcal{A}^*$ have full-support observational distribution $P$ and let $\varphi(x) = \mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}] \in \Phi_{\mathrm{int}}$. By Thm. 6, the agents $(M_+, \mathcal{U})$ and $(M_-, \mathcal{U})$ are both $\equiv_{\mathrm{obs}}$-equivalent to $\mathcal{A}^*$ and rank $x_0, x_1$ in opposite strict orders, so at least one of them differs from $\mathcal{A}^*$ on $\preceq^{\varphi}$. Hence $\varphi$ does not collapse. If $\varphi$ conditions on only some of the context variables, or on none, replace the event $\{\boldsymbol{C} = \boldsymbol{c}\}$ by the corresponding event throughout; nothing else changes. For the consequence, view observational

distributions as points of the probability simplex on $\mathrm{supp}_{\boldsymbol{V}}$. Interventional preferences can collapse only on its boundary, the set of distributions without full support, which is closed, nowhere dense, and of Lebesgue measure zero. □

**Remark 1.** *The argument does not use that the utility is binary. For any non-constant real-valued utility of the outcome, taking $y^+$ and $y^-$ to be a best and a worst outcome gives the same conclusion, so part (a) of Thm. 3 holds for all such utilities.*

# D EXPERIMENTAL DATA AND COUNTERFACTUAL PREFERENCES

This appendix proves part (b) of Thm. 3, Thms. 4 and 5, and Cors. 3 and 4. Throughout, Assump. 2 holds. The propositions we study have the form

$$\varphi_\beta(x) = \mathbb{E}_M\big[\mathcal{U}(Y_x) \mid \beta\big] = P_M(Y_x \in G \mid \beta) =: s_M(x),$$

where $\beta$ is a nonempty union of cells $\{\boldsymbol{C} = \boldsymbol{c}, X = x', Y = y'\}$ for a fixed context $\boldsymbol{c}$, with $P(\beta) > 0$. Unless $\beta = \{\boldsymbol{C} = \boldsymbol{c}\}$, such a $\beta$ is not determined by the context alone, so $\varphi_\beta \in \Phi_{\mathrm{cf}}$. For harm, for instance, $\beta = \{\boldsymbol{C} = \boldsymbol{c}, X = x_0, Y \notin G\}$ selects the units that took $x_0$ and failed, and $s(x)$ is the chance that such a unit would have succeeded under $x$. Here $\alpha$ is empty. For part (b) this is no loss, since one non-collapsing proposition suffices; propositions with $\alpha \neq \emptyset$ are discussed in Rem. 4.

**Overview.** The argument has three steps.

(1) *What the experiments leave free* (Lem. 1). No experiment shows a unit under two different actions. Among the units that did not take $x$, what they would do under $x$ can therefore be re-paired with the rest of their behaviour without changing any experimental distribution.

(2) *Exact ranges* (Lem. 2). For each action $x$, the experiments fix the share of units that did not take $x$ but would succeed under it, but not which units these are. So they do not fix how many of those successes fall inside $\beta$. This gives each score $s(x)$ an exact range, and these ranges can be attained independently for different actions.

(3) *When the ranking is fixed* (Lem. 4). The ranking of $a$ and $b$ is determined exactly when the range of $s(a) - s(b)$ does not contain two different signs.

The theorems then follow by checking, in each situation, whether two ranges can be made to cross.

**Notation.** Fix the context $\boldsymbol{c}$ (a value of all context variables) and write $P(\cdot)$ for $P(\cdot \mid \boldsymbol{C} = \boldsymbol{c})$. For an action $x$ let

$$\theta_x := P(Y_x \in G), \qquad r_x := P(X \neq x), \qquad q_x := \theta_x - P(X{=}x, Y{\in}G),$$
$$p := P(\beta), \qquad p'_x := P(\beta, X \neq x), \qquad F_x := P(\beta, X{=}x, Y{\in}G).$$

In words: $\theta_x$ is the success probability of action $x$ in the experiment; $r_x$ is the share of units that did not take $x$; $q_x$ is the share of units that did not take $x$ but would succeed under it (by consistency, $q_x = P(Y_x \in G, X \neq x)$); $p$ is the size of $\beta$; $p'_x$ is the part of $\beta$ that did not take $x$; and $F_x$ is the part of $\beta$ that took $x$ and succeeded, which the data record. The bounds (5) read $0 \leq q_x \leq r_x$, and we call $\theta_x$ *interior* if $0 < q_x < r_x$, i.e., if the experiment does not force all, or none, of the units that did not take $x$ to succeed under it.

## D.1 RE-PAIRING AND SHARP RANGES

**Response types.** Every SCM over finite variables is equivalent, for all counterfactual distributions, to one whose only exogenous variable is a *response type*: a list of functions, one per endogenous variable, giving its value for every value of the variables preceding it in a causal order (Balke & Pearl, 1994; Zhang et al., 2022). Under Assump. 2 the order can be taken to list $\boldsymbol{C}$ first, then $X$, then $\mathbf{D}$. A unit is thus described by the responses of the contexts, an action response $\boldsymbol{c}' \mapsto X$, and the responses of the variables in $\mathbf{D}$, which take as input the values of $\boldsymbol{C}$, $X$ and earlier variables in $\mathbf{D}$. A unit's *natural* context and action are those it has when nothing is intervened on. There are finitely many response types, and every probability $P(\boldsymbol{V_z} = \mathbf{v})$ is linear in their distribution, so the Layer-2 theories form a polytope $T$.

**Lemma 1** (Re-pairing)**.** *Fix a context $\boldsymbol{c}$ and an action $x$, and let $H$ be the set of units whose natural context is $\boldsymbol{c}$ and whose natural action is not $x$. For a unit, let its* slice *$S$ be the part of its $\mathbf{D}$-responses with inputs $\boldsymbol{C} = \boldsymbol{c}$ and $X = x$, and let $R$ be the rest of its response type. If a model $M'$ is obtained from $M$ by changing the joint distribution of $(R, S)$ on $H$ while keeping the distributions of $R$ and of $S$ on $H$, and keeping everything outside $H$, then $M' \sim_{\mathrm{int}} M$.*

*Proof. Strategy.* We show that in every experiment, the values of a unit of $H$ are computed either from its slice or from the rest of its response type, and that which of the two is used depends only on the experiment.

Consider a unit of $H$ under $\mathrm{do}(\boldsymbol{z})$. Its slice is used exactly when, in this experiment, its context is $\boldsymbol{c}$ and its action is $x$. Setting context variables to their natural values changes nothing, so its context is $\boldsymbol{c}$ exactly when $\boldsymbol{z}$ agrees with $\boldsymbol{c}$ on the context variables it sets; this depends only on $\boldsymbol{z}$. If $\boldsymbol{z}$ does not set $X$ and the context is $\boldsymbol{c}$, the unit takes its natural action, which is not $x$. So the slice is used exactly when $\boldsymbol{z}$ sets $X = x$ and agrees with $\boldsymbol{c}$ on the contexts it sets, a condition that depends only on $\boldsymbol{z}$. In that case the values of all variables are determined by $\boldsymbol{z}$ and $S$; otherwise they are determined by $\boldsymbol{z}$ and $R$. Hence the distribution of $\boldsymbol{V}_{\boldsymbol{z}}$ is the sum of a part computed from the distribution of $S$ on $H$, a part computed from the distribution of $R$ on $H$, and a part from units outside $H$. None of these changes. Finally, $M'$ is again a response-type model with the same causal order, so it satisfies Assump. 2. □

**Lemma 2** (Sharp ranges). *Fix a Layer-2 theory $\vartheta$ with $P(\beta) > 0$, and for each action $x$ let*

$$\Lambda_x := \big[\max\{0,\ q_x - (r_x - p'_x)\},\ \min\{p'_x,\ q_x\}\big].$$

*(i) For every $M$ satisfying Assump. 2 and realizing $\vartheta$, $p\, s_M(x) = F_x + \lambda_x$ with $\lambda_x := P_M(Y_x \in G, \beta, X \neq x) \in \Lambda_x$.*

*(ii) Conversely, for every choice of $\lambda_x \in \Lambda_x$, one for each action, some model realizing $\vartheta$ attains all of them at once.*

*(iii) Consequently, for $a \neq b$ the set $I_{ab}(\vartheta)$ of values of $s_M(a) - s_M(b)$ over models $M$ realizing $\vartheta$ is the interval*

$$I_{ab}(\vartheta) = \tfrac{1}{p}\Big[F_a - F_b + \min\Lambda_a - \max\Lambda_b,\ \ F_a - F_b + \max\Lambda_a - \min\Lambda_b\Big],$$

*whose endpoints are continuous functions of $\vartheta$ wherever $P(\beta) > 0$.*

The score $s(x)$ has a known part and an unknown part. Units in $\beta$ that took $x$ contribute $F_x$, which is recorded. Units in $\beta$ that did not take $x$ contribute $\lambda_x$, the share of the $q_x$ hypothetical successes of $x$ that fall inside $\beta$. The interval $\Lambda_x$ lists the possible values of $\lambda_x$: at most $p'_x$ of these successes fit inside $\beta$, and at most $r_x - p'_x$ fit outside it.

*Proof. Strategy.* Part (i) is counting: the $q_x$ successes must be split between the inside and the outside of $\beta$. Part (ii) uses Lem. 1 to place them wherever we like, one action after another.

(i) By consistency $Y_x = Y$ on $\{X = x\}$, so $P_M(Y_x \in G, \beta, X = x) = F_x$, which gives the decomposition. The event $\{Y_x \in G, X \neq x\}$ has probability $q_x$ and splits into a part inside $\beta \cap \{X \neq x\}$, of probability $\lambda_x \leq p'_x$, and a part outside it, of probability $q_x - \lambda_x \leq r_x - p'_x$. Hence $\lambda_x \in \Lambda_x$.

(ii) Start from any model realizing $\vartheta$, written with response types, and fix an action $x$. The set $H$ of Lem. 1 consists of the units in context $\boldsymbol{c}$ that did not take $x$, and has mass $r_x$. For a unit of $H$, whether it lies in $\beta$ is determined by $R$, since its observed behaviour does not use its slice, and whether $Y_x \in G$ is determined by $S$, since under $\mathrm{do}(x)$ its context stays $\boldsymbol{c}$. On $H$, the first event has mass $p'_x$ and the second mass $q_x$. Re-pairing $R$ and $S$ on $H$, the mass $\lambda_x$ of units having both can be given any value between the smallest and the largest possible overlap of two sets of masses $p'_x$ and $q_x$ inside a set of mass $r_x$, which is exactly $\Lambda_x$. By Lem. 1 the new model still realizes $\vartheta$. Now repeat for the next action $x'$. The values $\lambda_x$ already chosen are unchanged, because for units that did not take $x'$ they are determined by the rest of the response type (which contains the slice for $x$), and the other units are not touched.

(iii) By (i) and (ii), $(\lambda_a, \lambda_b)$ ranges over all of $\Lambda_a \times \Lambda_b$, and $s(a) - s(b) = (F_a - F_b + \lambda_a - \lambda_b)/p$. The quantities $F_x, p'_x, r_x, q_x, p$ are continuous in $\vartheta$, and so are $\min$, $\max$ and division by $p > 0$. □

**Remark 2** (Tian's bounds). *For $\boldsymbol{V} = \{X, Y\}$, $k = 2$ and $\beta = \{X = 0, Y \notin G\}$ we have $F_1 = 0$, $p'_1 = p = P(X{=}0, Y{\notin}G)$ and $r_1 = P(X{=}0)$. Then $\Lambda_1 = [\max\{0, q_1 - P(X{=}0, Y{\in}G)\}, \min\{p, q_1\}]$, which is the interval of Tian & Pearl (2000) for the probability of sufficiency,* $\mathrm{PS} = \lambda_1/p$. *Ex. 1 is the case $\beta = \{Y \notin G\}$.*

**Realizing prescribed experiments.** Several proofs below choose an observational distribution and success probabilities and need a model that has them. Call success probabilities $\{\theta_x(\boldsymbol{c})\}_{x,\boldsymbol{c}}$ *feasible* for $P$ if each lies within its bounds (5), i.e. $0 \le q_x(\boldsymbol{c}) \le r_x(\boldsymbol{c})$.

**Construction 2** (Fill-in models)**.** Let $P$ be an observational distribution and $\{\theta_x(\boldsymbol{c})\}$ feasible success probabilities. Let $U_{\boldsymbol{C}} \sim P(\boldsymbol{C})$ and $U' \sim \mathrm{Unif}[0,1]$ be independent. For each context $\boldsymbol{c}'$ choose a function $f^{\boldsymbol{c}'} = (f_X^{\boldsymbol{c}'}, f_{\mathbf{D}}^{\boldsymbol{c}'})$ on $[0,1]$ such that $f^{\boldsymbol{c}'}(U')$ has distribution $P(X, \mathbf{D} \mid \boldsymbol{c}')$. For each $(x, \boldsymbol{c}')$, the set $A_x^{\boldsymbol{c}'} := \{u' : f_X^{\boldsymbol{c}'}(u') \neq x\}$ has length $r_x(\boldsymbol{c}')$; choose a function $g_x^{\boldsymbol{c}'}$ on it that takes values in $G$ on a part of length $q_x(\boldsymbol{c}')$ and values outside $G$ on the rest. Set

$$\boldsymbol{C} \leftarrow U_{\boldsymbol{C}}, \qquad X \leftarrow f_X^{\boldsymbol{C}}(U'), \qquad D \leftarrow f_D^{\boldsymbol{C}}(U') \;\; (D \in \mathbf{D} \setminus \{Y\}),$$

$$Y \leftarrow \begin{cases} f_Y^{\boldsymbol{C}}(U') & \text{if } X = f_X^{\boldsymbol{C}}(U'), \\ g_X^{\boldsymbol{C}}(U') & \text{otherwise.} \end{cases}$$

In words: each unit keeps its observed behaviour, and $g_x^{\boldsymbol{c}}$ decides whether a unit that did not take $x$ would succeed under $x$. The model satisfies Assump. 2. With no intervention the second branch is never taken, so its observational distribution is $P$. Under $\mathrm{do}(x)$ in context $\boldsymbol{c}$, the probability of success is $P(X{=}x, Y{\in}G \mid \boldsymbol{c}) + q_x(\boldsymbol{c}) = \theta_x(\boldsymbol{c})$. So any observational distribution together with any feasible success probabilities is realized by some model. In a fill-in model the contexts are roots with their own independent exogenous variable; this is a property of the construction, not an assumption on the agents.

With this extra structure, and no variables after the action other than $Y$, Layer 2 has a simple explicit form. We use this only for examples (Rem. 3).

**Lemma 3** (Layer-2 theories in the basic case)**.** *Let $M, M'$ satisfy Assump. 2 with $\mathbf{D} = \{Y\}$, and suppose in addition that in both models the contexts are roots whose exogenous variables are independent of those of $X$ and $Y$. Then $M \sim_{\mathrm{int}} M'$ if and only if $P_M(\boldsymbol{V}) = P_{M'}(\boldsymbol{V})$ and $P_M(Y_x{=}y \mid \boldsymbol{c}) = P_{M'}(Y_x{=}y \mid \boldsymbol{c})$ for all $x, y, \boldsymbol{c}$.*

*Proof. Strategy.* Because the contexts are unconfounded roots, intervening on a context is the same as conditioning on it. Every interventional distribution therefore reduces to one of four simple cases, depending on whether $X$ and $Y$ are set.

The listed quantities are Layer-2 quantities: $P(Y_x = y \mid \boldsymbol{c})$ can be read off $P(\boldsymbol{V}_x)$ because $X$ does not affect $\boldsymbol{C}$. Conversely, let $\boldsymbol{Z} \subseteq \boldsymbol{V}$, split $\boldsymbol{C} = (\boldsymbol{C}_1, \boldsymbol{C}_2)$ with $\boldsymbol{C}_1 = \boldsymbol{Z} \cap \boldsymbol{C}$, and let $\boldsymbol{U}'$ be the exogenous variables of $(X, Y)$. Under $\mathrm{do}(\boldsymbol{z})$ the roots $\boldsymbol{C}_2$ keep their natural values, and $(X_{\boldsymbol{z}}, Y_{\boldsymbol{z}})$ is a function of $(\boldsymbol{c}_1, \boldsymbol{C}_2, \boldsymbol{U}')$ with $\boldsymbol{U}'$ independent of $\boldsymbol{C}_2$. Given $\boldsymbol{C}_2 = \boldsymbol{c}_2$, $(X_{\boldsymbol{z}}, Y_{\boldsymbol{z}})$ is therefore distributed as $(X, Y)$ under the part of $\boldsymbol{z}$ that sets $X$ or $Y$, in context $\boldsymbol{c} = (\boldsymbol{c}_1, \boldsymbol{c}_2)$. That distribution is $P(X, Y \mid \boldsymbol{c})$ if neither $X$ nor $Y$ is set, $P(Y_x \mid \boldsymbol{c})$ if only $X$ is, $P(X \mid \boldsymbol{c})$ if only $Y$ is (because $Y$ is not a cause of $X$), and a point mass otherwise. Hence $P(\boldsymbol{V}_{\boldsymbol{z}})$ is determined by $P(\boldsymbol{C}_2)$ and these distributions. □

**Lemma 4** (Collapse criterion)**.** *$\varphi_\beta$ collapses at $\mathcal{A}^*$ if and only if, for every pair $a \neq b$, the interval $I_{ab}$ at the Layer-2 theory of $\mathcal{A}^*$ is contained in $(0, \infty)$, contained in $(-\infty, 0)$, or equal to $\{0\}$.*

*Proof. Strategy.* A ranking is fixed once the sign of every pairwise difference is fixed, and Lem. 2 tells us exactly which differences are possible.

A ranking of $\mathrm{supp}_X$ is determined by the signs (negative, zero, positive) of all pairwise differences. If every $I_{ab}$ contains a single sign, every agent $\mathcal{A} \equiv_{\mathrm{int}} \mathcal{A}^*$ has the same ranking as $\mathcal{A}^*$. If some $I_{ab}$ contains two different signs, the two models attaining them rank $a, b$ differently, so at least one of them differs from $\mathcal{A}^*$. □

In particular, $\varphi_\beta$ does not collapse whenever, for some $a \neq b$,

$$F_a + \min \Lambda_a \;<\; F_b + \max \Lambda_b \quad \text{and} \quad F_b + \min \Lambda_b \;<\; F_a + \max \Lambda_a, \tag{6}$$

since then 0 lies strictly inside $I_{ab}$. Both conditions are strict inequalities between continuous functions of $\vartheta$, so they hold on an open set.

### D.2 Proof of the layer-level theorem

*Proof of Thm. 3(b). Strategy.* Layer-level collapse fails as soon as one counterfactual proposition fails to collapse, so it suffices to find one. We use harm: among units that took some other action $x_0$ and failed (or succeeded), how many would have succeeded under $x_1$? The score of $x_0$ is then known exactly (0 or 1), while the score of $x_1$ has a nontrivial range because $\theta_{x_1}$ is interior. One endpoint of that range ties with $x_0$ and the other does not.

Let $\boldsymbol{c}$ and $x_1$ be such that $\theta_{x_1}$ is interior, and write $q := q_{x_1} \in (0, r)$ with $r := r_{x_1}$. Choose any $x_0 \neq x_1$ and consider the two harm-type propositions with

$$\beta_F := \{\boldsymbol{C} = \boldsymbol{c}, X = x_0, Y \notin G\}, \qquad \beta_S := \{\boldsymbol{C} = \boldsymbol{c}, X = x_0, Y \in G\}.$$

By full support their probabilities $p_F, p_S$ are positive, and $p_F + p_S = P(X{=}x_0) \leq r$. For both events, $p'_{x_0} = 0$, so $s(x_0)$ is the same for all compatible models, and $F_{x_1} = 0$.

If $q \leq r - p_F$, use $\beta_F$. Then $s(x_0) = 0$, and $s(x_1) = \lambda / p_F$ with $\lambda \in [0, \min\{p_F, q\}]$, since the lower end of $\Lambda_{x_1}$ is 0. Both $\lambda = 0$ (a tie) and $\lambda > 0$ (strictly $x_0 \prec x_1$) are attainable.

Otherwise $q > r - p_F \geq p_S$, and we use $\beta_S$. Then $s(x_0) = 1$, and $s(x_1) = \lambda / p_S$ with $\lambda \in [\max\{0, q - (r - p_S)\}, p_S]$, whose lower end is below $p_S$ because $q < r$. Both a tie and strictly $x_1 \prec x_0$ are attainable.

In either case Lem. 4 shows that this proposition in $\Phi_{\mathrm{cf}}$ does not collapse at $\mathcal{A}^*$.

For the consequence, recall that the Layer-2 theories form a polytope $T$ in which every probability $P(\boldsymbol{V}_{\boldsymbol{z}} = \mathbf{v})$ is a linear function. By what we just proved, counterfactual preferences can collapse only where some $P(\boldsymbol{V} = \mathbf{v})$ vanishes, or where for every $x$ and $\boldsymbol{c}$ one of $P(\boldsymbol{c})\, q_x(\boldsymbol{c})$ and $P(\boldsymbol{c})\, (r_x(\boldsymbol{c}) - q_x(\boldsymbol{c}))$ vanishes. These are linear functions on $T$, and none vanishes on all of $T$ (Con. 2). The collapse set is therefore closed and contained in finitely many hyperplanes, none of which contains $T$, so it is nowhere dense in $T$ and of measure zero for Lebesgue measure on the affine hull of $T$. This holds for every $k \geq 2$. □

The witnesses above realize a tie against a strict order. Strict reversals require two actions whose scores both depend on units that did not take them, as in Ex. 3 and case (a) of Thm. 5.

### D.3 Proof of the dichotomy

*Proof of Thm. 4. Strategy.* The only freedom in a score $s(a)$ is how many of the units that did not take $a$, and would succeed under it, lie in $\beta$. If, for every action, $\beta$ contains none or all of these units, there is no freedom and the ranking is fixed. Otherwise some score has a range of positive width, and we show, case by case according to which cells $\beta$ contains, that the experiment can be chosen so that two scores have no fixed order. The sets of experiments we exhibit are described by strict inequalities, so they are open.

Call $\beta$ *degenerate* if, for every action $a$, it contains either none or all of the cells $\{\boldsymbol{C} = \boldsymbol{c}, X = x', Y = y'\}$ with $x' \neq a$.

*If $\beta$ is degenerate, (i) holds.* For each action $a$, either $p'_a = 0$ and $\Lambda_a = \{0\}$, or $p'_a = r_a$ and $\Lambda_a = \{q_a\}$, whatever the distribution. So every $s(a)$ is the same for all compatible models, every $I_{ab}$ is a single point, and $\varphi_\beta$ collapses at every agent by Lem. 4.

*If $\beta$ is not degenerate, (ii) holds.* Fix any full-support $P$. Then there is an action $a$ with $0 < p'_a < r_a$; fix it and any $b \neq a$. We exhibit a nonempty open set of Layer-2 theories at which $I_{ab}$ contains two signs. Each set below is described by strict inequalities between continuous functions of the Layer-2 theory, so it is open in $T$, and it is nonempty because Con. 2 realizes $P$ together with any feasible values of $q_a$ and $q_b$. Which of the cases below applies depends only on which cells $\beta$ contains, not on $P$. As $q_a$ ranges over $[0, r_a]$, the attainable values of $p\, s(a) = F_a + \lambda_a$ sweep $[F_a, F_a + p'_a]$, and similarly for $b$. Always

$$F_a + p'_a \;\geq\; P(\beta, X{=}b) \;\geq\; F_b, \qquad F_b + p'_b \;\geq\; F_a, \tag{7}$$

and $F_a + p'_a = F_b$ holds exactly when $\beta$ contains no success cell of $a$, no failure cell of $b$, and no cell of any third action.

*Case 1:* $p'_b = 0$, i.e. $\beta \subseteq \{X = b\}$. Then $F_a = 0$, $p'_a = p < r_a$, and $s(b) = F_b/p =: f$ is the same for all compatible models. If $0 < f < 1$, any $q_a \in (fp,\ fp + r_a - p)$ gives $\min \Lambda_a < fp < \max \Lambda_a$, so (6) holds. If $f = 0$, then on $\{0 < q_a < r_a - p\}$ we have $\min \Lambda_a = 0 < \max \Lambda_a$, so $I_{ab} = [0, h]$ with $h > 0$. If $f = 1$, then on $\{p < q_a < r_a\}$ we have $\max \Lambda_a = p > \min \Lambda_a$, so $I_{ab} = [-h, 0]$ with $h > 0$.

*Case 2:* $p'_b > 0$, $F_a + p'_a > F_b$ *and* $F_b + p'_b > F_a$. The open intervals $(F_a, F_a + p'_a)$ and $(F_b, F_b + p'_b)$ then overlap; pick $t$ in both. With $v := t - F_a \in (0, p'_a)$, any $q_a \in (v,\ v + r_a - p'_a)$ gives $F_a + \min \Lambda_a < t < F_a + \max \Lambda_a$. If $p'_b < r_b$, choose $q_b$ in the same way. If $p'_b = r_b$, then $\Lambda_b = \{q_b\}$, and we choose $q_b := t - F_b \in (0, r_b)$. In both cases $F_b + \min \Lambda_b \le t \le F_b + \max \Lambda_b$, so (6) holds.

*Case 3:* $p'_b > 0$ *and* $F_a + p'_a = F_b$. Then $\beta$ contains only failure cells of $a$ and only success cells of $b$, and no other cells. Hence $p\, s(a) = \lambda_a \le p'_a = F_b \le p\, s(b)$. On $\{p'_a < q_a < r_a\} \cap \{0 < q_b < r_b - p'_b\}$ both a tie ($\lambda_a = p'_a$, $\lambda_b = 0$) and strictly $a \prec b$ ($\lambda_a < p'_a$) are attainable. This set is nonempty: $p'_a < r_a$ by the choice of $a$, and $p'_b < r_b$ because the success cells of $a$ are not in $\beta$.

*Case 4:* $p'_b > 0$ *and* $F_b + p'_b = F_a$. This is Case 3 with the roles of success and failure exchanged: $\beta$ contains only success cells of $a$ and only failure cells of $b$, $p\, s(b) \le F_a \le p\, s(a)$, and on $\{p'_b < q_b < r_b\} \cap \{0 < q_a < r_a - p'_a\}$ both a tie and strictly $b \prec a$ are attainable. This set is nonempty because $p'_a < r_a$, and $p'_b < r_b$ since the failure cells of $a$ are not in $\beta$.

Cases 3 and 4 cannot hold together, since then $\beta$ would be empty. By (7) the four cases cover all possibilities. In each case the exhibited set is open, nonempty, and consists of points of non-collapse by Lem. 4. □

**Remark 3** (The non-collapse set need not be open)**.** *Take* $\boldsymbol{V} = \{X, Y\}$, $k = 2$, *the uniform table* $P(x, y) = \frac{1}{4}$ *and harm* $\beta = \{X = 0, Y = 0\}$, *so* $p = \frac{1}{4}$, $r_1 = \frac{1}{2}$ *and* $q_1 = \theta_1 - \frac{1}{4}$. *At* $\theta_1 = \frac{1}{2}$ *we have* $\Lambda_1 = [0, \frac{1}{4}]$*: a tie and strictly* $0 \prec 1$ *are both attainable. At* $\theta_1 = \frac{1}{2} + \epsilon$ *we have* $\min \Lambda_1 = \epsilon > 0$, *so every compatible agent has strictly* $0 \prec 1$, *and* $\varphi_\beta$ *collapses. By Lem. 3, with no contexts, these two Layer-2 theories differ only in* $\theta_1$, *so they are arbitrarily close in* $T$. *This is why Thm. 4 asserts that the non-collapse set contains an open set, rather than that it is open.*

**Example 5** (Ambiguity vs identification)**.** Under the same uniform table with $\beta = \{X = 0, Y = 0\}$: if $\theta_1 = \frac{2}{5}$, then PS $\in [0, \frac{3}{5}]$, and Layer 2 does *not* fix $\preceq^\varphi$. If instead $\theta_1 = \frac{3}{5}$, then PS $\ge \frac{2}{5} > 0$, so every compatible model has strictly $0 \prec 1$. Both outcomes occur on open sets of Layer-2 theories, as in Ex. 1. In Ex. 3 ($k = 3$, $\beta = \{X = 0, Y = 0\}$, $\theta_1 = \theta_2 = \frac{1}{2}$), both actions $1, 2$ have $F = 0$, $p' = p = \frac{1}{6}$, $r = \frac{2}{3}$ and $q = \frac{1}{3}$, so $\Lambda_1 = \Lambda_2 = [0, p]$ and $I_{12} = [-1, 1]$.

### D.4 Proof of the many-actions theorem

*Proof of Thm. 5. Strategy.* When no action dominates, each action is taken by only a small share of units, so the range of each score is close to a simple function of its success probability $\theta_x$ alone, with width bounded below. Two actions with nearly equal $\theta$'s then have overlapping ranges, so the ranking is not fixed. Since the $\theta$'s are drawn independently, a long list of actions contains such a near-tie with probability tending to one (a pigeonhole argument). In case (a) the scores can even take every value in $[0, 1]$, so any two actions whose $\theta$'s are bounded away from 0 and 1 suffice.

Write $\delta_k := \max_x P^{(k)}(X{=}x)$, which tends to 0, and $K := [\varepsilon, 1 - \varepsilon]$. Let $N_k$ be the number of actions (other than $x_0$ in case (a)) with $\theta_x \in K$. Under $\mu_k$, $N_k$ is a sum of independent indicators, each equal to one with probability at least $\rho$, so $\mu_k(N_k \le n) \to 0$ for every fixed $n$. Take $k$ large enough that $\delta_k < \varepsilon/2$.

*(a)* Let $x \ne x_0$ with $\theta_x \in K$. Since $\beta \subseteq \{X = x_0\}$, $F_x = 0$ and $p'_x = p \le P(X{=}x_0) \le \delta_k$. Moreover $q_x \ge \theta_x - \delta_k \ge \varepsilon - \delta_k \ge p$ and $q_x \le \theta_x \le 1 - \varepsilon \le 1 - 2\delta_k \le r_x - p$. Hence $\Lambda_x = [0, p]$, and $s(x)$ can take any value in $[0, 1]$. If $N_k \ge 2$, two such actions give $I_{ab} = [-1, 1]$, and $\varphi$ does not collapse by Lem. 4. So $\mu_k(\varphi \text{ collapses}) \le \mu_k(N_k \le 1) \to 0$.

*(b)* Let $p = P(\beta) \in [p_0, 1 - p_0]$ and put $L(\theta) := \max\{0, \theta - (1 - p)\}$ and $H(\theta) := \min\{p, \theta\}$. These change by at most $|\theta - \theta'|$ when $\theta$ changes to $\theta'$, and $H(\theta) - L(\theta) = \min\{p, 1 - p, \theta, 1 - \theta\} \ge w := \min\{p_0, \varepsilon\}$ on $K$. For every action, $F_x \le \delta_k$ and $|p'_x - p|, |r_x - 1|, |q_x - \theta_x| \le \delta_k$. Hence the range $[F_x + \min \Lambda_x, F_x + \max \Lambda_x]$ of $p\, s(x)$ has endpoints within $4\delta_k$ of $L(\theta_x)$ and $H(\theta_x)$. Take $k$

also large enough that $8\delta_k < w/2$. For $a, b$ with $\theta_a, \theta_b \in K$ and $|\theta_a - \theta_b| \le w/4$,

$$F_a + \min \Lambda_a \;\le\; L(\theta_b) + \tfrac{w}{4} + 4\delta_k \;\le\; H(\theta_b) - \tfrac{3w}{4} + 4\delta_k \;<\; H(\theta_b) - 4\delta_k \;\le\; F_b + \max \Lambda_b,$$

and the same holds with $a, b$ exchanged, so (6) holds. Cover $K$ by $n := \lceil 4/w \rceil$ intervals of length at most $w/4$. If $N_k > n$, two of the actions with $\theta_x \in K$ fall in the same interval, and $\varphi$ does not collapse. Hence $\mu_k(\varphi \text{ collapses}) \le \mu_k(N_k \le n) \to 0$. □

### D.5 Proofs of the corollaries

*Proof of Cor. 3. Strategy.* We choose experimental success probabilities that realize the prescribed rankings and are all interior, and then apply the harm construction from the proof of Thm. 3(b), which changes only the counterfactual part of the model.

Let $\beta = \{\boldsymbol{C} = \boldsymbol{c}, X = x_0, Y \notin G\}$ be the harm event. Choose a full-support $P$ with $P(X{=}x, Y{\in}G \mid \boldsymbol{c}') = P(X{=}x, Y{\notin}G \mid \boldsymbol{c}') = \frac{1}{2k}$ for every $x$ and $\boldsymbol{c}'$. Then $r_x = 1 - \frac{1}{k}$ for every $x$, $p_F := P(\beta \mid \boldsymbol{c}) = \frac{1}{2k}$, and each $\theta_x(\boldsymbol{c}')$ must lie in $[\frac{1}{2k}, 1 - \frac{1}{2k}]$ by (5). For each context $\boldsymbol{c}'$, let $\ell_{\boldsymbol{c}'}(x) \in \{0, \dots, k-1\}$ be the position of $x$ in $R_{\boldsymbol{c}'}$ (tied actions share a position), and set $\theta_x(\boldsymbol{c}') := v + \eta\, \ell_{\boldsymbol{c}'}(x)$, where $v$ is the midpoint of $(\frac{1}{2k}, 1 - \frac{1}{k})$ and $\eta > 0$ is small enough that all values stay in this interval, which is nonempty for $k \ge 2$. By Con. 2, some model has observational distribution $P$ and these success probabilities. Since $\mathbb{E}[\mathcal{U}(Y_x) \mid \boldsymbol{c}'] = \theta_x(\boldsymbol{c}')$, this Layer-2 theory induces $R_{\boldsymbol{c}'}$ in every context. For any $x_1 \ne x_0$, $q_{x_1} = \theta_{x_1}(\boldsymbol{c}) - \frac{1}{2k} \in (0,\, 1 - \frac{3}{2k}) = (0,\, r_{x_1} - p_F)$, so the first case of the proof of Thm. 3(b) yields two models realizing this theory that differ on $\preceq^{\varphi}$. □

*Proof of Cor. 4. Strategy.* As in the previous proof, but now we only need the prescribed action to be the unique best one in each context. We give it a high success probability and all other actions a success probability just above their lower bound, keeping everything interior.

Take any full-support $P$ and a small $\epsilon > 0$. All probabilities below are conditional on the context under discussion. In a context $\boldsymbol{c}'$ other than the harm context $\boldsymbol{c}$, set $\theta_{\pi(\boldsymbol{c}')} := P(X{=}\pi(\boldsymbol{c}'), Y{\in}G) + r_{\pi(\boldsymbol{c}')} - \epsilon$ and $\theta_x := P(X{=}x, Y{\in}G) + \epsilon$ for $x \ne \pi(\boldsymbol{c}')$. For small $\epsilon$, $\pi(\boldsymbol{c}')$ is the unique maximizer, because $P(X{=}x, Y{\in}G) < P(X{=}x) \le r_{\pi(\boldsymbol{c}')}$ for $x \ne \pi(\boldsymbol{c}')$. In the harm context, let $p_F := P(X{=}x_0, Y{\notin}G)$. If $\pi(\boldsymbol{c}) \ne x_0$, let $x_1 := \pi(\boldsymbol{c})$, set $\theta_{x_1} := P(X{=}x_1, Y{\in}G) + r_{x_1} - p_F - \epsilon$, and $\theta_x := P(X{=}x, Y{\in}G) + \epsilon$ for $x \ne x_1$. Since $r_{x_1} - p_F = P(X{=}x_0, Y{\in}G) + \sum_{x \notin \{x_0, x_1\}} P(X{=}x) \ge P(X{=}x, Y{\in}G)$ for every $x \ne x_1$, and $P(X{=}x_1, Y{\in}G) > 0$, action $x_1$ is the unique maximizer for small $\epsilon$. If $\pi(\boldsymbol{c}) = x_0$, pick any $x_1 \ne x_0$ and proceed as in the other contexts, with $x_0$ as the maximizer. In both cases every $\theta$ is interior and $q_{x_1} \in (0, r_{x_1} - p_F)$. Realize $P$ and these success probabilities by Con. 2. The first case of the proof of Thm. 3(b) yields two models realizing this Layer-2 theory, hence with policy $\pi$, that differ on $\preceq^{\varphi}$. □

**Corollary 5** (No universal recovery)**.** *There is no map $F$ from observational distributions to rankings such that $\preceq^{\varphi}_{\mathcal{A}} = F(P_{\mathcal{A}}(\boldsymbol{V}))$ for every agent and a fixed $\varphi \in \Phi_{\text{int}}$; any candidate fails at every full-support distribution. Likewise, for every $\varphi_\beta$ in case (ii) of Thm. 4, any map from Layer-2 theories to rankings fails on a nonempty open set.*

*Proof.* At every such distribution there are two agents with the same lower-layer theory and different rankings (Thm. 6 and the proof of Thm. 4), and $F$ cannot equal both. The claim concerns a single proposition, as in Def. 8. It does not say that collapse is unverifiable: for a fixed $\varphi$ the collapse set can also contain an open set (Exs. 1 and 5). □

**Remark 4** (Scope)**.** *Three extensions are not covered by the proofs above. (1)* Knowledge of the causal graph. *The competing models range over all SCMs that respect the temporal order of Assump. 2. If the observer also knows the causal graph, for instance an exclusion restriction such as "$X$ affects $Y$ only through a mediator", re-paired models may violate it, and the ranges $\Lambda_x$ can shrink. Such cases need an analysis for each graph, as in partial identification (Zhang et al., 2022). (2)* Non-binary utilities. *For a real-valued utility with more than two values, Lem. 2 extends by re-pairing so that the lowest- or highest-utility outcomes under $x$ fall on $\beta$, but the case analysis behind Thm. 3(b) and Thm. 4 has not been carried out. (3)* Additional interventions $\alpha$. *If $\beta$ is determined by the context and*

*the contexts are roots with independent exogenous variables, the proposition reduces to Layer 2 and collapses (Appendix B); without that condition this case is also open. If $\alpha$ changes the context and $\beta$ involves $X$ or $Y$, e.g. $\mathbb{E}[\mathcal{U}(Y_{x,\mathbf{c}'}) \mid \boldsymbol{C} = \mathbf{c}, X = x', Y = y']$ with $\mathbf{c}' \neq \mathbf{c}$, no unit in $\beta$ is observed in context $\mathbf{c}'$, and we expect wide ranges as in case (a) of Thm. 5. A proof would need to control how different actions in context $\mathbf{c}'$ share the same observed data, and is left open.*