COUNTERFACTUAL REALIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

It is commonly believed that, in a real-world environment, samples can only be drawn from observational and interventional distributions, corresponding to Layers 1 and 2 of the *Pearl Causal Hierarchy*. Layer 3, representing counterfactual distributions, is believed to be inaccessible by definition. However, Bareinboim, Forney, and Pearl (2015) introduced a procedure that allows an agent to sample directly from a counterfactual distribution, leaving open the question of what other counterfactual quantities can be estimated directly via physical experimentation. We resolve this by introducing a formal definition of *realizability*, the ability to draw samples from a distribution, and then developing a complete algorithm to determine whether an arbitrary counterfactual distribution is realizable given fundamental physical constraints, such as the inability to go back in time and subject the same unit to a different experimental condition. We illustrate the implications of this new framework for counterfactual data collection using motivating examples from causal fairness and causal reinforcement learning. While the baseline approach in these motivating settings typically follows an interventional or observational strategy, we show that a counterfactual strategy provably dominates both.

1 INTRODUCTION

The *Pearl Causal Hierarchy*, or PCH, is an important recent milestone in our understanding of causality (Pearl & Mackenzie, 2018; Bareinboim et al., 2022). The three layers of the PCH represent the distinct regimes of *seeing*, *doing*, and *imagining*, with regard to an environment. Consider an environment involving a decision variable X and an outcome Y. Layer 1 ($\mathcal{L}_1$) represents *observational* distributions, such as $P(Y \mid x)$. Layer 2 ($\mathcal{L}_2$) represents *interventional* distributions, such as $P(Y; do(x))$, using the $do()$ operator. Layer 3 ($\mathcal{L}_3$) represents *counterfactual* distributions dealing with conflicting realities, such as $P(Y_x \mid x', y')$: the distribution of Y had X been fixed as x, given that X, Y were in fact naturally observed to be x', y'. Higher layers subsume lower ones, but are underdetermined by them (Ibeling & Icard, 2020; Bareinboim et al., 2022).

Reasoning about $\mathcal{L}_3$-quantities plays a vital role in personalized decision-making (Mueller & Pearl, 2023), analysing a causal effect into direct and indirect pathways (Pearl, 2005; Rubin, 2004), and constructing explanations for decisions, among other topics, in applications such as healthcare (Mueller & Pearl, forthcoming), economics (Li & Pearl, 2019), epidemiology (Robins & Greenland, 1992) etc. Suppose an economist were interested in estimating $P(y_x \mid x')$, an important $\mathcal{L}_3$-quantity called the effect of the treatment on the treated, or ETT (Heckman & Robb Jr., 1985; 1986). One approach to computing such quantities is through *identification* (Pearl, 2000, §3.2.4): leveraging causal knowledge about the environment, typically a causal graph or parametric assumptions, to infer the higher-layer quantity using lower-layer data. This approach fails when the quantity is nonidentifiable, e.g. ETT in the general setting (Shpitser & Pearl, 2009; Correa et al., 2021).

However, another approach uses physical experimentation to attempt to directly draw samples from the relevant distribution, $P(Y_x, X)$ in the case of ETT, and then uses statistical methods to estimate $P(Y_x = y, X = x')$. This approach is only possible if there is some sequence of physical actions by which an agent can measure these random variables simultaneously for a single unit. It is generally believed to be feasible to draw samples only from $\mathcal{L}_1$- and $\mathcal{L}_2$-distributions, the latter by interventions like randomized controlled trials (RCT), à la Fisher (Fisher, 1935), and the former by simply observing the natural behaviour of the system. $\mathcal{L}_3$-distributions like $P(Y_x, X)$ are deemed non-realizable in general, as the potential response Y_x and natural decision X belong to different "worlds". Once a unit naturally adopts decision $X = x'$, Y_x cannot be evaluated in the $do(x)$ regime

for the same unit.[1] However, Bareinboim, Forney & Pearl have shown it is feasible to draw samples from the ETT distribution $P(Y_x, X)$ through a *counterfactual randomization* procedure (Bareinboim et al., 2015; Forney et al., 2017). This leaves open the possibility that other $\mathcal{L}_3$-distributions, say perhaps $P(Y_x, X, Y)$, are also realizable through clever experimental setups, allowing one to estimate important quantities like the probability of sufficiency, $P(y_x \mid y', x')$ (Pearl, 1999).

This brings us to the central question motivating this work: *from which $\mathcal{L}_3$-distributions is it possible to draw samples given fundamental physical constraints like the inability to travel back in time and subject the original unit to a different experimental condition?* We resolve this open question with a rigorous formal treatment of the *realizability of an $\mathcal{L}_3$-distribution* (Def. 3.4).

Our main contributions in this work are as follows:

- In Sec. 2 we introduce a physical procedure called *counterfactual randomization* (Def. 2.3) by which an agent can gather counterfactual data, subsuming previous similar notions.
- In Sec. 3 we develop the **CTF-REALIZE** algorithm (Algo. 1) to determine whether an $\mathcal{L}_3$-distribution is physically realizable. We prove the algorithm is complete (Thm. 3.5), and derive important corollaries characterizing realizable distributions (Cors. 3.7,3.8). For instance, we show that our main result generalizes an influential notion in the causal inference literature, known as the *fundamental problem of causal inference* (Holland, 1986).
- In Sec. 4 we discuss important practical implications of counterfactual realizability. The traditional route of computing $\mathcal{L}_3$-quantities through identification often fails. Our work suggests opportunities for novel experiment-design ideas to directly estimate these quantities, as illustrated through Examples 1,2 and 3. More concretely,
  - In Sec. 4.1, we describe an application in causal fairness, where the naive approach of constraining a classifier using an interventional ($\mathcal{L}_2$) fairness metric fails to prevent disparities in outcomes across groups, but where a counterfactual ($\mathcal{L}_3$) approach works.
  - In Sec. 4.2, we show how counterfactual randomization can be used to improve RL algorithms. The baseline approach in a multi-arm bandit setting is to use allocation procedures (e.g., UCB, EXP3, Thompson Sampling) to discover which arm x optimizes the expected outcome $\mathbb{E}[Y; do(x)]$, which is an interventional ($\mathcal{L}_2$) strategy (Sutton & Barto, 1998; Lattimore & Szepesvári, 2020). It turns out there are provably superior strategies (w.r.t expected outcome) based on directly optimizing counterfactual ($\mathcal{L}_3$) objectives, as we demonstrate in Example 3. We prove optimality of our proposed strategy in a bandit setting with a generic causal template (Thm. F.2, Cor. F.3).

Finally, Sec. 5 discusses important themes, future directions, and the limitations of our work. Proofs and details of simulations are included in Appendices.

Preliminaries. We denote variables by capital letters, X, and values by small letters, x. Bold letters, $\mathbf{X}$, are sets of variables and $\mathbf{x}$ sets of values. $P(\mathbf{x})$ is shorthand for $P(\mathbf{X} = \mathbf{x})$. $\mathbb{1}[.]$ is the indicator function. We use *Structural Causal Models* (SCM) to describe the generative process for a system of interest (Bareinboim et al., 2022, Def. 1)(Pearl, 2000). An SCM $\mathcal{M}$ is a tuple $\langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$. $\mathbf{V}$ is the set of observable variables. $\mathbf{U}$ is the set of unobservable variables exogenous to the system, distributed according to $P^{\mathcal{M}}(\mathbf{U})$. $\mathcal{F} = \{f_V\}$ is a set of functions s.t. each f_V causally generates the value of $V \in \mathbf{V}$ as $V \leftarrow f_V(\mathbf{U}_V, \mathbf{Pa}_V)$, where $\mathbf{U}_V \subseteq \mathbf{U}$ and $\mathbf{Pa}_V \in \mathbf{V} \setminus V$. Each $\mathcal{M}$ induces a *causal diagram* $\mathcal{G}$ (Bareinboim et al., 2022, Def. 13), which is a graph containing a vertex for each $V \in \mathbf{V}$, a directed edge from each node in $\mathbf{Pa}_V$ to V, and a bidirected edge between V, V' if $\mathbf{U}_V, \mathbf{U}_{V'}$ are not independent. Given a graph $\mathcal{G}$, $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{W}}}$ is the result of removing edges coming into variables in $\mathbf{X}$, and edges coming out of $\mathbf{W}$. We use standard terminology like parents, descendants of a node (see App. A). Our treatment is limited to *recursive* SCMs, which implies acyclic diagrams, with finite discrete domains over $\mathbf{V}$. The $do(\mathbf{x})$ operator indexes a sub-model $\mathcal{M}_{\mathbf{x}}$ where the functions generating variables $\mathbf{X}$ are replaced with constant values $\mathbf{x}$. A variable $Y \notin \mathbf{X}$ evaluated in this regime is called a *potential response*, denoted $Y_{\mathbf{x}}$. $(\mathbf{W}_\star = \mathbf{w})$ denotes an arbitrary counterfactual event, e.g. $(Y_x = y \wedge Y_{x'} = y' \wedge X = x'')$.

[1] E.g., "The problem with counterfactuals like $[P(Y_x \mid x')]$ is [that] ... we simply cannot perform an experiment where the same person is both given and not given treatment." (Shpitser & Pearl, 2007) Also, "By definition, one can never observe [counterfactuals], nor assess empirically the validity of any modeling assumptions made about them..." (Dawid, 2000)

The probability of such an event is given by the $\mathcal{L}_3$-valuation (Bareinboim et al., 2022, Def. 7): $P^{\mathcal{M}}(\mathbf{W}_\star = \mathbf{w}) = \sum_{\mathbf{u}} \left(\prod_{W_{\mathbf{t}} \in \mathbf{W}_\star} \mathbb{1}[W_{\mathbf{t}}(\mathbf{u}) = w] \right) P^{\mathcal{M}}(\mathbf{u})$, with w taken from $\mathbf{w}$.

2 Data-collection procedures

In this section, we define a procedure, *counterfactual randomization*, that extends the scope of traditional *Fisherian* experimentation (discussed below). Consider a system of interest modeled by unknown SCM $\mathcal{M}$. Interventions and counterfactual events are typically defined in terms of *symbolic* operations on $\mathcal{M}$. To conceptually separate this from the *physical* constraints experienced by an agent (natural or artificial), we define the following physical actions that an agent can perform in the system. These are simply the physical counterparts to symbolic procedures.

We call each discrete episode of the system's behaviour a *unit*. Examples of units are patients in a clinical trial, neighbourhoods in a social science experiment, rounds played on a slot machine etc. We index units WLOG by $i = 1, 2, 3...$, which constitute a target population in the system.

Definition 2.1 (Physical actions)**.** (1) $\textsc{Select}^{(i)}$: randomly choosing, without replacement, a unit i from the target population, to observe in the system; (2) $\textsc{Read}(V)^{(i)}$: measuring the realized feature $V^{(i)}$ of unit i, produced by a causal mechanism $f_V \in \mathcal{F}$ operating on i; (3) $\textsc{Rand}(X)^{(i)}$: erasing and replacing i's natural mechanism f_X for a decision variable X with an enforced value drawn from a randomizing device having support over $\text{Domain}(X)$. ■

$\textsc{Read}(V)^{(i)} = v$ and $\textsc{Rand}(X)^{(i)} = x$ are also overloaded to refer to the values read and enforced, respectively. $\textsc{Rand}(X)^{(i)}$ is the standard Fisherian randomization of a decision variable X, corresponding to the symbolic procedure of a *stochastic* intervention on X (Correa & Bareinboim, 2020).[2] As $\textsc{Rand}(X)^{(i)}$ erases the unit i's natural decision, $\textsc{Read}(X)^{(i)}$ will yield the value randomly assigned to unit i. The discovery of this procedure marked an important achievement in the history of science and experiment-design (Fisher, 1925; 1935). Since the use of a randomizing device eliminates by design any confounding between the assigned decision and the unit's latent attributes $\mathbf{U}^{(i)}$, it allows researchers to estimate causal effects.

It is evident that the actions in Def. 2.1 are sufficient for an agent to physically draw samples from any $\mathcal{L}_1$- or $\mathcal{L}_2$-distribution, as discussed in App. C.3. Until recently, it was generally presumed these were the only physical actions possible on units in a system. However, we discuss some important extensions of experimental capabilities next.

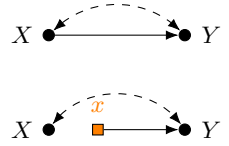


Figure 1: (Top) Causal diagram with decision variable X; (Bottom) Procedure of randomizing the actual decision without erasing the unit's natural decision.

Counterfactual data-collection procedures. In an early work from the causal reinforcement learning literature, Bareinboim, Forney & Pearl describe an experimental setting in which it is possible to both randomize a unit's actual decision, and also record the natural decision the unit *would have normally* taken (Bareinboim et al., 2015; Forney et al., 2017). Subsequently, this procedure has been used to establish benchmarks in counterfactual decision making (Zhang & Bareinboim, 2022). These settings involve an agent introspecting to gauge their natural choice, or otherwise revealing their natural choice by some indication, e.g. physical gestures prior to decision-time. Importantly, this form of randomization does not erase the unit's natural choice of decision variable X, as schematically illustrated in Fig. 1.

Building on the idea, we formalize this into a more general extension of the agent's capabilities: the ability to intervene on a variable X's value *as perceived* by its causal children. To illustrate this, consider the $\mathcal{L}_3$-quantity known as natural direct effect, or NDE, which is used in mediation analysis to measure the effect of X on Y via a "direct" path, as opposed to an "indirect" path via a mediator Z (Pearl, 2001) – highly relevant in several fields, as discussed in Sec. 1. The NDE is generally considered as identifiable from experimental data only under certain conditions (Pearl, 2005; Correa et al., 2021). The following example details an experiment design where it is possible to compute the NDE even when these identification conditions are not met, by randomizing the *perception* of X.

[2] If the device used for enforcing the value of X is a constant function, this action simply becomes $\textsc{Write}(X : x)^{(i)}$, corresponding to the atomic intervention $do(x)$. See Preliminaries in Sec. 1.

Example 1 (Mediation analysis). A computer vision company's tool is being evaluated for an automated speeding ticket system that uses footage from traffic cameras. But the government's audit team has a concern: it is possible the model is trained on footage with a strong correlation between the color of the car and speeding (perhaps due to color preference of different socioeconomic neighbourhoods), and unfairly penalizes certain car colors.

This amounts to a hypothesis that X (car's color) affects Y (AI decision to issue a ticket) via a direct path as opposed to the indirect path via Z (speeding). The indirect path describes the causal effects of, say, how pedestrians and other drivers react to a red car and affect its speeding. This hypothesis is true iff NDE is measured to be non-0, where NDE is defined as the following expression: $\text{NDE}_{x,x'}(y) = P(y_{x'Z_x}) - P(y_x)$ (Pearl, 2001). The second term, $P(y_x)$, can be estimated from a Fisherian randomization of X (say, an experiment recruiting drivers and assigning them random cars). Inconveniently, the first term, $P(y_{x'Z_x})$, is nonidentifiable for Fig. 2(a), even using RCT data. So it is unclear how to make progress with this hypothesis test.

However, the audit team recognizes there exists a special mediator, viz. the features W in the video which reveal the car's color to the model (say, RGB values of pixels in the video frames). They use standard video-editing tools to randomly swap the color of the car in the footage. By randomly assigning a particular car $W \leftarrow$ red, they are able to affect the mechanism f_Y's *perception* of X:

$$
\begin{aligned}
&P(Y_{W=\text{red}} \mid X = \text{blue}) && \text{est. from } \mathcal{L}_2 \text{ data} && (1)\\
=&P(Y_{W=\text{red},Z} \mid X = \text{blue}) && Z: \text{natural value} && (2)\\
=&P(Y_{W=\text{red},Z_{X=\text{blue}}} \mid X = \text{blue}) && \text{consistency property} && (3)\\
=&P(Y_{X=\text{red},Z_{X=\text{blue}}} \mid X = \text{blue}) && \text{Def. 2.2}, X \equiv W && (4)\\
=&P(Y_{X=\text{red},Z_{X=\text{blue}}}) && \text{d-separation} && (5)
\end{aligned}
$$

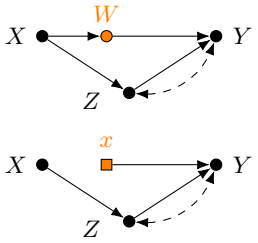


Figure 2: (a) "Expanded" diagram for Example 1, where W is *counterfactual mediator* for X; (b) Randomizing the value of X as perceived by Y.

Eq. 4 is justified because W controls Y's perception of X given a fixed z (formalized in Lemma D.4). Thus, they are able to directly sample from the $\mathcal{L}_3$-distribution $P(Y_{x'Z_x}, X)$ via a physical procedure, and use identification rules to obtain $P(y_{x'Z_x})$. Using the formula for NDE, they can evaluate whether a car's color has a direct effect on the odds of getting a speeding ticket. ■

Here, one is able to randomize X as perceived by one of its children, by leveraging the variable W (RGB values) that fully encodes information about X (color) and mediates its effect on Y. We capture this intuition with the following (informal) definition.

Definition 2.2 (Counterfactual mediator (informal))**.** We call W a *counterfactual mediator* of X w.r.t $Y \in Ch(X)$ if the value of X can be retrieved from W by the mechanism generating Y. ■

Other examples of interventions on perceived attributes via counterfactual mediators include changing details on a job application (name, pronouns, keywords) to simulate a perceived alternate demographic identity (Bertrand & Mullainathan, 2003), or editing specific portions of text input to a language model (Feder et al., 2022). Randomizing perception has been discussed in Pearl et al. (2016, §4.4.4). For a detailed discussion of the causal semantics of intervening on perceptions, and the related literature, see (Plecko & Bareinboim, 2024, App. D.1). For the interested reader, we provide a rigorous treatment in App. D, including a formal Def. D.2 of a counterfactual mediator.

This important extension to experimental capabilities is captured in the following definition of a new physical action that an agent might be able to perform in an environment.

Definition 2.3 (Counterfactual (ctf-) randomization)**.** $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$: fixing the value of X *as an input to* the mechanisms generating $\mathbf{C} \subseteq Ch(X)_{\mathcal{G}}$ using a randomizing device having support over Domain(X), for unit i, given causal diagram $\mathcal{G}$. ■

The key differences between the Fisherian $\textsc{Rand}(X)^{(i)}$ and $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$ are (1) $\textsc{ctf-Rand}$ does not erase the unit i's natural decision $X^{(i)}$; and (2) while $\textsc{Rand}$ affects all children of X, $\textsc{ctf-Rand}$ does not affect $Ch(X) \setminus \mathbf{C}$. $\textsc{ctf-Rand}$ can only be enacted under certain structural conditions, viz., either in environments which permit the measurement of a unit's natural decision while simultaneously randomizing the actual decision (Bareinboim et al., 2015), or where counterfactual mediators can be used to alter X as perceived by a subset of children. Whether the agent is indeed able to perform this action thus depends on the specific experimental setting.

Note: Def. 2.3 implies that it is possible to physically perform multiple randomizations involving the same variable X on a single unit i, with each intervention affecting a different subset of children. Further, CTF-RAND may only be performed w.r.t a graphical child variable; it is not possible to bypass a child and directly affect a descendant's perception of X.

3 COUNTERFACTUAL REALIZABILITY

Given the possibility of performing ctf-randomization (Def. 2.3), we are interested in knowing which $\mathcal{L}_3$-distributions can be accessed directly by experimentation. In this section, we discuss the constraints imposed by nature on an agent. We then formally define *realizability* and develop a complete algorithm to determine whether an $\mathcal{L}_3$-distribution is realizable.

The most basic constraints experienced by the agent (natural or artificial) are physical. Each mechanism $f_V \in \mathcal{F}$ represents some physical process that transforms a unit i according to the laws of nature. For instance, taking a drug, X, produces a side effect in the patient, Y, by a biochemical reaction $f_Y(X, U_Y)$, which depends on the drug and the patient's latent health condition, U_Y. Once patient i has been subjected to mechanism f_Y under $X = x$, there appears to be no way to go back in time and subject the same patient to mechanism f_Y under $X = x'$. Even if technologically feasible to reverse the process (e.g., by taking an antidote to the drug), the latent factors $\mathbf{U} = \mathbf{u}$ might have changed after the experiment (e.g., the patient could have developed tolerance to the drug). Repeating the experiment on this patient is tantamount to testing a *new* unit with unknown latent features $\mathbf{U} = \mathbf{u}'$.[3] This observation is made more formal through the following assumption.

Assumption 3.1 (Fundamental constraint of experimentation (FCE))**.** A unit i in the target population can physically undergo a causal mechanism $f_V \in \mathcal{F}$ at most once. ■

Remark 3.2. The FCE assumption entails that a unit i can only be submitted to a particular mechanism $f_V(\mathbf{Pa}_V, \mathbf{U}_V)$ under a single set of experimental conditions, received as input to f_V. By implication, the physical actions in Defs. 2.1, 2.3 can only be performed at most once per unit i. ■

Once unit i has been subjected to f_V, it is not possible to re-run f_V with differently fixed inputs. $\text{READ}(V)^{(i)}$ thus only yields one value for i. Although ctf-randomization permits multiple interventions involving the same variable X, each such intervention can only be performed once, since it impacts different child mechanisms that can each only occur once for unit i. We also assume that the agent can only perform the physical actions in Defs. 2.1, 2.3, up to isomorphism.

Definition 3.3 (I.i.d sample)**.** Given an $\mathcal{L}_3$-distribution $Q = P(\mathbf{W}_\star)$ and a sequence of physical actions $\mathcal{A}^{(i)}$ performed on unit i in an environment modeled by SCM $\mathcal{M}$, producing a vector of realized values $\mathbf{W}_\star^{(i)} = \mathbf{w}$ for the variables in $\mathbf{W}_\star$, the vector is said to be an *i.i.d sample* from Q if $P^{\mathbb{C}}(\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{W}_\star = \mathbf{w}), \forall \mathbf{w}$, where $P^{\mathbb{C}}$ is the probability measure over the beliefs of the acting agent $\mathbb{C}$, and the l.h.s is the probability of physical actions $\mathcal{A}^{(i)}$ producing the vector $\mathbf{w}$ when performed on some unit i. ■

Definition 3.4 (Realizability)**.** Given a causal diagram $\mathcal{G}$ and the set of physical actions $\mathbb{A}$, an $\mathcal{L}_3$-distribution $P(\mathbf{W}_\star)$ is *realizable given* $\mathbb{A}$ *and* $\mathcal{G}$ iff there exists a sequence of actions $\mathcal{A}$ from $\mathbb{A}$ by which an agent can draw an i.i.d sample (Def. 3.3) from $P^{\mathcal{M}}(\mathbf{W}_\star)$, for any $\mathcal{M} \in M(\mathcal{G})$, the class of SCMs compatible with $\mathcal{G}$. ■

We emphasize the *distinction between realizability and identifiability*. Identifiability (Pearl, 2000, Def. 3.2.3) from $\mathcal{G}$ states that a distribution (say, $P(\mathbf{v}; do(x))$) can be uniquely computed from the available data (say, $P(\mathbf{v})$) for any SCM compatible with the assumptions in $\mathcal{G}$. Realizability of a distribution states that it is physically possible for an agent to actually gather data samples according to this distribution.

We next develop an algorithm to decide whether a distribution is realizable. As an intuition pump, suppose that an agent is able to perform $\text{CTF-RAND}(V \to C), \forall V, C \in Ch(V)$, w.r.t an input causal diagram, and wants to obtain samples from $P(Z_x, W_t)$. Consider the diagram $\mathcal{G}_2$ in Fig. 3. By performing $\text{CTF-RAND}(T \to W)$ and $\text{CTF-RAND}(X \to Z)$, the distribution is realizable. However, suppose the input diagram is $\mathcal{G}_1$. A necessary condition to measure Z_x for a unit is for mechanism f_A

[3] In the philosophy of science literature, similar ideas have been discussed under the topic of the temporal asymmetry of causation (Reichenbach, 1956, §III-IV).

Algorithm 1 CTF-REALIZE

1: **Input:** $\mathcal{L}_3$-distribution $Q = P(\mathbf{W}_\star)$; causal diagram $\mathcal{G}$; action set $\mathbb{A}$
2: **Output:** I.i.d sample $\mathbf{W}_\star^{(i)}$ from Q; FAIL if Q is not realizable given $\mathcal{G}, \mathbb{A}$
3: Fix a topological ordering Top($\mathcal{G}$)
4: $\textsc{Select}^{(i)}$ for a new unit i
5: **for** V in order Top($\mathcal{G}$) **do**
6: $\quad$ $\text{INT}_V \leftarrow \emptyset$ {Interventions for V}
7: $\quad$ $\text{OUTPUT}_V \leftarrow \emptyset$ {Index in output vector}
8: $\quad$ **for** each term $W_\mathbf{t}$ in expression $\mathbf{W}_\star$ **do**
9: $\quad\quad$ **if** $V \in An(W)_{\mathcal{G}_{\overline{\mathbf{T}}}}$ **and** $V \neq W$ **then**
10: $\quad\quad\quad$ Call **COMPATIBLE**$(V, W_\mathbf{t})$ 2
11: $\quad\quad$ **end if**
12: $\quad\quad$ **if** $V = W$ **then**
13: $\quad\quad\quad$ Add $\{W_\mathbf{t}\}$ to OUTPUT_V
14: $\quad\quad$ **end if**
15: $\quad$ **end for**
16: $\quad$ **for** each {action : tag} $\in \text{INT}_V$ **do**
17: $\quad\quad$ Perform the randomization on unit i
18: $\quad\quad$ If the random-generated value $\neq$ tag, discard the unit and return to Line 4
19: $\quad$ **end for**
20: $\quad$ **for** each $W_t \in \text{OUTPUT}_V$ **do**
21: $\quad\quad$ **if** $\{\textsc{Rand}(V) : .\} \in \text{INT}_V$ **then**
22: $\quad\quad\quad$ Return FAIL
23: $\quad\quad$ **else**
24: $\quad\quad\quad$ Perform $\textsc{Read}(V)^{(i)} = v'$
25: $\quad\quad\quad$ Assign v' to each index $W_t^{(i)}$ in output vector $\mathbf{W}_\star^{(i)} = \mathbf{w}$
26: $\quad\quad$ **end if**
27: $\quad$ **end for**
28: **end for**
29: Return i.i.d sample $\mathbf{W}_\star^{(i)} = \mathbf{w}$

to receive the natural value of T, illustrated in green. While a necessary condition to simultaneously measure W_t is for f_W to receive A_t, which in turn requires f_A to receive a fixed t, shown in red. This conflict in necessary conditions renders the query non-realizable.[4]

This "edge-coloring" intuition is formalized in Algo. 1. The algorithm **CTF-REALIZE** takes as input an $\mathcal{L}_3$-distribution $P(\mathbf{W}_\star)$, a graph $\mathcal{G}$, and a set of physical actions $\mathbb{A}$ the agent is able to perform in the environment (viz., the RAND and CTF-RAND actions which are possible in the environment). It returns an i.i.d sample if the distribution is realizable, and FAIL otherwise.

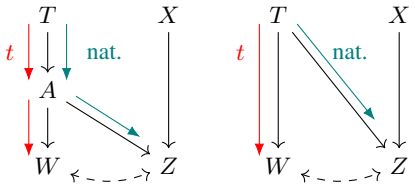


Figure 3: Testing realizability of $P(Z_x, W_t)$ for $\mathcal{G}_1$ (left) and $\mathcal{G}_2$ (right). $\mathcal{G}_1$ yields conflicting requirements.

The algorithm works as follows (a more detailed walk-through is presented in App. B.2): going over each node V in topological order, the inner loops gather the necessary and sufficient conditions needed w.r.t V for realizing each $W_\mathbf{t}$ in the input query $\mathbf{W}_\star$. If there is a conflict in the necessary conditions for evaluating two terms (as we saw for $P(Z_x, W_t)$ in Fig. 3, $\mathcal{G}_1$), the query is non-realizable. The algorithm is fully general and does not make assumptions about the ability to perform any particular interventions. If the agent cannot perform any counterfactual randomization, the algorithm returns FAIL for non-$\mathcal{L}_2$ queries. If the agent cannot perform any interventions at all, the algorithm returns FAIL for non-$\mathcal{L}_1$ queries (we assume the ability to READ all variables). Details about the time and space complexity of Algo. 1 are provided in App. B.3, for the interested reader.

Theorem 3.5 (Correctness and Completeness)**.** *An $\mathcal{L}_3$-distribution $Q = P(\mathbf{W}_\star)$ is realizable given action set $\mathbb{A}$ and causal diagram $\mathcal{G}$ iff the algorithm **CTF-REALIZE**$(Q, \mathcal{G}, \mathbb{A})$ returns a sample.* ■

A further question we may ask is which $\mathcal{L}_3$-distributions are realizable if we assume maximum experimental capabilities, notably, the ability to perform separate ctf-randomization for each child of each variable. Given a causal diagram $\mathcal{G}$, we define the *maximal feasible action set* $\mathbb{A}^\dagger(\mathcal{G})$ as the set containing all of the following actions: $\textsc{Select}^{(i)}$, $\textsc{Read}(V)^{(i)}$, $\forall V$, and $\textsc{ctf-Rand}(X \to C)^{(i)}$, $\forall X$ and $C \in Ch(X)$. $\mathbb{A}^\dagger(\mathcal{G})$ thus gives the agent the most granular interventional capabilities.

Definition 3.6 (Ancestors of a counterfactual (Correa et al., 2021))**.** Given a causal diagram $\mathcal{G}$ and a potential response $Y_\mathbf{x}$, the set of (counterfactual) ancestors of $Y_\mathbf{x}$, denoted $An(Y_\mathbf{x})$, consists of each

[4]To be clear, the input to the algorithm is a graph and an accurate set of actions the agent can perform in the environment. If the graph is per $\mathcal{G}_1$ in Fig. 3, then $\textsc{ctf-Rand}(T \to Z)$ is not possible in this environment. Marginalizing out A and providing graph $\mathcal{G}_2$ as input does not help.

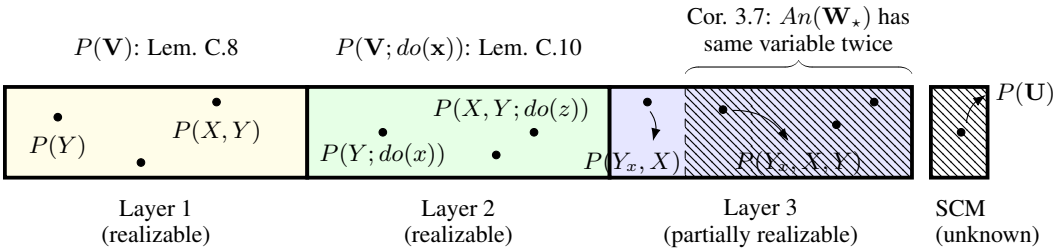


Figure 4: Pearl Causal Hierarchy (PCH) induced by an unknown SCM $\mathcal{M}$. An $\mathcal{L}_3$-distribution is realizable given a graph $\mathcal{G}$ and the maximal feasible action set $\mathbb{A}^\dagger(\mathcal{G})$ iff the ancestor set $An(\mathbf{W}_\star)$ does not contain the same variable under different regimes.

$W_\mathbf{z}$ s.t. $W \in An(Y)_{\mathcal{G}_{\overline{\mathbf{X}}}}$, and $\mathbf{z} = \mathbf{x} \cap An(W)_{\mathcal{G}_{\overline{\mathbf{X}}}}$. For a set $\mathbf{W}_\star$, $An(\mathbf{W}_\star)$ is defined to be the union of the ancestors of each potential response in the set. ■

Corollary 3.7. *An $\mathcal{L}_3$-distribution $Q = P(\mathbf{W}_\star)$ is realizable given causal diagram $\mathcal{G}$ and action set $\mathbb{A}^\dagger(\mathcal{G})$ iff the ancestor set $An(\mathbf{W}_\star)$ does not contain a pair of potential responses $W_\mathbf{t}, W_\mathbf{s}$ of the same variable W under different regimes.* ■

For instance, if $\mathbf{W}_\star = \{Z_x, W_t\}$ w.r.t graph $\mathcal{G}_1$ in Fig. 3, then $An(\mathbf{W}_\star) = \{Z_x, A, T, W_t, A_t\}$, which contains both A, A_t. Thus, $P(\mathbf{W}_\star)$ is not realizable even with maximal experimentation capabilities. In App. B.4, we provide further examples of using the **CTF-REALIZE** algorithm, and the graphical criterion, to demonstrate the realizability of the ETT distribution $P(Y_x, X)$, the non-realizability of the probability of sufficiency distribution $P(Y_x, X, Y)$.

We believe this is an important contribution to causal inference. Cor. 3.7 provides a graphical criterion to delineate how far up the PCH an agent can go via experimental methods, in principle. Often, counterfactuals have been criticized as being hypothetical, untestable, or unscientific assumptions. Our analysis counters this claim, as summarized in Fig. 4.

Corollary 3.8 (Fundamental problem of causal inference (FPCI) (Holland, 1986))**.** *The distribution $Q = P(Y_x, Y_{x'})$ is not realizable given maximal feasible action set $\mathbb{A}^\dagger(\mathcal{G})$, for any causal diagram $\mathcal{G}$, and any variables $X, Y \in Desc(X)$.* ■

The FPCI is an influential notion in the literature, and is often taken as a primitive, or in an axiomatic fashion. We show that it is rather a specific consequence of the more general FCE assumption 3.1, and follows from Thm. 3.5 and Cor. 3.7. By itself, the FPCI does not translate to an operational criterion for determining which $\mathcal{L}_3$-distributions are realizable (Def. 3.4). For instance, it does not clarify that a distribution with potential responses under conflicting regimes like $P(Y_x, Z_{x'})$ may indeed be realizable via counterfactual randomization, as we show in Example 2. It also does not tell us that $P(Z_x, W_t)$ may be realizable given causal diagram $\mathcal{G}_2$ in Fig. 3, but not realizable given $\mathcal{G}_1$.

4 Applications: Counterfactual Decision-Making and Fairness

Next, we highlight the practical relevance of our results with some concrete use-cases. We already discussed in Example 1 how realizability can be used to design experiments for performing **mediation analysis** of direct/indirect effects, an important task in several fields. We now discuss applications in **causal fairness analysis** and **causal reinforcement learning (RL)**. Our goal is to underscore that the standard/baseline approaches in these areas, even among approaches that incorporate counterfactual reasoning, typically use observational ($\mathcal{L}_1$) or interventional ($\mathcal{L}_2$) data only, whereas a counterfactual ($\mathcal{L}_3$) data-collection approach can lead to demonstrably better results. Due to space constraints, we include in App. E the full specification of SCMs used and algorithms implemented.

4.1 Causal fairness - using counterfactual data for fairer decisions

Causal fairness analysis is a burgeoning field and a full survey is beyond the scope of this paper (see, e.g., Plecko & Bareinboim (2024) for a review of related works). We limit our discussion to an example where counterfactual realizability is directly relevant.

A common concern is that models trained to make automated decisions often reveal problematic biases (Angwin et al., 2016; Kodiyan, 2019, e.g.). The causal approach to address this is typically to constrain a classifier to obey some causally-sensitive *fairness measure*, μ (Plecko & Bareinboim, 2024, Def. 3.3). Some measures in the literature involve $\mathcal{L}_3$-quantities, and thus face the familiar issue of nonidentifiability (Kusner et al., 2017; Imai & Jiang, 2023). Other approaches acknowledge this limitation and try to construct interventional fairness measures that solely use $\mathcal{L}_2$-quantities (Salimi et al., 2019). We present next an example where relying only on $\mathcal{L}_2$-data can misleadingly approve a classifier as fair, but where a realizable $\mathcal{L}_3$ fairness measure actually ensures fairness. This scenario is inspired by a classic experiment in labor economics (Bertrand & Mullainathan, 2003).

Example 2 (Causal fairness). A college is developing an automated system to screen candidates in the first round of college applications, receiving as input a standardized CV per candidate. The system contains two models: model 1 outputs Y and model 2 outputs Z, which are binary decisions of whether the applicant cleared the first review stage for admission and for financial scholarship, respectively. The two models are respectively trained using data from previous years where an admissions team and a separate scholarship team reviewed applications manually. The college wants to ensure fairness w.r.t X, a candidate's race (a binary variable, for simplicity). In particular, they want to ensure equitable financial access to education for all qualified candidates: a candidate of race $X = 1$ who cleared the admissions screening $(Y = 1)$ but was rejected for financial aid $(Z = 0)$ should still receive $Z = 0$ had they been of race $X = 0$. The causal diagram is in Fig. 5(a), where the models' decisions Y, Z might reflect the unconscious race bias of the two committees in previous years (including possibly shared biases, represented by the latent confounder).

The $\mathcal{L}_3$ fairness measure they ought to minimize is thus

$$\mu_{ctf} = |P(Y_{x_1} = 1, Z_{x_1} = 0) - P(Y_{x_1} = 1, Z_{x_0} = 0)| \quad (6)$$

But the second term $P(y_x, z'_{x'})$ is nonidentifiable from the causal diagram in 5(a). So the college instead uses the following $\mathcal{L}_2$ measures, as an approximation for the fairness condition:

$$\begin{aligned} \mu_{int1} =& |P(Y = 1; do(x_1)).P(Z = 0; do(x_1)) \quad (7) \\ &- P(Y = 1; do(x_1)).P(Z = 0; do(x_0))| \\ \mu_{int2} =& |P(Y = 1, Z = 0; do(x_1)) \quad (8) \\ &- P(Y = 1, Z = 0; do(x_0))| \end{aligned}$$

They train the models, adding $\mu_{int1} + \mu_{int2}$ as a penalty in the objective. μ_{int1}, μ_{int2} are estimated using a holdout set of fake CVs, with the intervention $do(x)$ being enacted by randomly choosing an applicant name from an equivalence class which stereotypically indicates one unique race group $X = x$, e.g. names like Lakisha and Jamal for Blacks, or last names like Nguyen or Xi for Asians (cf. Bertrand & Mullainathan (2003)). Since the holdout set's CV body is independent of X, any effect of X on Y and Z is solely via the perception of race from the candidate name. We show in 5(c) simulations of such an optimization. In blue is the distribution of the true score μ_{ctf}, when the models are trained using μ_{int1}, μ_{int2}. Out of 1000 simulations of classifiers, we see $\mu_{ctf} > 5\%$ for nearly half the $\mathcal{L}_2$ simulations, indicating statistically significant discrimination roughly 50% the time.

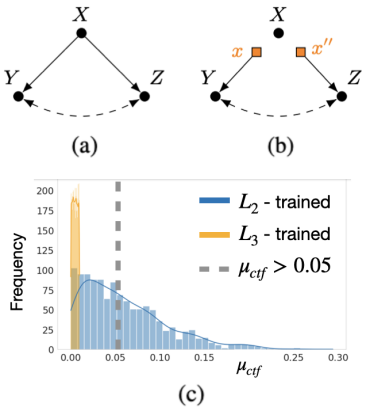


Figure 5: (a) Causal diagram for Example 2; (b) $P(Y_x, Z_{x'})$ is realizable using the interventions CTF-RAND$(X \to Y)$ and CTF-RAND$(X \to Z)$; (c) Histogram of 1000 classifiers trained on $\mathcal{L}_2$ (blue) and $\mathcal{L}_3$ (orange) fairness measures. $\mathcal{L}_2$ classifiers show statisically significant discrimination ($\mu_{ctf} > 0.05$).

However, the distribution $P(Y_x, Z_{x'})$ is indeed realizable (Def. 3.4) via the interventions CTF-RAND$(X \to Y)$, CTF-RAND$(X \to Z)$. The data science team notices that they can *separately and simultaneously* randomize the candidate name as an input to the respective models, and enact these interventions, as shown in 5(b). Thus, they are able to directly use the counterfactual measure μ_{ctf} as a fairness constraint in training. Results from 1000 simulations show that the classifiers trained directly using μ_{ctf} (shown in orange) nearly always meet the fairness requirement.

Details of the implementation are in App. E.2. Note: as in the original experiment, this example requires the structural assumption of race being revealed at the screening stage only by candidate name, which may be more defensible in highly standardized and controlled application processes. ■

4.2 CAUSAL RL - COUNTERFACTUAL POLICIES FOR OPTIMAL DECISION-MAKING

Consider a multi-arm bandit problem in which X represents the choice of bandit arm and Y the outcome. The default online learning approach is for the agent to adopt an algorithm like EXP3, UCB or Thompson Sampling to converge to some arm $x^\star := \arg\max_x \mathbb{E}[Y; do(x)]$ (Lattimore & Szepesvári, 2020; Sutton & Barto, 1998). Even in methods that explicitly incorporate causal knowledge, the common approach is to use a combination of offline ($\mathcal{L}_1$) and online ($\mathcal{L}_2$) data to converge more efficiently to the $\mathcal{L}_2$ optimization target $\arg\max_x \mathbb{E}[Y; do(x)]$ (Zhang & Bareinboim, 2017, e.g.). It was already shown in (Bareinboim et al., 2015; Forney et al., 2017) that it is possible to perform better by deploying a counterfactual strategy based on sampling each unit's natural choice $X = x'$ and randomizing actual choice in the *same* round, thus seeking to converge to $\arg\max_x \mathbb{E}[Y_x \mid x'], \forall x'$, as we discussed in Sec. 2. We call this the ETT baseline strategy, as it relies on drawing samples from the $\mathcal{L}_3$ ETT distribution, $P(Y_x, X)$, mentioned in Sec. 1.

Table 1: Performance of different strategies in Example 3.

Strategy	$\mathbb{E}[Y]$
Behavioral policy ($\mathcal{L}_1$)	0.65
Naive randomization ($\mathcal{L}_2$)	0.7
ETT baseline strategy ($\mathcal{L}_3$)	0.75
Optimal $\mathcal{L}_3$ strategy (this work)	0.80

We improve on this baseline by showing how an agent can leverage the realizability (Def. 3.4) of more nuanced counterfactuals like $P(Y_x, X, D_{x''})$ to construct superior counterfactual strategies. The following scenario involves an agent faced with adversarial latent confounding.

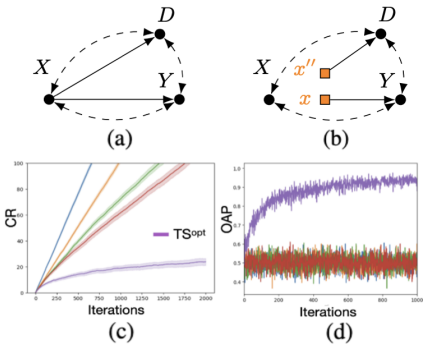


Figure 6: (a) Causal diagram for Example 3; (b) $P(Y_x, X, D_{x''})$ is realizable using the interventions CTF-RAND$(X \to Y)$ and CTF-RAND$(X \to D)$; (c) Cumulative Regret (CR) for $\mathcal{L}_1$ strategy (blue) and Thompson Sampling algorithms implementing naive $\mathcal{L}_2$ (yellow, green), ETT baseline (red), and optimal $\mathcal{L}_3$ strategy (purple); (d) Optimal Arm Probability (OAP) for all algorithms.

Example 3 (Counterfactual bandit policies). Consider a user of a social media platform which uses surveillance and predictions to increase user engagement through addictive notifications and recommendations (Zuboff, 2018). The user chooses every evening whether to use the platform via desktop ($X = 0$) or mobile ($X = 1$). Y is a binary indicator of whether she stays within her self-determined social media usage limit per day. She also notices that she receives ads when she logs in each evening as D (0: streaming service, 1: food delivery ads). The usage type X affects D, Y, as shown in Fig. 6(a).

On average, the user experiences $\mathbb{E}[Y] = 0.65$ from the observational ($\mathcal{L}_1$) policy of following her natural inclination each day. She suspects that the company could be tracking and exploiting her latent preferences, so she decides to randomize her daily choice and pick the best "arm". Sure enough, this naive $\mathcal{L}_2$ strategy breaks the adversarial confounding, and incurs a better avg. performance of $\mathbb{E}[Y; do(x)] = 0.7, \forall x$. She then decides to test the ETT-based strategy ($\mathcal{L}_3$) described earlier, by recording what she naturally feels like doing each day ($X = x'$), and subsequently randomizing her actual choice on the same day to optimize $\mathbb{E}[Y_x \mid x']$, getting an avg. performance of 0.75. However, at this point, she notices that she can do even better. The $\mathcal{L}_3$-distribution $P(Y_x, X, D_{x''})$ is realizable (Def. 3.4), since she can perform *another* counterfactual randomization, by sampling her natural choice ($X = x'$), randomly logging in to just see what ads she gets ($D_{x''} = d$), and again randomizing how she actually uses the platform that day to get Y_x. This strategy seeks an optimal $x^\star = \arg\max_x \mathbb{E}[Y_x \mid x', d_{x''}]$, which performs best as shown in Table 1. Details of the SCM, latent confounders, and the optimal $\mathcal{L}_3$-strategy are in App. E.3.

Simulations in the online setting corroborate this finding. Fig. 6(c,d) shows the cumulative regret (CR) and optimal arm probability (OAP) over 2000 iterations averaged over 200 epochs (CI=95%). We adapt Thompson Sampling to implement the strategies in Table 1. Details of implementation are in App. E.3.1. The optimal $\mathcal{L}_3$ strategy (purple) performs best, improving on the performance of the baseline ETT-based strategy (red). Naive randomizations, the standard $\mathcal{L}_2$ bandit strategy, are shown in yellow and green. All other algorithms fail to improve in OAP after 2000 iterations. ■

We make two remarks. First, the optimal counterfactual strategy is not simply a contextual Thompson Sampling, where $X, D_{x''}$ are used "merely" as extra context variables per round; indeed, treating this merely as a contextual bandit problem is one of the naive $\mathcal{L}_2$-strategies that we test (green plot in Fig. 6(c-d)), which ignores the counterfactual relationship between these variables and incurs dramatically higher regret, as we discuss in App. E.3.1.

Second, an interesting follow-up is whether we can guarantee that our strategy based on maximizing $\mathbb{E}[Y_x \mid x', d_{x''}]$ is optimal in this problem. Perhaps there are more refined $\mathcal{L}_3$-distributions like $P(Y_x, X, D_x, D_{x''})$ etc. that could yield better algorithms? It turns that this is indeed optimal, since most other $\mathcal{L}_3$-distributions are not realizable (Def. 3.4). As a bonus, we prove this claim for all bandit problems that fit a specific causal template in App. F. Thereby, we avoid having to conduct an intractable search over the space of all possible $\mathcal{L}_3$-strategies, trying to assess their realizability.

5 Discussion

Finally, we discuss some important implications, future directions, and limitations of our work.

Identification and bounding. Much work has been done in the area of $\mathcal{L}_3$ identification and estimation (Shpitser, 2008; Correa et al., 2021; Geneletti & Dawid, 2011). A natural extension to our work is to investigate the relationship between realizability and identification: which *additional* $\mathcal{L}_3$-quantities now become identifiable if the environment permits even some counterfactual randomization? This warrants an update to existing identification algorithms to allow (some) $\mathcal{L}_3$-data as input. Another fascinating research question involves "partial identification", where an input query is tightly bounded within a range that can be computed from available data (Zhang et al., 2022): how would the new $\mathcal{L}_3$-data further tighten the bounds for nonidentifiable $\mathcal{L}_3$-quantities?

Experiment design. One of the goals of this paper is to instigate new experiment design ideas that leverage ctf-randomization (Def. 2.3) and go beyond the standard RCT methodology, as in Examples 1-2. For instance, the increasingly automated HR pipeline in companies suggests opportunities for targeted interventions to randomize demographic details in virtual interviews, in standardized aptitude tests, or in performance-evaluation systems for remote workers, to track fairness metrics.

Causal reinforcement learning (CRL). While counterfactual strategies have been studied in CRL, the literature currently focuses on ETT-related strategies based on optimizing $\mathbb{E}[Y_x \mid x']$ (Bareinboim et al., 2015; Forney et al., 2017; Zhang & Bareinboim, 2022)(Richardson & Robins, 2013, §5.1). We presented an important extension by formalizing ctf-randomization (Def. 2.3) via counterfactual mediators (Def. 2.2), subsuming the previous approach. An ETT-based approach only allows one randomization of a variable X, affecting all downstream mechanisms. Our approach recognizes the possibility of isolating specific causal pathways and randomizing X multiple times per unit, demonstrably surpassing the ETT baseline in Example 3. We proved in App. F an optimality guarantee for our proposed strategy in bandit problems following a causal template. Generalizing this to sequential decision-making settings with arbitrary graphs is an important, non-trivial extension.

Limitations. The first obvious limitation of our framework is that it requires causal knowledge in the form of a graph (or equivalent). This is a standard assumption, needed to make progress in several areas of causal machine learning. Subsequent work could accommodate partial knowledge or model misspecification. Second, it may not always be feasible to perform counterfactual randomization (Def. 2.3) in a given setting. This is why Algo. 1 and Thm. 3.5 are general and do not assume this capability a priori. But where it is possible, even in principle, our work pinpoints opportunities for novel experiment design, as discussed above.

6 Conclusion

In this paper, we tackle the open question of which counterfactual distributions are directly accessible by experimental methods - what we define as the *realizability* of a distribution. Countering prevalent belief, we provide a complete algorithm and a graphical criterion for when a counterfactual can indeed be physically sampled from (Fig. 4). We demonstrate the practical relevance of this new framework with examples from causal fairness and causal RL, highlighting that ignoring this possibility could lead to poor outcomes. We believe that switching from an interventional to a counterfactual mindset could help researchers spot opportunities for *counterfactual randomization* that permit exciting new types of experiments, and improved, more personalized decisions.

REFERENCES

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i10.26435. URL https://doi.org/10.1609/aaai.v37i10.26435.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pp. 1342–1350, 2015.

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, New York, NY, USA, 1st edition, 2022.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Working Paper 9873, National Bureau of Economic Research, July 2003. URL http://www.nber.org/papers/w9873.

J. Correa and E. Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

Juan Correa, Sanghack Lee, and Elias Bareinboim. Nested counterfactual identification from arbitrary surrogate experiments. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6856–6867. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/36bedb6eb7152f39b16328448942822b-Paper.pdf.

A Philip Dawid. Causal Inference Without Counterfactuals (with Comments and Rejoinder). *Journal of the American Statistical Association*, 95(450):407–448, 2000.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022. doi: 10.1162/tacl_a_00511. URL https://aclanthology.org/2022.tacl-1.66.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1925.

Ronald A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual Data-Fusion for Online Reinforcement Learners. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. ISBN 9781510855144. doi: http://dx.doi.org/10.1037/a0022750.

Sara Geneletti and A. Philip Dawid. 72834 defining and identifying the effect of treatment on the treated. In *Causality in the Sciences*. Oxford University Press, 03 2011. ISBN 9780199574131. doi: 10.1093/acprof:oso/9780199574131.003.0034. URL https://doi.org/10.1093/acprof:oso/9780199574131.003.0034.

James J. Heckman and Richard Robb Jr. Alternative Methods for Evaluating the Impact of Interventions. In J J Heckman and B Singer (eds.), *Longitudinal Analysis of Labor Market Data*. Cambridge University Press, New York, NY, 1985.

James J. Heckman and Richard Robb Jr. Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In H. Wainer (ed.), *Drawing Inference From Self Selected Samples*, pp. 63–107. Springer-Verlag, New York, NY, 1986.

P W Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81 (396):945–960, 12 1986.

Duligur Ibeling and Thomas Icard. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10170–10177, 2020.

Kosuke Imai and Zhichao Jiang. Principal Fairness for Human and Algorithmic Decision-Making. *Statistical Science*, 38(2):317 – 328, 2023. doi: 10.1214/22-STS872. URL https://doi.org/10.1214/22-STS872.

Akhil Alfons Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. *ResearchGate preprint*, pp. 1–19, 11 2019.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Ang Li and Judea Pearl. Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 1793–1799. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/248. URL https://doi.org/10.24963/ijcai.2019/248.

Scott Mueller and Judea Pearl. Personalized decision making – a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023. doi: doi:10.1515/jci-2022-0050. URL https://doi.org/10.1515/jci-2022-0050.

Scott Mueller and Judea Pearl. Perspective on 'harm' in personalized medicine – an alternative perspective. forthcoming.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1–2):93–149, 11 1999.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000. ISBN 978-0-521-89560-6.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann, San Francisco, CA, 2001.

Judea Pearl. Direct and indirect effects. In *Proceedings of the American Statistical Association, Joint Statistical Meetings*, pp. 1572–1581. {MIRA} Digital Publishing, Minn., MN, 2005.

Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018. ISBN 978-0-465-09760-9.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A Primer*. John Wiley & Sons, 2016. ISBN 9781119186847.

Drago Plecko and Elias Bareinboim. Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3):304–589, Jan 2024. ISSN 1935-8237. doi: 10.1561/2200000106. URL https://doi.org/10.1561/2200000106.

Hans Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Working Paper 128, Center for the Statistics and the Social Sciences, 2013.

James M Robins and Sander Greenland. Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143–155, 1992.

Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 2004.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. SIGMOD '19, pp. 793–810, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450356435. doi: 10.1145/3299869.3319901. URL https://doi.org/10.1145/3299869.3319901.

Ilya Shpitser. *Complete Identification Methods for Causal Inference*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, 4 2008.

Ilya Shpitser and Judea Pearl. What Counterfactuals Can Be Tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359. AUAI Press, Vancouver, BC, Canada, 2007.

Ilya Shpitser and Judea Pearl. Effects of Treatment on the Treated: Identification and Generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Quebec, 2009. AUAI Press.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Kevin Xia and Elias Bareinboim. Neural causal abstractions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20585–20595, Mar. 2024. doi: 10.1609/aaai.v38i18.30044. URL https://ojs.aaai.org/index.php/AAAI/article/view/30044.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: A Causal approach. In *IJCAI International Joint Conference on Artificial Intelligence*, 2017. ISBN 9780999241103.

Junzhe Zhang and Elias Bareinboim. Can humans be out of the loop? In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 1010–1025. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/zhang22a.html.

Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26548–26558. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22ab.html.

Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st edition, 2018. ISBN 1610395697.

APPENDICES

Appendix A: Graphical terminology

Appendix B: Details on the **CTF-REALIZE** algorithm

Appendix C: Assumptions and realizability proofs

Appendix D: Details on counterfactual randomization

Appendix E: Details on examples

Appendix F: Optimality result and proof

A GRAPHICAL TERMINOLOGY

Structural Causal Models (SCM) and causal diagrams are described in the preliminaries in Sec. 1. See (Bareinboim et al., 2022) for full treatment. We use the following graphical kinship nomenclature w.r.t causal diagram $\mathcal{G}$:

- Parent(s) of V, denoted $\mathbf{Pa}_V$: the set of variables $\{V'\}$ s.t. there is a direct edge $V' \to V$ in $\mathcal{G}$. $\mathbf{Pa}_V$ does not include V.
- Children of V, denoted $Ch(V)$: the set of variables $\{V'\}$ s.t. there is a direct edge $V \to V'$ in $\mathcal{G}$. $Ch(V)$ does not include V.
- Ancestors of V, denoted $An(V)$: the set of variables $\{V'\}$ s.t. there is a path (possibly length 0) from V' to V consisting only of edges pointing toward V, $V' \to ... \to V$. $An(V)$ is defined to include V.
- Descendants of V, denoted $Desc(V)$: the set of variables $\{V'\}$ s.t. there is a path (possibly length 0) from V to V' consisting only of edges pointing toward V', $V \to ... \to V'$. $Desc(V)$ is defined to include V.
- Non-descendants of V, denoted $NDesc(V)$: the set $\mathbf{V} \setminus Desc(V)$. $NDesc(V)$ does not include V.

Given a graph $\mathcal{G}$, $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{W}}}$ is the result of removing edges coming into variables in $\mathbf{X}$, and edges coming out of $\mathbf{W}$.

B Details on the **CTF-REALIZE** algorithm

B.1 Sub-routine of **CTF-REALIZE** algorithm (Algo. 1)

Algorithm 2 COMPATIBLE (sub-routine)

1: **Input:** $V \in \mathbf{V}$ of $\mathcal{G}$; $W_{\mathbf{t}} \in \mathbf{W}_\star$ of Q
2: **for** each $C \in Ch(V)$ **do**
3: **if** $C \in An(W)$ **then**
4: **if** $V \in \mathbf{T}$ **then**
5: Let $v :=$ value of V in subscript $\mathbf{t}$
6: Find smallest $\mathbf{C} \ni C$ s.t. CTF-RAND$(V \to \mathbf{C}) \in \mathbb{A}$
7: **if** $\{$CTF-RAND$(V \to \mathbf{C}) : .\} \in$ INT$_V$ and its label is not "v" **then**
8: Return FAIL
9: **else**
10: Add $\{$CTF-RAND$(V \to \mathbf{C}) : v\}$ to INT$_V$, with the label "v"
11: **end if**
12: **if** no such $\mathbf{C} \ni C$ s.t. CTF-RAND$(V \to \mathbf{C}) \in \mathbb{A}$ **then**
13: **if** $\{$RAND$(V) : .\} \in$ INT$_V$ and its label is not "v" **then**
14: Return FAIL
15: **else if** RAND$(V) \notin \mathbb{A}$ **then**
16: Return FAIL
17: **else**
18: Add $\{$RAND$(V) : v\}$ to INT$_V$, with the label "v"
19: **end if**
20: **end if**
21: **end if**
22: **if** $V \notin \mathbf{T}$ **then**
23: **for** each $\mathbf{C} \ni C$ s.t. CTF-RAND$(V \to \mathbf{C}) \in \mathbb{A}$ **do**
24: **if** $\{$CTF-RAND$(V \to \mathbf{C}) : .\} \in$ INT$_V$ and its label is not "Natural" **then**
25: Return FAIL
26: **else**
27: Add $\{$CTF-RAND$(V \to \mathbf{C}) :$ Natural$\}$ to INT$_V$, with the label "Natural"
28: **end if**
29: **end for**
30: **if** $\{$RAND$(V) : .\} \in$ INT$_V$ and its label is not "Natural" **then**
31: Return FAIL
32: **else if** RAND$(V) \in \mathbb{A}$ **then**
33: Add $\{$RAND$(V) :$ Natural$\}$ to INT$_V$, with the label "Natural"
34: **end if**
35: **end if**
36: **end if**
37: **end for**

B.2 Walk-through of the **CTF-REALIZE** algorithm (Algo. 1)

General strategy. For each variable V in the input graph, we check what are the necessary and sufficient interventions (or lack of interventions) we need to perform w.r.t each term $W_{\mathbf{t}}$ in the input query $\mathbf{W}_\star$. This is what the inner loops and subroutine **COMPATIBLE** are doing - accumulating correct and complete conditions in topological order. If there is no conflict across these conditions collectively, and if the feasible action set contains the necessary actions, the query is realizable. Otherwise not. We use induction to prove that it is enough to check for conflicting conditions with regard to prior loops, in topological order. Proof of completeness results in Appendix C.

Walk-through for Algo. 1 **CTF-REALIZE**:

i. **Lines 5-7**: we go over each node V in the input graph, in topological order, and maintain a tracker of interventions (and lack of interventions) needed w.r.t V to realize $\mathbf{W}_\star$; we also check whether V needs to be added to the final output vector.

ii. **Line 10**: for each $W_{\mathbf{t}}$ in the input query $\mathbf{W}_\star$, we check if there is any conflict in necessary and sufficient conditions w.r.t V for realizing $W_{\mathbf{t}}$, by calling **COMPATIBLE**$(V, W_{\mathbf{t}})$. This only needs to be done if $V \in An(W)$; otherwise V has no effect on $W_{\mathbf{t}}$.

iii. **Line 13**: if $V = W$, the measured value of V needs to be added to the output vector $\mathbf{w}$.

iv. **Lines 17-18**: we perform all the interventions (if any) that are needed for V. Step [ii] has already checked whether these actions can be performed, and if there are any conflicts.

- **Note on rejection sampling**: since we framed our actions as randomizations, in order to enact an intervention like $do(x)$, we draw a random value and reject if the draw is not x. This is for clarity of presentation, aligned with the rest of the paper. We could easily

introduce deterministic actions, or add some concentration guarantees of drawing x within finite samples etc. but this is well-understood and would be a distraction.

v. **Lines 21-25**: if V itself is part of the output, the set of necessary actions cannot involve a Fisherian RAND of V (because unlike CTF-RAND, Fisherian RAND overrides the mechanism f_V generating V).

vi. **Line 29**: if the above steps have been completed w.r.t V for each $W_{\mathbf{t}}$, there will be no further conflicts arising w.r.t V and all nodes topologically prior to V, regardless of conditions needed w.r.t subsequent nodes (by an induction argument). If this is can be completed for all V, then the query is realizable and we output the vector $\mathbf{w}$.

Walk-through for sub-routine Algo. 2 **COMPATIBLE**$(V, W_{\mathbf{t}})$ called in Step [ii]:

ii.a. **Lines 2-3**: the necessary and sufficient conditions w.r.t V for $W_{\mathbf{t}}$ involve how the children $Ch(V)$ receive the value of V as an input. We only care about the children that belong in $An(W)$ for this sub-routine call; if a child is not in $An(W)$ it wouldn't affect $W_{\mathbf{t}}$.

ii.b. **Lines 4-19**: if $V \in \mathbf{T}$, this means the potential response $W_{\mathbf{t}}$ involves an intervention on V. We find the minimal interventions needed to achieve this (CTF-RAND for the smallest subset of children possible, and failing this, a Fisherian RAND); and we update our tracker of necessary actions for V.

ii.b.1. The necessary and sufficient conditions to measure $W_{\mathbf{t}}$ is that $\mathbf{T}$ should be fixed as $\mathbf{t}$ (by intervention) as an input to all children $C \in Ch(\mathbf{T}) \cap An(W)$; and that all other ancestors of W are received naturally by their respective children. These two conditions ensure that W is evaluated in the $do(\mathbf{t})$ regime, as defined in the SCM.

ii.b.2. In particular, for $V \in \mathbf{T}$ in this sub-routine call, this means that we need to fix $v \in \mathbf{t}$ (**Line 5**) as input to each relevant child $C \in Ch(V) \cap An(W)$. So for each such child C, we find the smallest subset $\mathbf{C}' \subseteq Ch(V)$ s.t. $C \in \mathbf{C}'$ and CTF-RAND$(V \to \mathbf{C}')$ is in the input action set $\mathbb{A}$ (**Line 6**), and we add this to the required actions, along with the tagged value "v" that needs to be fixed for V by this action (**Line 10**). If CTF-RAND is not available, we add Fisherian RAND(V) to the required action list (**Line 18**). We call this chosen action the "minimal action".

ii.b.3. If no such randomization action is available, return FAIL (**Line 16**). If the minimal action is already being used in a previous loop to enforce a different value $v' \neq v$ that is not compatible with $\mathbf{t}$, return FAIL (**Lines 7-8, 13-14**). Or if the minimal action has already been tagged with the value "Natural" in a previous loop, return FAIL - this means a previous loop already recorded that this action must *necessarily not* be performed (**Lines 7-8, 13-14**). Such a conflict means that $W_{\mathbf{t}}$ cannot be realised.

ii.c. **Lines 22-34**: if $V \notin \mathbf{T}$, this means the potential response $W_{\mathbf{t}}$ requires that V be received without intervention (i.e. "naturally") by the relevant child nodes; we also add this necessary condition to our tracker.

ii.c.1. Again, the necessary and sufficient conditions to measure $W_{\mathbf{t}}$ is that each ancestor $A \in An(W)_{\mathcal{G}_{\overline{\mathbf{T}}}}$, with $A \notin \mathbf{T} \cup \{W\}$, needs to be received "naturally" (i.e., without intervention) by its children $C \in Ch(A) \cap An(W)$; and that $\mathbf{T}$ needs to be fixed by intervention, as discussed earlier in [ii.b.1]. These two conditions ensure that W is evaluated in the $do(\mathbf{t})$ regime, as defined in the SCM.

ii.c.2. In particular, for $V \notin \mathbf{T}$ in this sub-routine, this means that we should *necessarily not* intervene on the value of V that is received as input by each relevant child $C \in Ch(V)$ which is also an ancestor of W.

ii.c.3. We track this requirement by adding that for every possible randomization CTF-RAND$(V \to \mathbf{C}')$ and RAND(V), where $\mathbf{C}'$ contains an ancestor of W, that this action *cannot* be performed, by using the tag "Natural" (**Lines 27, 33**).

ii.c.4. If any such action identified in [ii.c.3] has already been tagged with some value v', this means that a previous loop has recorded that this action *necessarily* needs to be performed in order to fix the value as v' - we return FAIL (**Lines 24-25, 30-31**). Such a conflict means that $W_{\mathbf{t}}$ cannot be realized.

B.3 Complexity analysis of **CTF-REALIZE** algorithm (Algo. 1)

The *time complexity* of Algorithm 1 is $\mathcal{O}(kn^2)$, where k is the number of terms in the input query Q, and n is the number of variables in the input graph $\mathcal{G}$.

k depends on the domain size of the variables. That is, $k \leq n.\prod_V \Big(|Domain(V)| + 1\Big) = \mathcal{O}(m^n)$, where m is the domain size of the variable with the most possible categorical values.

The *space complexity* is the same, as the algorithm needs to store up to all intermediate steps before terminating.

B.4 Examples using the **CTF-REALIZE** algorithm

Example B.1. (ETT realizability)

Query, $Q = P(Y_x, X)$

Graph, $\mathcal{G}$: Fig. 7

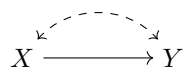


Figure 7: Graph for Example B.1

Suppose action set $\mathbb{A} = \mathbb{A}^\dagger(\mathcal{G}) := \{\text{CTF-RAND}(X \to Y)\}$

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y_x)
    - $Y \in Ch(X)$ and $Y \in An(Y)$
    - $X \in$ subscript of Y_x
    - CTF-RAND$(X \to Y) \in \mathbb{A}^\dagger(\mathcal{G})$
    - INT$_X \leftarrow \{\text{CTF-RAND}(X \to Y) : x\}$
- For the second term in $\mathbf{W}_\star$: X
  - OUTPUT$_X \leftarrow \{X\}$
- Moving to Y (next in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - OUTPUT$_Y \leftarrow \{Y_x\}$
- Perform interventions in INT$_X$, followed by READ, and assign output vector based on OUTPUT$_X$, OUTPUT$_Y$
- Return i.i.d sample

For simplicity, we don't show the steps SELECT$^{(i)}$ and the rejection sampling involving in the randomization procedure (steps 17-18 of Algo. 1).

Thus, Q is realizable given $\mathcal{G}, \mathbb{A}^\dagger$. This is validated by the ancestor set $An(Y_x, X)_\mathcal{G} = \{Y_x, X\}$, which doesn't repeat any variables. This is also illustrated in Fig. 8.

However, suppose the agent's action set is

$\mathbb{A} = \{\text{RAND}(X)\}$, i.e., does not permit any counterfactual randomization procedures.

In this case,

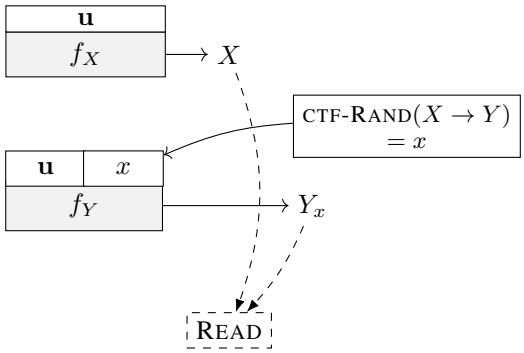


Figure 8: $P(Y_x, X)$ is realizable given the graph in Fig. 7 and $\mathbb{A}^\dagger(\mathcal{G})$.

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A})$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y_x)
    - $Y \in Ch(X)$ and $Y \in An(Y)$
    - $X \in$ subscript of Y_x
    - RAND$(X) \in \mathbb{A}$, and no other ctf-randomization procedure
    - INT$_X \leftarrow \{$RAND$(X) : x\}$
- For the second term in $\mathbf{W}_\star$: X
  - OUTPUT$_X \leftarrow \{X\}$
- Moving to Y (next in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - OUTPUT$_Y \leftarrow \{Y_x\}$
- OUTPUT$_X$ contains X, but the intervention set INT$_X$ contains RAND(X)
- FAIL (Line 22 of Algo. 1)

■

Example B.2. (Probability of sufficiency (PS) realizability)

Query, $Q = P(Y_x, X, Y)$

Graph, $\mathcal{G}$: Fig. 9

$$X \longrightarrow Y$$

Figure 9: Graph for Example B.2

Suppose action set $\mathbb{A} = \mathbb{A}^\dagger(\mathcal{G}) := \{$CTF-RAND$(X \to Y)\}$

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y_x)

    * $Y \in Ch(X)$ and $Y \in An(Y)$
    * $X \in$ subscript of Y_x
    * $\text{CTF-RAND}(X \to Y) \in \mathbb{A}^\dagger(\mathcal{G})$
    * $\text{INT}_X \leftarrow \{\text{CTF-RAND}(X \to Y) : x\}$

- For the second term in $\mathbf{W}_\star$: X
  - $\text{OUTPUT}_X \leftarrow \{X\}$
- For the third term in $\mathbf{W}_\star$: Y
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y)
    * $Y \in Ch(X)$ and $Y \in An(Y)$
    * $X \notin$ subscript of Y ; X needs to be received naturally
    * But INT_X already contains $\{\text{CTF-RAND}(X \to Y) : x\}$ with label $x \neq$ "Natural"
    * FAIL (Line 25 of Algo. 2)

Thus, Q is not realizable given $\mathcal{G}, \mathbb{A}^\dagger$. This is validated by the ancestor set $An(Y_x, X, Y)_\mathcal{G} = \{Y_x, X, Y\}$, which contains both Y_x, Y. This is also illustrated in Fig. 10.

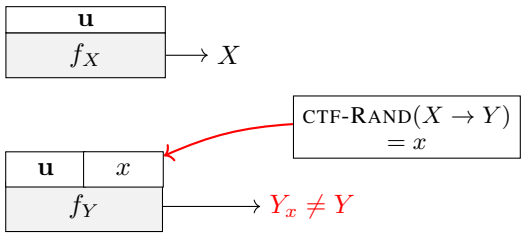


Figure 10: $P(Y_x, X, Y)$ is not realizable given the graph in Fig. 9 and $\mathbb{A}^\dagger(\mathcal{G})$.

■

Example B.3. Query, $Q = P(W_{xt}, Z_{x'})$

Graph, $\mathcal{G}$: Fig. 11

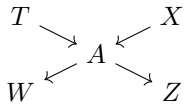


Figure 11: Graph for Example B.3

Suppose action set $\mathbb{A} = \mathbb{A}^\dagger(\mathcal{G}) := \{\text{CTF-RAND}(T \to A), \text{CTF-RAND}(X \to A), \text{CTF-RAND}(A \to W), \text{CTF-RAND}(A \to Z)\}$

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: W_{xt}
  - Since $X \in An(W)$, call Algo. 2 **COMPATIBLE**(X, W_{xt})
    * $A \in Ch(X)$ and $A \in An(W)$
    * $X \in$ subscript of W_{xt}
    * $\text{CTF-RAND}(X \to A) \in \mathbb{A}^\dagger(\mathcal{G})$
    * $\text{INT}_X \leftarrow \{\text{CTF-RAND}(X \to A) : x\}$
- For the second term in $\mathbf{W}_\star$: $Z_{x'}$

- Since $X \in An(Z)$, call Algo. 2 **COMPATIBLE**$(X, Z_{x'})$
  - $A \in Ch(X)$ and $A \in An(Z)$
  - $X \in$ subscript of $Z_{x'}$; X needs to be fixed as x'
  - But INT_X already contains $\{\text{CTF-RAND}(X \to A) : x\}$ with label $x \neq x'$
  - FAIL (Line 8 of Algo. 2)

Thus, Q is not realizable given $\mathcal{G}, \mathbb{A}^\dagger$. This is validated by the ancestor set $An(W_{xt}, Z_{x'})_{\mathcal{G}} = \{W_{xt}, A_{xt}, Z_{x'}, A_{x'}\}$, which contains both $A_{xt}, A_{x'}$. This is also illustrated in Fig. 12.

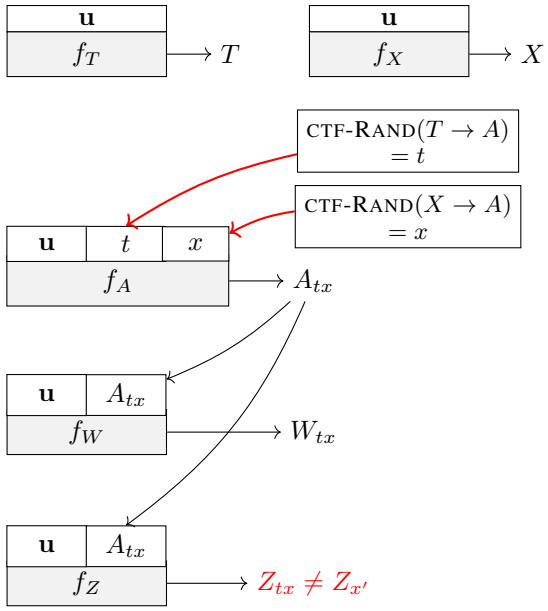


Figure 12: $P(W_{xt}, Z_{x'})$ is not realizable given the graph in Fig. 11 and $\mathbb{A}^\dagger(\mathcal{G})$.

■

Example B.4. Query, $Q = P(Y_x, Z_{x'}, W_{x''})$

Graph, $\mathcal{G}$: Fig. 13

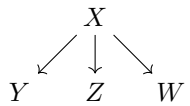


Figure 13: Graph for Example B.4

Suppose action set $\mathbb{A} = \{\text{RAND}(X), \text{CTF-RAND}(X \to \{Z, W\})\}$

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A})$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y_x)
    - $Y \in Ch(X)$ and $Y \in An(Y)$
    - $X \in$ subscript of Y_x
    - $\text{RAND}(X) \in \mathbb{A}$; no other ctf-randomization procedures affecting Y
    - $\text{INT}_X \leftarrow \{\text{RAND}(X) : x\}$

- For the second term in $\mathbf{W}_\star$: $Z_{x'}$
  - Since $X \in An(Z)$, call Algo. 2 **COMPATIBLE**$(X, Z_{x'})$
    - $Z \in Ch(X)$ and $Z \in An(Z)$
    - $X \in$ subscript of $Z_{x'}$
    - CTF-RAND$(X \to \{Z, W\}) \in \mathbb{A}$; no smaller ctf-randomization procedures affecting Z
    - INT$_X \leftarrow$ INT$_X \cup \{$CTF-RAND$(X \to \{Z, W\}) : x'\}$
- For the third term in $\mathbf{W}_\star$: $W_{x''}$
  - Since $X \in An(W)$, call Algo. 2 **COMPATIBLE**$(X, W_{x''})$
    - $W \in Ch(X)$ and $W \in An(W)$
    - $X \in$ subscript of $W_{x''}$; X needs to be fixed as x''
    - But INT$_X$ already contains $\{$CTF-RAND$(X \to \{Z, W\}) : x'\}$ with label $x' \neq x''$
    - No smaller ctf-randomization procedures affecting W
    - FAIL (Line 8 of Algo. 2)

Thus, Q is not realizable given $\mathcal{G}, \mathbb{A}$.

However, suppose instead that action set $\mathbb{A}' = \{$RAND(X), CTF-RAND$(X \to \{Z, W\})$, CTF-RAND$(X \to Z)\}$

CTF-REALIZE$(Q, \mathcal{G}, \mathbb{A}')$ trace:

- Start with X (first in topological order)
- For the first term in $\mathbf{W}_\star$: Y_x
  - Since $X \in An(Y)$, call Algo. 2 **COMPATIBLE**(X, Y_x)
    - $Y \in Ch(X)$ and $Y \in An(Y)$
    - $X \in$ subscript of Y_x
    - RAND$(X) \in \mathbb{A}$; no other ctf-randomization procedures affecting Y
    - INT$_X \leftarrow \{$RAND$(X) : x\}$
- For the second term in $\mathbf{W}_\star$: $Z_{x'}$
  - Since $X \in An(Z)$, call Algo. 2 **COMPATIBLE**$(X, Z_{x'})$
    - $Z \in Ch(X)$ and $Z \in An(Z)$
    - $X \in$ subscript of $Z_{x'}$
    - CTF-RAND$(X \to Z) \in \mathbb{A}$
    - INT$_X \leftarrow$ INT$_X \cup \{$CTF-RAND$(X \to Z) : x'\}$
- For the third term in $\mathbf{W}_\star$: $W_{x''}$
  - Since $X \in An(W)$, call Algo. 2 **COMPATIBLE**$(X, W_{x''})$
    - $W \in Ch(X)$ and $W \in An(W)$
    - $X \in$ subscript of $W_{x''}$
    - CTF-RAND$(X \to \{Z, W\}) \in \mathbb{A}$
    - INT$_X \leftarrow$ INT$_X \cup \{$CTF-RAND$(X \to \{Z, W\}) : x''\}$
- Moving to Y (next in topological order)
  - OUTPUT$_Y \leftarrow \{Y_x\}$
- Moving to Z (next in topological order)
  - OUTPUT$_Z \leftarrow \{Z_{x'}\}$
- Moving to W (next in topological order)
  - OUTPUT$_W \leftarrow \{W_{x''}\}$

- Perform interventions in INT_X, followed by READ, and assign output vector based on $\text{OUTPUT}_Y, \text{OUTPUT}_Z$,$\text{OUTPUT}_W$
- Return i.i.d sample

Thus, Q is realizable given $\mathcal{G}, \mathbb{A}'$.

Lastly, it is evident that Q is realizable given $\mathcal{G}, \mathbb{A}^\dagger$. This is validated by the ancestor set $An(Y_x, Z_{x'}, W_{x''})_\mathcal{G} = \{Y_x, Z_{x'}, W_{x''}\}$, which does not repeat any variables. This is also illustrated in Fig. 14.

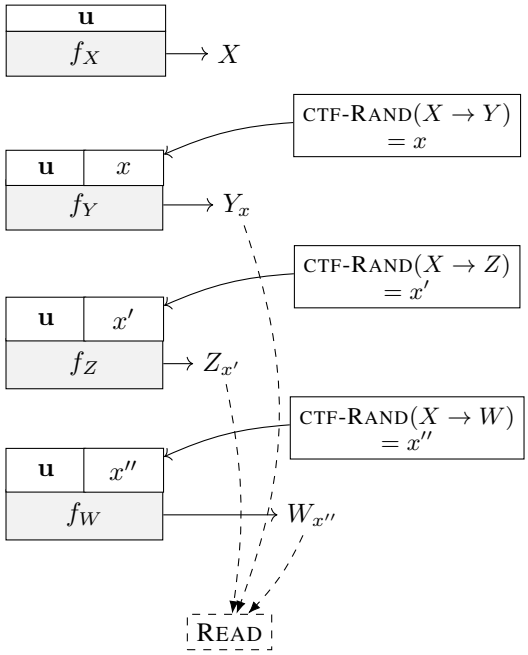


Figure 14: $P(Y_x, Z_{x'}, W_{x''})$ is realizable given the graph in Fig. 13 and $\mathbb{A}^\dagger(\mathcal{G})$.

■

C Assumptions and Realizability Proofs

C.1 Assumptions

In this section, we gather together all the structural assumptions we make in this paper, for ease of reference. We also include related remarks.

Assumption C.1 (Unobservability)**.** An agent deployed in the environment does not know the underlying SCM $\mathcal{M}$ of the environment, and does not know the latent features $\mathbf{U}^{(i)}$ of any unit i in the target population. ■

Assumption C.2 (Feasible actions)**.** Given causal diagram $\mathcal{G}$, the physical actions that an agent can perform on any unit i in the target population are limited to: $\textsc{Select}^{(i)}$, $\textsc{Read}(V)^{(i)}$, $\textsc{Rand}(X)^{(i)}$, and $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$, for some $V, X \in \mathbf{V}$ and $\mathbf{C} \subseteq Ch(X)_{\mathcal{G}}$, per Defs. 2.1,2.3. ■

Assumption 3.1 (Fundamental constraint of experimentation (FCE))**.** A unit i in the target population can physically undergo a causal mechanism $f_V \in \mathcal{F}$ at most once. ■

We define the probability measure $P^{\mathbb{C}}(.)$ from the perspective of an exogenous agent (i.e., an agent external to the system) $\mathbb{C}$'s beliefs about the environment, distinguished by superscript from $P^{\mathcal{M}}(.)$, the true unknown distribution.

Remark C.3. Let $\mathcal{A}^{(i)}$ be a sequence of actions taken by agent $\mathbb{C}$ on unit i that is not conditional on any data gathered regarding i. The assumption of $\mathbb{C}$ behaving exogenously means that $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{u})$. ■

Regarding the structural conditions involving counterfactual randomization (Def. 2.3), we make the following assumption, mainly as a simplifying step for use in the proofs.

Assumption D.3 (Tree structure)**.** Given a variable X, causal diagram $\mathcal{G}$, and an "expanded" diagram $\mathcal{G}^+$ (Def. D.1) including the set of all the counterfactual mediators $\mathbf{W}$ (Def. D.2) of X in the environment, each $W \in \mathbf{W}$ has only one parent in $\mathcal{G}^+$, and each $C \in Ch(X)_{\mathcal{G}}$ has at most one $W \in \mathbf{W}$ as a parent in $\mathcal{G}^+$. ■

From this assumption, and from the definition of a counterfactual mediator (Def. D.2), we can derive the following observations:

Remark C.4 (No bypassing children). Given causal diagram $\mathcal{G}$, the procedure $\textsc{ctf-Rand}(X \to \mathbf{C})$, either by eliciting a unit's natural decision or via a counterfactual mediator, can only be performed w.r.t $\mathbf{C} \subseteq Ch(X)_{\mathcal{G}}$. It cannot by-pass child mechanisms and directly affect a descendant. This is elaborated in App. D.3, and specifically in Lemma D.7. ■

Remark D.6 (Procedure containment). Assumption D.3 implies that if an agent is capable of performing both $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$ and $\textsc{ctf-Rand}(X \to \mathbf{C}')^{(i)}$ s.t. $\mathbf{C} \neq \mathbf{C}'$ and $\mathbf{C} \cap \mathbf{C}' \neq \emptyset$, then either $\mathbf{C} \subseteq \mathbf{C}'$ or $\mathbf{C}' \subseteq \mathbf{C}$. ■

Remark D.5 (Superseding action). Given a decision variable X, the action $\textsc{ctf-Rand}(X \to \mathbf{C}')^{(i)}$ can *supersede* the action $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$ if $\mathbf{C}' \subsetneq \mathbf{C}$, where *supersede* means that the former action $\textsc{ctf-Rand}(X \to \mathbf{C}')^{(i)}$ blocks any effect that the latter action has on the variables $\mathbf{C}'$. Additionally, the action $\textsc{ctf-Rand}(X \to \mathbf{C})^{(i)}$ *supersedes* the action $\textsc{Rand}(X)^{(i)}$. ■

Counterfactual randomization permits multiple randomizations for the same variable X for a single unit i. But some randomizations block the effects of others. See App. D.2.

C.2 PROOFS FOR SECTION 3

Recall, $P^{\mathbb{C}}(.)$ is the probability measure from the perspective of an exogenous agent (i.e., an agent external to the system) $\mathbb{C}$'s beliefs about the environment, distinguished by superscript from $P^{\mathcal{M}}(.)$, the true unknown distribution.

Since unit selection is randomized, $\text{SELECT}^{(i)}$ yields an unbiased sample of a unit with latent features distributed according to the target population frequency $P(\mathbf{u})$. I.e., $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \text{SELECT}^{(i)}) = P^{\mathcal{M}}(\mathbf{u})$.

Lemma C.5 (I.i.d requirement)**.** *Consider a sequence of actions $\mathcal{A}^{(i)}$ performed on unit i in the target population, that yields a vector of realized values $\mathbf{W}_\star^{(i)}$. $\mathbf{W}_\star^{(i)}$ is an i.i.d sample from $P^{\mathcal{M}}(\mathbf{W}_\star)$, for arbitrary $\mathcal{M}$ iff*

i. $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{U} = \mathbf{u})$*; and*

ii. $\mathbb{1}[\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}, \mathbf{U}^{(i)} = \mathbf{u}] = \mathbb{1}[\mathbf{W}_\star(\mathbf{u}) = \mathbf{w}]$.

Proof. Recall from Def. 3.3 that $\mathbf{W}_\star^{(i)}$ being an i.i.d sample from $P^{\mathcal{M}}(\mathbf{W}_\star)$ means that

$$P^{\mathbb{C}}(\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{W}_\star = \mathbf{w}), \forall \mathbf{w} \tag{9}$$

<u>Reverse direction:</u>

We simply multiply respective l.h.s and r.h.s of conditions [i] and [ii] and sum over all $\mathbf{u}$ to get

$$\sum_{\mathbf{u}} P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \mathcal{A}^{(i)}).\mathbb{1}[\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}, \mathbf{U}^{(i)} = \mathbf{u}] = \sum_{\mathbf{u}} P^{\mathcal{M}}(\mathbf{U} = \mathbf{u}).\mathbb{1}[\mathbf{W}_\star(\mathbf{u}) = \mathbf{w}] \tag{10}$$

$$= P^{\mathcal{M}}(\mathbf{W}_\star = \mathbf{w}), \tag{11}$$

which we get from the Layer 3 valuation formula (see preliminaries in Sec. 1). On the l.h.s, we apply the chain rule to get the result we need.

$$P^{\mathbb{C}}(\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{W}_\star = \mathbf{w}) \tag{12}$$

<u>Forward direction:</u>

Assume Eq. 9.

From Remark C.3, we conclude that condition [i] automatically holds. Since the agent acts exogenously to the system, $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \mathcal{A}^{(i)}) = P^{\mathcal{M}}(\mathbf{U} = \mathbf{u}), \forall \mathbf{u}$.

Applying the chain rule on both sides of Eq. 9,

$$\sum_{\mathbf{u}} P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \mathcal{A}^{(i)}).\mathbb{1}[\mathbf{W}_\star^{(i)} = \mathbf{w} \mid \mathcal{A}^{(i)}, \mathbf{U}^{(i)} = \mathbf{u}] = \sum_{\mathbf{u}} P^{\mathcal{M}}(\mathbf{U} = \mathbf{u}).\mathbb{1}[\mathbf{W}_\star(\mathbf{u}) = \mathbf{w}] \tag{13}$$

The probability terms are equal, for each $\mathbf{u}$.

Since the probability terms are free parameters, and we need this equation to hold for any arbitrary probability simplex, it must be the case that the indicator terms are also equal. Thus begetting condition [ii].

■

Lemma C.6. *Given a causal diagram $\mathcal{G}$, for any SCM $\mathcal{M}$ compatible with $\mathcal{G}$, the jointly necessary and sufficient conditions to measure a potential response $W_{\mathbf{t}}(\mathbf{u})$ are [i] $\mathbf{T}$ is fixed as $\mathbf{t}$ (by intervention) as an input to all children $C \in Ch(\mathbf{T}) \cap An(W)$; [ii] each $A \in An(W)_{\mathcal{G}_{\overline{\mathbf{T}}}}$, $A \notin \{\mathbf{T}, W\}$ is received "naturally" (i.e., without intervention) by its children $C \in Ch(A) \cap An(W)$; and [iii] the mechanism f_W is not erased and overwritten (by a Fisherian intervention).*

Proof. $W_{\mathbf{t}}$ is the variable W evaluated in the sub-model $\mathcal{M}_{\mathbf{t}}$, where the equations for $\mathbf{T}$ are replaced by constant values in $\mathbf{t}$.

For any changes to the function for $T \in \mathbf{T}$, the function f_W is only affected by any effect on the children of T which are also ancestors of W. Any effect of T on some $C' \in Ch(T)$ s.t. $C' \notin An(W)$ has no effect on W.

Further, in the submodel $\mathcal{M}_{\mathbf{t}}$ there are no interventions on any other ancestors of W in $\mathcal{G}_{\overline{\mathbf{T}}}$, besides $\mathbf{T}$. Even if there were tnterventions involving some $X \notin An(W)_{\mathcal{G}_{\overline{\mathbf{T}}}}$, this would have no effect on W in the sub-model $\mathcal{M}_{\mathbf{t}}$, by Rule 3 of do-calculus.

It is evident that f_W evaluated according to the sub-model $\mathcal{M}_{\mathbf{t}}$, and evaluated according to a sub-model satisfying conditions [i] and [ii] are identical for each $\mathbf{u}$, since the sequence of structural equations that eventually generate W are the same.

Finally, in order to measure $W_{\mathbf{t}}$, we need to measure the output of the mechanism f_W in the real world. The mechanism cannot not be erased and overwritten, as per condition [iii]. ■

Lemma C.7. *Given a set $\mathbf{W}_\star$ and graph $\mathcal{G}$, where each member $W_{\mathbf{t}} \in \mathbf{W}_\star$ has its respective conditions [i-iii] (per Lemma C.6), suppose these conditions introduce conflicts when combined across $\mathbf{W}_\star$. Removing X from $\mathbf{W}_\star$ and from all subscripts in $\mathbf{W}_\star$ to get a new set $\mathbf{W}'_\star$ does not introduce new conflicts between the terms, if X is first in a topological ordering of $\mathcal{G}$.*

Proof. Let us consider the conditions in the necessary-and-sufficient set given in Lemma C.6 for each $W_{\mathbf{t}} \in \mathbf{W}_\star$.

Condition [i] add requirements for each $t \in \mathbf{t}$ of some $W_{\mathbf{t}} \in \mathbf{W}_\star$. Since X is removed from every subscript, this no longer applies to any term in $\mathbf{W}'_\star$.

Condition [ii] requires that if X is an ancestor of some $W_{\mathbf{t}} \in \mathbf{W}_\star$, and it doesn't appear in $\mathbf{T}$, then X must be received without intervention by mediating children. Since X is removed from every subscript, X not being intervened upon at all meets this condition [ii] for every term $\mathbf{W}'_\star$, without conflicting with condition [i], which no longer applies.

Importantly, even though X is no longer being intervened upon, for each $W_{\mathbf{t}} \in \mathbf{W}_\star$, removing X from $\mathbf{T}$ (if it appears) does not add any additional ancestors in $\mathcal{G}_{\overline{T}}$ that need to be tracked for condition [ii], since X is first in topological order.

Condition [iii] would only apply to X itself, which is not present as a potential response in $\mathbf{W}'_\star$. ■

Theorem 3.5 :

Let $\mathcal{A}^{(i)}$ be a sequence of actions conducted by an exogenous agent to beget a vector of values $\mathbf{W}_\star^{(i)}$ for a unit i.

By Lemma C.5, if an agent wants $\mathbf{W}_\star^{(i)}$ to be an i.i.d sample from $P(\mathbf{W}_\star)$, then for each possible $\mathbf{U}^{(i)} = \mathbf{u}$, the vector $\mathbf{W}_\star^{(i)}$ needs to be identical to the $\mathbf{W}_\star(\mathbf{u})$ as evaluated according to the SCM. Essentially, this says that the agent's actions need to output the same vector $\mathbf{W}_\star(\mathbf{u})$ as if it has been evaluated according to the SCM.

By Lemma C.6, $\mathbf{W}_\star(\mathbf{u})$ can be evaluated if and only if the following three conditions are met for each $W_{\mathbf{t}} \in \mathbf{W}_\star$ simultaneously:

i $\mathbf{T}$ is fixed as $\mathbf{t}$ (by intervention) as an input to all children $C \in Ch(\mathbf{T}) \cap An(W)$;

ii Each $A \in An(W)_{\mathcal{G}_{\overline{\mathbf{T}}}}$, $A \notin \{\mathbf{T}, W\}$ is received "naturally" (i.e., without intervention) by its children $C \in Ch(A) \cap An(W)$; and

iii The mechanism f_W is not erased and overwritten (by a Fisherian intervention).

Inductive hypothesis (IH):

CTF-REALIZE$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A})$ returns FAIL if and only if conditions [i-iii] are not met simultaneously, when combined across all $W_\mathbf{t} \in \mathbf{W}_\star$ w.r.t a causal diagram $\mathcal{G}$ having $\leq n$ nodes

Base case:

Consider an SCM with only one variable $V \in \mathbf{V}$. IH is trivially true, since the conditions are always met, and since **CTF-REALIZE** will just return the value $\text{READ}(V)$.

Assume IH is true for any SCM with causal diagram having $\leq n$ nodes.

n+1 case:

Consider an SCM whose causal diagram $\mathcal{G}$ has $n+1$ nodes. Let X be the first in some topological ordering of $\mathcal{G}$. Consider an action set $\mathbb{A}$ that the agent can perform in the environment, and an arbitrary distribution $P(\mathbf{W}_\star)$.

WLOG, we can begin the outer loop of **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}))$ with X (first in topological order).

- The inner loop calls **COMPATIBLE**$(X, W_\mathbf{t})$ for each $W \in Desc(X), X \neq W$.
- It maintains a tracker INT_X of the smallest counterfactual interventions needed to satisfy condition [i] for each $W_\mathbf{t}$, resorting to Fisherian intervention if needed. Note (per Remark D.6), interventions follow a tree-like structure, so conflicts can be tracked by tagging the smallest available intervention that is needed for each $W_\mathbf{t}$ w.r.t each child of X.
- Note also (per Remark D.5) that if there are two simultaneous interventions added to INT_X, $\text{CTF-RAND}(X \to \mathbf{C})$, $\text{CTF-RAND}(X \to \mathbf{C}')$, where $\mathbf{C}' \subseteq \mathbf{C}$, then the set $\mathbf{C}'$ is unaffected by the first procedure.
- This inner loop exactly checks if there are any conflicts in conditions [i-ii] among $\mathbf{W}_\star$ w.r.t X, by "tagging" each procedure with the fixed value x needed for that intervention (including the requirement of no intervention).
- Finally the outer loop in Line 20 of Algo. 1 checks if X appears as a potential response anywhere in $\mathbf{W}_\star$. If so, INT_X cannot contain the requirement of Fisherian $\text{RAND}(X)$, since this violates condition [iii] w.r.t X.
- X does not appear anywhere else in subsequent algorithm iterations.

Thus, we conclude that **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}))$ does not return FAIL on the outer loops evaluated for X, if and only if there are no conflicts in the conditions [i-iii] for $\mathbf{W}_\star$ w.r.t X. *In other words, all conflicts w.r.t X, in the conditions [i-iii] combined across the terms in $\mathbf{W}_\star$, are identified in the algorithm steps that involve X.*

Next, define the new set $\mathbf{W}'_\star$ by dropping x from the subscript (if it appears) for each $W_\mathbf{t} \in \mathbf{W}_\star$, and dropping X from $\mathbf{W}_\star$ (if it appears). Since X is first in topological order of $\mathcal{G}$, this does not add any *new* conflicts across conditions [i-iii] induced by each term in $\mathbf{W}'_\star$ (by Lemma C.7).

It is also clear that if there are conflicts *not* involving X, that are induced by conditions [i-iii] across the terms in $\mathbf{W}_\star$, then these conflicts are also induced by $\mathbf{W}'_\star$. Suppose there are two terms $W_\mathbf{t}, Y_\mathbf{h} \in \mathbf{W}_\star$ s.t. $\mathbf{T} \setminus X$ needs to be received as $\mathbf{t} \setminus X$ by mediating children (condition [i]) for $W_\mathbf{t}$, and this conflicts with the requirement that $\mathbf{T} \setminus X$ needs to be received as $\mathbf{t}' \setminus X$ (or naturally) by the same mediating children, for $Y_\mathbf{h}$. Removing X does not affect this conflict, since X is topologically prior.

Next, define the graph $\mathcal{G}'$ as the projection of $\mathcal{G}$ that marginalizes out X (and adds bidirected edges if needed). $\mathcal{G}'$ has $\leq n$ nodes. Therefore, from the IH, we conclude that **CTF-REALIZE**$(P(\mathbf{W}'_\star), \mathcal{G}', \mathbb{A})$ does not return FAIL if and only if there are no conflicts induced by conditions [i-iii], combined across terms in $\mathbf{W}'_\star$.

Now, we note that **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A})$ is merely **CTF-REALIZE**$(P(\mathbf{W}'_\star), \mathcal{G}', \mathbb{A})$, plus all the steps involving X that we discussed earlier (can be verified from inspecting the algorithm - the former has an outer loop involving X and then contains the same steps as the latter). *Therefore, all conflicts induced by conditions [i-iii] that involve X and do not involve X are identified in the algorithm steps when run on $\mathcal{G}$ and $\mathbf{W}_\star$.*

Thus, we show that **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A})$ returns FAIL if and only if conditions [i-iii] are not met simultaneously, when combined across all $W_\mathbf{t} \in \mathbf{W}_\star$ w.r.t a causal diagram having $\leq n+1$ nodes. The IH stands proved.

By Lemma C.6, we know that conditions [i-iii] are necessary and sufficient to evaluate each term in $\mathbf{W}_\star(\mathbf{u})$ simultaneously, for any SCM compatible with $\mathcal{G}$. By Lemma C.5, we know that this is equivalent to drawing an i.i.d sample from $P(\mathbf{W}_\star)$. This gives us the proof of the theorem.

Note: we don't discuss the rejection sampling steps involved steps 17-18 of Algo. 1 as this is trivially equivalent to intervening using a fixed value. ■

Corollary 3.7 :

The proof intuition is as follows: given a graph $\mathcal{G}$ and a potential response $Y_\mathbf{x}$, the set of (counterfactual) ancestors of $Y_\mathbf{x}$ (Correa et al., 2021) lists each ancestor of Y and *what regime it must be realized in*, in order for $Y_\mathbf{x}$ to be evaluated. In other words $An(Y_\mathbf{x})$ tracks the regimes necessary and sufficient for its ancestors to be evaluated under to beget $Y_\mathbf{x}$.

For instance, in graph $\mathcal{G}_1$ in Fig. 3, in order to evaluate W_t, we need A_t to be evaluated in the regime $\mathcal{M}_t$. In order to evaluate Z_x, we need A, T to both be evaluated naturally. This reveals a conflict at the bottleneck f_A, which renders the distribution non-realizable.

Thus, Corollary 3.7 provides a *sufficient* condition to conclude that a distribution is non-realizable, if $An(\mathbf{W}_\star)$ contains two potential responses of the same variable under different regimes. It also becomes a *necessary* condition for non-realizability, if the agent can perform CTF-RAND$(X \to C)$, separately for each $C \in Ch(X)$, for all X. I.e., if the action set is $\mathbb{A}^\dagger(\mathcal{G})$.

The proof steps are similar to Theorem 3.5.

Inductive Hypothesis (IH):

Given a graph $\mathcal{G}$ with $\leq n$ nodes, and an arbitrary distribution $\mathbf{W}_\star$, **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ if and only if $An(\mathbf{W}_\star)$ does not contain a pair of potential responses $W_\mathbf{t}, W_\mathbf{s}$ of the same variable W under different regimes.

Base case:

For a graph containing only one variable Y, this is trivially true. $An(Y) = Y$, and the distribution $P(Y)$ is realizable.

Assume IH is true for a graph of $\leq n$ nodes.

n+1 case:

Consider an SCM whose causal diagram $\mathcal{G}$ has $n+1$ nodes. Let X be the first in some topological ordering of $\mathcal{G}$. The agent can perform $\mathbb{A}^\dagger$ in the environment, and the distribution is some arbitrary $P(\mathbf{W}_\star)$. WLOG, we can begin the outer loop of **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A})$ with X (first in topological order).

From Lemmas C.6 and C.5, we know that conditions [i-iii] for each $W_t \in \mathbf{W}_\star$, combined across $\mathbf{W}_\star$ form a necessary and sufficient set to realize $P(\mathbf{W}_\star)$.

Note that condition [iii] is always satisfied because the agent need never perform a Fisherian RAND(V) for any V. It can get the same effect by performing CTF-RAND$(V \to C)$ for each $C \in Ch(V)$. Step 12 of the sub-routine, Algo. 2 would never be invoked.

From Theorem 3.5, we know that **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ returns FAIL if and only if there are conflicts in conditions [i-ii] when combined across all the terms $\mathbf{W}_\star$.

Define the new set $\mathbf{W}'_\star$ by dropping x from the subscript (if it appears) for each $W_\mathbf{t} \in \mathbf{W}_\star$, and dropping X from $\mathbf{W}_\star$ (if it appears). Since X is first in topological order of $\mathcal{G}$, this does not add any *new* conflicts across conditions [i-ii] induced by each term in $\mathbf{W}'_\star$ (by Lemma C.7). It also doesn't *remove* any conflicts that are not related to X, as argued in the proof of Theorem 3.5, since X comes topologically first.

Define the graph $\mathcal{G}'$ as the projection of $\mathcal{G}$ that marginalizes out X (and adds bidirected edges if needed). $\mathcal{G}'$ has $\leq n$ nodes. From the IH, we conclude that **CTF-REALIZE**$(P(\mathbf{W}'_\star), \mathcal{G}', \mathbb{A}^\dagger(\mathcal{G}'))$ does not return FAIL if and only if $An(\mathbf{W}'_\star)$ does not contain two potential responses $W_\mathbf{t}, W_\mathbf{s}$ of the same variable under different regimes.

However, note that (as discussed in the proof of Theorem 3.5, and from inspecting the algorithm), the only difference between **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ and **CTF-REALIZE**$(P(\mathbf{W}'_\star), \mathcal{G}', \mathbb{A}^\dagger(\mathcal{G}'))$ is that in the former, the outer loop of **CTF-REALIZE** first checks for conflicts in the conditions [i-ii] across $\mathbf{W}_\star$ w.r.t X. After that, the steps for both algorithms are the identical.

Therefore, any conflicts detected by **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ that are *not* detected by **CTF-REALIZE**$(P(\mathbf{W}'_\star), \mathcal{G}', \mathbb{A}^\dagger(\mathcal{G}'))$ must be conflicts w.r.t X. By the IH, these additional conflicts (unrelated to X) cannot be because of a pair of conflicting potential responses in $An(\mathbf{W}'_\star)$.

We have already established that removing X to make $\mathbf{W}'_\star$ does not remove or add any conflicting potential response pairs that don't involve X. Therefore, our task is to now show that each of these additional conflicts (involving X) must correspond to at least one conflicting pair of potential responses in $An(\mathbf{W}_\star)$, that are not present in $An(\mathbf{W}'_\star)$. And conversely, we need to show that each pair of conflicting potential responses in $An(\mathbf{W}_\star)$ involving X (i.e., that is not present in $An(\mathbf{W}'_\star)$) corresponds to at least one conflict detected by **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ in the outer loop involving X.

<u>Forward direction</u>:

As discussed in the proof of Theorem 3.5, **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ returns FAIL in the outer loop involving X if and only if the input of X to some $C \in Ch(X)$ is required to be some x to satisfy condition [i] w.r.t some $W_\mathbf{t} \in \mathbf{W}_\star$, but also required to be x' or "natural" to satisfy condition [i/ii] w.r.t some $Y_\mathbf{h} \in \mathbf{W}_\star$.

Note that the action set is $\mathbb{A}^\dagger(\mathcal{G})$. Therefore step 6 of sub-routine Algo. 2 would always pick only the procedure CTF-RAND$(X \to C)$ whenever C needs to receive a fixed value. The interventions affecting other $C' \in Ch(X)$ would not affect C.

In this case, it is easy to see that the set $An(W_\mathbf{t})$ must contain $C_{x\ldots}$ per Def. 3.6, and $An(Y_\mathbf{h})$ must contain $C_{x'\ldots}$ or a potential response of C without X in the subscript. Thus, if **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ returns FAIL in the outer loop involving X, there must be a pair of conflicting potential responses in $An(\mathbf{W}_\star)$.

<u>Reverse direction</u>:

Assume there exists a conflicting pair of potential responses $A_\mathbf{t}, A_\mathbf{s} \in An(\mathbf{W}_\star)$ where $x \in \mathbf{t}$ and $\mathbf{s}$ contains some x' or does not contain X at all.

This means there is some $W_\mathbf{h} \in \mathbf{W}_\star$ s.t. $A \in An(W)_{\mathcal{G}_{\underline{X}}}$ and some $Y_\mathbf{j} \in \mathbf{W}_\star$ s.t. $A \in An(Y)_{\mathcal{G}_{\underline{X}}}$. I.e., A mediates the effect of X on W, Y. Further, from Def. 3.6, it means that A needs to be realized in conflicting regimes w.r.t X.

From Lemma C.6, such conflict happens because for $A_\mathbf{t}$, condition [i] requires that X is fixed by intervention to be x for all children $C \in Ch(X) \cap An(A)$. Whereas, for $A_\mathbf{s}$, condition [i] or [ii] requires that each child $C \in Ch(X) \cap An(A)$ receives X either fixed as x', or naturally (as the case may be, for $\mathbf{s}$). For any such $C \in Ch(X) \cap An(A)$, it is clear from the proof of Theorem 3.5 that this conflict will trigger a FAIL from **CTF-REALIZE**$(P(\mathbf{W}_\star), \mathcal{G}, \mathbb{A}^\dagger(\mathcal{G}))$ in the first outer loop involving X.

Thus, we have shown that the IH holds for any $\mathbf{W}_\star$ involving a graph with $n+1$ nodes.

Since Theorem 3.5 shows **CTF-REALIZE** is complete, we have thus proved that $\mathbf{W}_\star$ is realizable given $\mathcal{G}$ and $\mathbb{A}^\dagger(\mathcal{G})$ if and only if $An(\mathbf{W}_\star)$ does not contain a pair of conflicting potential responses for the same variable under different regimes.

■

Corollary 3.8 :

This follows from Corollary 3.7. For any causal diagram, the ancestral set of $\{Y_x, Y_{x'}\}$ would include both these potential responses.

Thus, the query is not realizable.

■

C.3 Realizability of $\mathcal{L}_1$- and $\mathcal{L}_2$-distributions

It is widely known and acknowledged that it is possible to draw samples from $\mathcal{L}_1$- and $\mathcal{L}_2$-distributions: the former by simply observing a system's natural behaviour, and the latter by intervening in the system through interventions like Fisherian randomization.

Still, we find it educational to derive these proofs from first principles. This sub-section is not strictly needed to follow the main contributions in Secs. 2 and 3.

We define the probability measure $P^{\mathbb{C}}(.)$ from the perspective of an exogenous agent (i.e., an agent external to the system) $\mathbb{C}$'s beliefs about the environment, distinguished by superscript from $P^{\mathcal{M}}(.)$, the true unknown distribution.

Since unit selection is randomized, $\textsc{Select}^{(i)}$ yields an unbiased sample of a unit with latent features distributed according to the target population frequency $P(\mathbf{u})$. I.e., $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \textsc{Select}^{(i)}) = P^{\mathcal{M}}(\mathbf{u})$. We also assume that target population size is large enough that $\textsc{Select}^{(i)}$ does not significantly change the distribution of the remaining population.

Further, we assume that the actions $\textsc{Read}(V)^{(i)}$ and $\textsc{Rand}(V)^{(i)}$ do not disrupt any other mechanism $f_{V'}$ for unit i.

Lemma C.8 (Observational sample)**.** *An agent $\mathbb{C}$ can draw an i.i.d sample distributed according to the $\mathcal{L}_1$ query $P(\mathbf{V})$ associated with an SCM $\mathcal{M}$, by the following actions:*

i. $\textsc{Select}^{(i)}$

ii. $\textsc{Read}(\mathbf{V})^{(i)} = \mathbf{v} \sim P(\mathbf{V})$

Given N i.i.d samples, the consistent unbiased estimate of $P(\mathbf{v})$ is

$$\hat{P}(\mathbf{v}) := \frac{1}{N} \sum_i \prod_{v \in \mathbf{v}} \mathbb{1}[\textsc{Read}(V)^{(i)} = v] \tag{14}$$

Proof. This follows directly from the definitions of the actions. $\textsc{Select}^{(i)}$ chooses a unit at random from the population. By Remark C.3, $P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \textsc{Select}^{(i)}) = P^{\mathcal{M}}(\mathbf{u})$. For randomly selected unit i,

$$P^{\mathbb{C}}(\textsc{Read}(\mathbf{V})^{(i)} = \mathbf{v} \mid \textsc{Select}^{(i)}) \tag{15}$$

$$= \sum_{\mathbf{u}} P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \textsc{Select}^{(i)}). \tag{16}$$

$$P^{\mathbb{C}}(\textsc{Read}(\mathbf{V})^{(i)} = \mathbf{v} \mid \mathbf{U}^{(i)} = \mathbf{u}, \textsc{Select}^{(i)}) \qquad \text{Chain rule}$$

$$= \sum_{\mathbf{u}} P^{\mathbb{C}}(\mathbf{U}^{(i)} = \mathbf{u} \mid \textsc{Select}^{(i)}).\mathbb{1}^{\mathcal{M}}[\mathbf{V}(\mathbf{u}) = \mathbf{v}] \qquad \text{Def. 2.1(ii)} \tag{17}$$

$$= \sum_{\mathbf{u}} P^{\mathcal{M}}(\mathbf{u}).\mathbb{1}^{\mathcal{M}}[\mathbf{V}(\mathbf{u}) = \mathbf{v}] \qquad \text{Rem. C.3} \tag{18}$$

$$= P^{\mathcal{M}}(\mathbf{v}) \qquad \text{Definition} \tag{19}$$

I.e., this record is an i.i.d. sample from $P^{\mathcal{M}}(\mathbf{V})$. Now consider the estimator below.

$$\hat{P}(\mathbf{v}) := \frac{1}{N}\sum_{n}\prod_{v\in\mathbf{v}} \mathbb{1}^{\mathbb{C}}[\textsc{Read}(V)^{(i)} = v] \tag{20}$$

$$= \frac{1}{N}\sum_{n}\sum_{\mathbf{u}}\prod_{v\in\mathbf{v}} \mathbb{1}^{\mathcal{M}}[\mathbf{U}^{(i)} = \mathbf{u}].\mathbb{1}^{\mathcal{M}}[V(\mathbf{u}) = v] \tag{21}$$

Un-biasedness is established by taking expectation on either side, w.r.t the agent $\mathbb{C}$'s actions (choice of units to observe):

$$\mathbb{E}_{\mathbb{C}}[\hat{P}(\mathbf{v})] = \mathbb{E}_{\mathbb{C}}\left[\frac{1}{N}\sum_{n}\sum_{\mathbf{u}}\prod_{v\in\mathbf{v}} \mathbb{1}^{\mathcal{M}}[\mathbf{U}^{(i)} = \mathbf{u}].\mathbb{1}^{\mathcal{M}}[V(\mathbf{u}) = v]\right] \tag{22}$$

$$= \sum_{\mathbf{u}}\frac{1}{N}\mathbb{E}_{\mathbb{C}}\left[\sum_{n}\mathbb{1}^{\mathcal{M}}[\mathbf{U}^{(i)} = \mathbf{u}]\prod_{v\in\mathbf{v}}.\mathbb{1}^{\mathcal{M}}[V(\mathbf{u}) = v]\right] \quad \text{Linearity of expectation} \tag{23}$$

$$= \sum_{\mathbf{u}}\frac{1}{N}\mathbb{E}_{\mathbb{C}}\left[\sum_{n}\mathbb{1}^{\mathcal{M}}[\mathbf{U}^{(i)} = \mathbf{u}]\right]\prod_{v\in\mathbf{v}}\mathbb{1}^{\mathcal{M}}[V(\mathbf{u}) = v] \quad V(\mathbf{u}) \text{ constant wrt } \mathbb{C} \tag{24}$$

$$= \sum_{\mathbf{u}}\frac{1}{N}\left[N.P^{\mathcal{M}}(\mathbf{u})\right]\prod_{v\in\mathbf{v}}\mathbb{I}^{\mathcal{M}}[V(\mathbf{u}) = v] \quad \text{Def. 2.1(i), Rem. C.3} \tag{25}$$

$$= P^{\mathcal{M}}(\mathbf{v}) \quad \text{Definition} \tag{26}$$

Consistency is established by the fact that as $\mathcal{N}$ (target population size) $\to \infty$, and N (sample size) $\to \infty$,

$$\frac{1}{N}\sum_{n}\mathbb{I}^{\mathcal{M}}[\mathbf{U}^{(i)} = \mathbf{u}] \to P^{\mathcal{M}}(\mathbf{u}) \tag{27}$$

■

Lemma C.9. *The $\mathcal{L}_2$ distribution of an atomic intervention is equivalent to the $\mathcal{L}_2$ distribution of the corresponding conditional stochastic intervention.*

$$P^{\mathcal{M}}(\mathbf{v}; do(\mathbf{x})) = P^{\mathcal{M}}(\mathbf{v}|\mathbf{x}; \sigma_{\mathbf{X}}) \tag{28}$$

$$= \sum_{\mathbf{u}}\underbrace{\mathbb{1}[\mathbf{V}_{\sigma_{\mathbf{X}}}(\mathbf{u}) = \mathbf{v} \mid X_{\sigma_{\mathbf{X}}} = \mathbf{x}]}_{①}.\underbrace{P(\mathbf{u})}_{②} \tag{29}$$

Proof. The step from the r.h.s of Eq. 28 to Eq. 29 is derived as follows: in the submodel $\mathcal{M}_{\sigma_{\mathbf{X}}}$, if we are given that $\mathbf{X}$ has been randomly assigned $\mathbf{x}$, then the remaining variables are deterministically generated as a function of $\mathbf{u}$ and $\mathbf{x}$ via their respective equations. The probability mass is collected over all the $\mathbf{u}$ which produce the output $\mathbf{v}$ over all these equations.

$$P^{\mathcal{M}}(\mathbf{v}|\mathbf{x}; \sigma_{\mathbf{X}}) = \sum_{\mathbf{u}}\mathbb{I}[\mathbf{V}_{\sigma_{\mathbf{X}}}(\mathbf{u}) = \mathbf{v} \mid X_{\sigma_{\mathbf{X}}} = \mathbf{x}].P^{\mathcal{M}}(\mathbf{u}) \tag{30}$$

Notice: if $\mathbf{v}$ is incompatible with $\mathbf{x}$, the indicator in the r.h.s evaluates to 0. Next, we prove. Eq. 28.

In $\mathcal{M}_{\sigma\mathbf{X}}$, as defined, $\mathbf{X}$ is assigned according to an independent random vector. Notate this vector as $\mathbf{X}_{\sigma_{\mathbf{X}}}$ and let the distribution of this vector be $P_{\sigma_{\mathbf{X}}}(\mathbf{X})$, defined by the assignment frequency over the target population.

$\mathcal{M}_{\sigma\mathbf{X}}$ is defined such that the target population is split into groups, each assigned $(X_{\sigma_{\mathbf{X}}} = \mathbf{x})$ for some $\mathbf{x}$. Note, the assignment vector $\mathbf{X}_{\sigma_{\mathbf{X}}}$ is independent of the latent features $\mathbf{U}$ across the target population iff each finite group assigned $(X_{\sigma_{\mathbf{X}}} = \mathbf{x})$ has the same distribution of latent features $P(\mathbf{U})$ as in the overall target population.

The above discussion handles the finite size of the target population. Starting with the r.h.s of Eq. 28,

$$P^{\mathcal{M}}(\mathbf{v}|\mathbf{x}; \sigma_{\mathbf{X}}) = \frac{P(\mathbf{v}, \mathbf{x}; \sigma_{\mathbf{X}})}{P(\mathbf{x}; \sigma_{\mathbf{X}})} = \begin{cases} P(\mathbf{v}; \sigma_{\mathbf{X}})/P(\mathbf{x}; \sigma_{\mathbf{X}}) & \text{if } \mathbf{v} \text{ compatible with } \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

Evaluating for when $\mathbf{v}$ is compatible with $\mathbf{x}$:

$$\frac{P(\mathbf{v};\sigma_{\mathbf{X}})}{P(\mathbf{x};\sigma_{\mathbf{X}})} = \frac{P(\mathbf{v};\sigma_{\mathbf{X}})}{P_{\sigma_{\mathbf{X}}}(\mathbf{x})} \tag{32}$$

$$= \frac{\sum_{\mathbf{u}} \left(P(\mathbf{u}) \prod_{V_i \in \mathbf{V}\setminus\mathbf{X}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i).P_{\sigma_{\mathbf{X}}}(\mathbf{x}) \right)}{P_{\sigma_{\mathbf{X}}}(\mathbf{x})} \quad \text{Truncated factorization product} \tag{33}$$

$$= \sum_{\mathbf{u}} P(\mathbf{u}) \prod_{V_i \in \mathbf{V}\setminus\mathbf{X}} P(v_i \mid \mathbf{pa}_i, \mathbf{u}_i) \tag{34}$$

$$= P^{\mathcal{M}}(\mathbf{v}; do(\mathbf{x})) \quad \text{Truncated factorization product} \tag{35}$$

Eq. 33 uses the fact that each sub-group assigned $(X_{\sigma_{\mathbf{X}}} = \mathbf{x})$, by independence, has the same frequency of latent features $P(\mathbf{u})$. ■

Lemma C.10 (Interventional sample)**.** *An agent $\mathbb{C}$ can draw an i.i.d sample distributed according to the $\mathcal{L}_2$ query $P(\mathbf{V}; do(\mathbf{x}))$ associated with an SCM $\mathcal{M}$, by the following actions:*

i. $\textsc{Select}^{(i)}$

ii. $\textsc{Rand}(\mathbf{X})^{(i)}$

iii. If $\textsc{Rand}(\mathbf{X})^{(i)} = \mathbf{x}$*, then* $\textsc{Read}(\mathbf{V})^{(i)} = \mathbf{v} \sim P(\mathbf{V}; do(\mathbf{x}))$*, else repeat i-iii.*

Given $N_{\mathbf{x}}$ i.i.d samples, the consistent unbiased estimate of Eq. 29 is given by

$$\hat{P}(\mathbf{v}; do(\mathbf{x})) = \underbrace{\frac{1}{N_{\mathbf{x}}} \sum_{i}}_{②} \underbrace{\mathbb{1}[\textsc{Read}(\mathbf{V})^{(i)} = \mathbf{v}, \textsc{Rand}(\mathbf{X})^{(i)} = \mathbf{x}]}_{①}, \tag{36}$$

Proof. The proof steps are similar to the ones used for the Observational i.i.d sample case. Note that Remark C.3 still hold since even though the agent is conditioning on the value randomly assigned to a particular unit i, this value is independent of the unit's latent features $\mathbf{U}^{(i)}$. ■

D DETAILS ON COUNTERFACTUAL RANDOMIZATION

In this appendix, we provide a formal account for the procedure we define in Sec. 2, called CTF-RAND$(X \to \mathbf{C})$ (Def. 2.3). This appendix is intended for the interested reader, and details some useful remarks to be used in the proofs of our results in Appendix C. These details are not strictly needed to follow the results presented in the main body of the paper.

- In Sec. D.1, we lay out the structural conditions under which it is possible to perform this procedure, and we provide an algorithm (Algorithm 3) by which an agent can translate the structural conditions in the environment into a list of CTF-RAND procedures that it is able to perform in the given setting.
- In Sec. D.2, we emphasize that it is possible for an agent to enact multiple randomization procedures involving the same variable X for a single unit i, and illustrate this with an example.
- In Sec. D.3, we discuss the constraints implied by the assumptions we make. In particular, we discuss why CTF-RAND$(X \to \mathbf{C})$ can only be performed on some $\mathbf{C} \subseteq Ch(X)$, and not by-pass children to directly affect some distant descendants of X.

D.1 STRUCTURAL CONDITIONS REQUIRED FOR COUNTERFACTUAL RANDOMIZATION

Counterfactual randomization (Def. 2.3) can be performed under two circumstances:

i. CTF-RAND$(X \to Ch(X))$ can be performed by eliciting a unit's natural decision X, while simultaneously randomizing its actual enforced decision. Thus, the agent can affect the value of the decision X as received by all the children of X, whilst also recording the natural realization of X. As discussed in Sec. 2, this was established in (Bareinboim et al., 2015; Forney et al., 2017; Zhang & Bareinboim, 2022).

ii. CTF-RAND$(X \to \mathbf{C})$ can also be performed for some $\mathbf{C} \subseteq Ch(X)$ if there is a special *counterfactual mediator* (defined below) by which the mechanisms generating $\mathbf{C}$ perceive the value of X. This counterfactual mediator then allows the agent to intervene on the value of X as *perceived* by $\mathbf{C}$, thus mimicking an actual intervention on X.

Definition D.1 (Expanded SCM)**.** Given an SCM $\mathcal{M}$ containing observable variables $\mathbf{V}$, we define an *expanded SCM* $\mathcal{M}^+$ of the same environment to be a model containing a bigger set of observable variables $\mathbf{V}^+ \supset \mathbf{V}$, and which relaxes the positivity requirement. I.e., it is possible that $P^{\mathcal{M}^+}(\mathbf{v}^+) = 0$, for some $\mathbf{v}^+$ in $\mathcal{L}_1$. We call the causal diagram of $\mathcal{M}^+$ an *expanded* causal diagram $\mathcal{G}^+$. ■

Definition D.2 (Counterfactual mediator (formal))**.** Given a variable X in a causal diagram $\mathcal{G}$, we call any variable $W \notin \mathbf{V}$ a *counterfactual mediator* of X w.r.t $Y \in Ch(X)_{\mathcal{G}}$ if

i. In an "expanded" SCM of the environment $\mathcal{M}^+$ (Def. D.1), W is generated according to an invertible mechanism $W \leftarrow f_W(X, \mathbf{U}_W)$ with $\mathbf{U}_W$ possibly empty, s.t. $f_W^{-1}(W) = X$;

ii. It is physically possible to perform RAND$(W)^{(i)}$ (Def. 2.1); and

iii. In $\mathcal{M}^+$, Y is generated by the mechanism $Y \leftarrow f_Y(f_W^{-1}(W), \mathbf{A}, \mathbf{U}_Y)$, where $\mathbf{A}$ is the set $\mathbf{Pa}_Y \setminus X$ in $\mathcal{G}$. ■

The intuition behind Def. D.2 is that a *counterfactual mediator* is a real variable in the environment which fully encodes information about the variable X, and which mediates how Y perceives the value of X via the "direct" causal path. For instance, in Example 1 (*Mediation analysis*), the RGB values of the video frames W are a counterfactual mediator for the mechanism f_Y (decision to issue a speeding ticket) to perceive the car's color X via the "direct" path, not via the actual speeding of the car).

Condition [i] of Def. D.2 divides the domain of W into *equivalence classes* s.t. each value w belongs to an equivalence class $\{w' : f_W^{-1}(w) = x\}$ for some value x.

Condition [iii] of Def. D.2 essentially says that the mechanism f_Y only cares about which equivalence class W belongs to. I.e., Y only cares about what W reveals about X.

Note: these conditions does not require an agent to have full knowledge of the SCM. They are rather structural assumptions about the underlying mechanisms which can be verified in a given setting. In Example 1, treating W as counterfactual mediator means the assumption that (1) the video features W uniquely map back to the actual color of the car in the footage; and (2) the computer vision system only cares W reveals about X, and is indifferent to any stochasticity *within* some equivalence class $\{w' : f_W^{-1}(w) = x\}$.

Assumption D.3. (Tree structure) Given a variable X, causal diagram $\mathcal{G}$, and an "expanded" diagram $\mathcal{G}^+$ (Def. D.1) including the set of all the counterfactual mediators $\mathbf{W}$ (Def. D.2) of X in the environment, each $W \in \mathbf{W}$ has only one parent in $\mathcal{G}^+$, and each $C \in Ch(X)_{\mathcal{G}}$ has at most one $W \in \mathbf{W}$ as a parent in $\mathcal{G}^+$. ■

Assumption D.3 enforces that each child of X perceives X through at most one proxy pathway. This assumption rules out possible structures like Fig. 15(a) where a child perceives X through multiple proxy pathways.

This assumption is general enough to allow most cases of interest. If X is a construct like gender identity, then it is possible that a child perceives X via a cluster of personal attributes $\mathbf{W}$ which indicate X. In this case, no single attribute solely satisfies Def. D.2 of a counterfactual mediator. However, the cluster of attributes $\mathbf{W}$ could be collapsed into a single variable having domain equal to the cartesian product of the sub-domains (Anand et al., 2023; Xia & Bareinboim, 2024). This single node $\mathbf{W}$ would indeed satisfy the definition of a counterfactual mediator and would comply with the tree structure in Assumption D.3. For a comprehensive discussion of the semantics of interventions on the perception of a compound attribute such as race or gender identity, see (Plecko & Bareinboim, 2024, App. D.1). The following Lemma is the key property that enables path-specific randomization.

Lemma D.4. *Given a causal diagram $\mathcal{G}$ containing variables X and $Y \in Ch(X)_{\mathcal{G}}$. Let W be a counterfactual mediator of X w.r.t Y (Def. D.2). For any value x, we have*

$$Y_{w\mathbf{a}}(\mathbf{u}) = Y_{x\mathbf{a}}(\mathbf{u}), \qquad \forall \mathbf{u}, \forall w \in \{w' : f_W^{-1}(w) = x\}, \tag{37}$$

where $\mathbf{A} := \mathbf{Pa}_Y \setminus X$ in $\mathcal{G}$.

Proof. This follows from Def. D.2. Suppose we are given values (w, x) where $f_W^{-1}(w) = x$. Let $\mathbf{A} := \mathbf{Pa}_Y \setminus X$ in $\mathcal{G}$.

The variable $W_x(\mathbf{u}) = W_{x\mathbf{a}}(\mathbf{u}) = f_W(x, \mathbf{u})$, in the enhanced submodel $\mathcal{M}^+_{x\mathbf{a}}$. Adding $\mathbf{a}$ to the subscript does not matter - by Assumption D.3 and Lemma D.7, $\mathbf{A}$ cannot be an ancestor of W in $\mathcal{M}^+$.

Since f_W is invertible by condition [i] in $\mathcal{M}^+$, it is also invertible in submodel submodel $\mathcal{M}^+_{x\mathbf{a}}$. Therefore, we have $f_W^{-1}(W_{x\mathbf{a}}(\mathbf{u})) = x$.

$$\begin{aligned} Y_{w\mathbf{a}}(\mathbf{u}) &= f_Y(f_W^{-1}(w), a, \mathbf{u}) \\ &= f_Y(x, a, \mathbf{u}) \\ Y_{W_{x\mathbf{a}}\mathbf{a}}(\mathbf{u}) &= f_Y(f_W^{-1}(W_{x\mathbf{a}}(\mathbf{u})), a, \mathbf{u}) \\ &= f_Y(x, a, \mathbf{u}) \end{aligned}$$

The r.h.s is identical, giving us $Y_{w\mathbf{a}} = Y_{W_{x\mathbf{a}}\mathbf{a}}$. Finally, we argue that $Y_{W_{x\mathbf{a}}\mathbf{a}} = Y_{x\mathbf{a}}$.

The counterfactual $Y_{x\mathbf{a}}$ is evaluated in a submodel of $\mathcal{M}^+$, where f_W receives input x and this value of W_x is an input to f_Y, while $\mathbf{A}$ is fixed to be $\mathbf{a}$. Structurally, this is identical to how the counterfactual $Y_{W_{x\mathbf{a}}\mathbf{a}} = Y_{W_x\mathbf{a}}$ is evaluated. Therefore, it is evident that $Y_{W_x\mathbf{a}} = Y_{x\mathbf{a}}$. ■

Given a variable X, the way an agent actually performs the action CTF-RAND is as follows:

i. **Performing CTF-RAND by eliciting natural decision:** The agent can perform CTF-RAND$(X \to Ch(X))^{(i)}$ by randomizing the unit's decision. The agent can further perform READ$(X)^{(i)}$ to elicit the unit's natural decision, which has not been erased. This is described in in (Bareinboim et al., 2015; Forney et al., 2017; Zhang & Bareinboim, 2022).

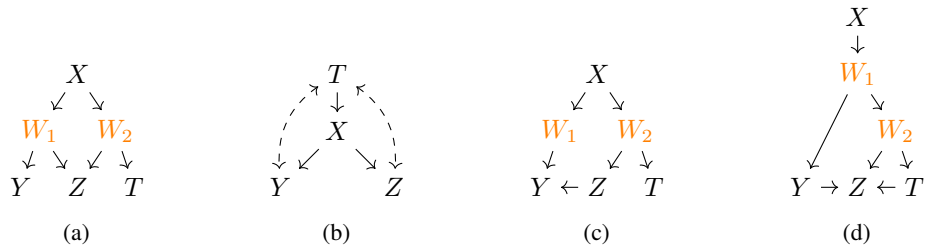


Figure 15: "Expanded" causal diagrams, where counterfactual mediators (labeled W_i) of X have been marked in red. (a) is not permitted by Assumption D.3. Assume the environment in diagram (b) permits the agent to elicit the unit's natural decision X even when randomizing the actual decision.

ii. **Performing CTF-RAND using counterfactual mediators:** If X has a counterfactual mediator W in the environment, and $\mathbf{C} \subseteq Ch(X)$ are the children which perceive X via W, then the agent can perform CTF-RAND$(X \to \mathbf{C})^{(i)}$ by randomizing W. Each value w mimics randomizing X as perceived by $\mathbf{C}$, per Lemma D.4. The agent can still perform READ$(X)^{(i)}$ by measuring $X^{(i)}$ to get the unit's natural decision, which has not been erased.

Having described the structural conditions that permit counterfactual randomization, we want to abstract away the mediators and succinctly describe the agent's physical actions via the definition of CTF-RAND. Given a variable X, and assumptions/knowledge about X in the environment stated in points **[i]** and **[ii]** above, we translate this knowledge into a set of counterfactual randomizations that the agent is physically able to perform in the environment, using Algorithm 3.

Algorithm 3 CTF-PROCEDURES

Input: Causal diagram $\mathcal{G}$ with decision variable X; "expanded" diagram $\mathcal{G}^+$ (Def. D.1) including the counterfactual mediators of X in the environment
Output: $\mathbb{A}_X$ - the set of CTF-RAND actions that can be performed involving X
$\mathbb{A}_X \leftarrow \emptyset$
if environment allows eliciting natural decision X even when randomizing actual decision **then**
  if X can be randomized **then**
    $\mathbb{A}_X \leftarrow \mathbb{A}_X \cup \{\text{CTF-RAND}(X \to Ch(X)_{\mathcal{G}})\}$
  end if
end if
for each counterfactual mediator W of X **do**
  Let $\mathbf{C} := \{C \mid C \in Ch(X)_{\mathcal{G}}$ and perceives X via $W\}$
  $\mathbb{A}_X \leftarrow \mathbb{A}_X \cup \{\text{CTF-RAND}(X \to \mathbf{C})\}$
end for
Return $\mathbb{A}_X$

Consider Fig. 15(a-d). (a) is not permitted by Assumption D.3. We assume that in the environment represented by (b) X can be randomized for a unit in the target population without erasing the unit's natural decision, satisfying condition **[i]** mentioned earlier. Thus, when applying Algorithm 3 to diagrams (b-d), we get the following resulting set of counterfactual randomization procedures which are permitted by the structural assumptions made (unit superscript i is omitted for legibility):

(b) $\{\text{CTF-RAND}(X \to \{Y, Z\})\}$
(c) $\{\text{CTF-RAND}(X \to Y), \text{CTF-RAND}(X \to \{Z, T\})\}$
(d) $\{\text{CTF-RAND}(X \to \{Y, Z, T\}, \text{CTF-RAND}(X \to \{Z, T\})\}$

D.2 MULTIPLE SIMULTANEOUS RANDOMIZATIONS ARE POSSIBLE, FOR A SINGLE UNIT

For a particular decision variable X, there could be multiple randomization procedures which an agent can perform. Consider the example in Fig. 16. The "expanded" diagram on the left shows two counterfactual mediators, W_1, W_2 in a causal structure which permit an agent to perform all

of the following randomization procedures: $\text{RAND}(X)^{(i)}$, $\text{CTF-RAND}(X \to \{Z, T, B\})^{(i)}$ and $\text{CTF-RAND}(X \to \{T, B\})^{(i)}$ for the same unit i.

However, if all three actions are performed in parallel, randomizing W_1 to enact $\text{CTF-RAND}(X \to \{Z, T, B\})^{(i)}$ will only affect variable Z. This is since the action of randomizing W_2 to further enact $\text{CTF-RAND}(X \to \{T, B\})^{(i)}$ blocks any effect on T, B from the previous action. Similarly, $\text{RAND}(X)^{(i)}$ ends up affecting only variable Y, because $\text{CTF-RAND}(X \to \{Z, T, B\})^{(i)}$ blocks any effect from the previous action on Z, T, B. We formalize this observation in Remark D.5.

Remark D.5 (Superseding action). Given a decision variable X, the action $\text{CTF-RAND}(X \to \mathbf{C}')^{(i)}$ can *supersede* the action $\text{CTF-RAND}(X \to \mathbf{C})^{(i)}$ if $\mathbf{C}' \subsetneq \mathbf{C}$, where *supersede* means that the former action $\text{CTF-RAND}(X \to \mathbf{C}')^{(i)}$ blocks any effect that the latter action has on the variables $\mathbf{C}'$. Additionally, the action $\text{CTF-RAND}(X \to \mathbf{C})^{(i)}$ *supersedes* the action $\text{RAND}(X)^{(i)}$. ■

Further, Assumption D.3 ensures that all such procedures follow a "nested" structure. I.e., given any two randomization procedures involving the same variable, the sets of children affected by one will be a subset of the set affected by the other, as shown in Fig. 16.

Remark D.6. (Procedure containment) Assumption D.3 implies that if an agent is capable of performing both $\text{CTF-RAND}(X \to \mathbf{C})^{(i)}$ and $\text{CTF-RAND}(X \to \mathbf{C}')^{(i)}$ s.t. $\mathbf{C} \neq \mathbf{C}'$ and $\mathbf{C} \cap \mathbf{C}' \neq \emptyset$, then either $\mathbf{C} \subseteq \mathbf{C}'$ or $\mathbf{C}' \subseteq \mathbf{C}$. ■

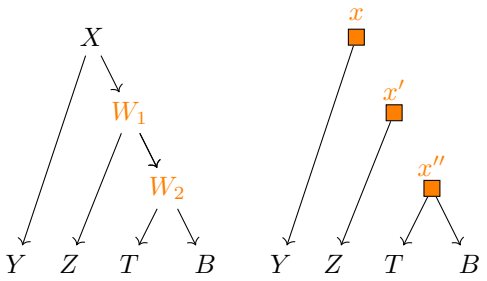


Figure 16: (Left) "Expanded" causal diagram showing counterfactual mediators W_1, W_2 of X; (Right) Agent performing actions $\text{RAND}(X)^{(i)}$, $\text{CTF-RAND}(X \to \{Z, T, B\})^{(i)}$ and $\text{CTF-RAND}(X \to \{T, B\})^{(i)}$ all together on the single unit i.

D.3 Counterfactual Randomization is only possible for direct children of X

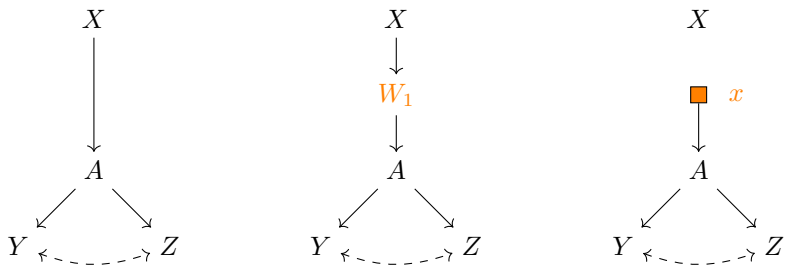
Our definition of $\text{CTF-RAND}(X \to \mathbf{C})^{(i)}$, is only valid for some $\mathbf{C} \subseteq Ch(X)$ in the causal diagram (Def. 2.3). This action essentially randomizes the value of decision variable X *as an input* to the mechanisms generating its causal children $\mathbf{C}$, while leaving open the possibility of measuring the unit i's natural decision (what it would have normally decided in the $\mathcal{L}_1$ regime), and also the possibility of separately and in parallel randomizing the value of X as an input to other causal children $\mathbf{C}' = Ch(X) \setminus \mathbf{C}$.

However, the notion of "child" is an abstraction w.r.t a specific diagram of the environment under study. Consider Fig. 17(a-Left), where $\mathcal{G}_1$ is the diagram of some environment. Assume there exists a counterfactual mediator W_1 of X (Def. D.2) as shown in Fig. 17(a-Middle), which means an agent is able to perform the physical action $\text{CTF-RAND}(X \to A)^{(i)}$, while still being able to measure the natural value of X for unit i.

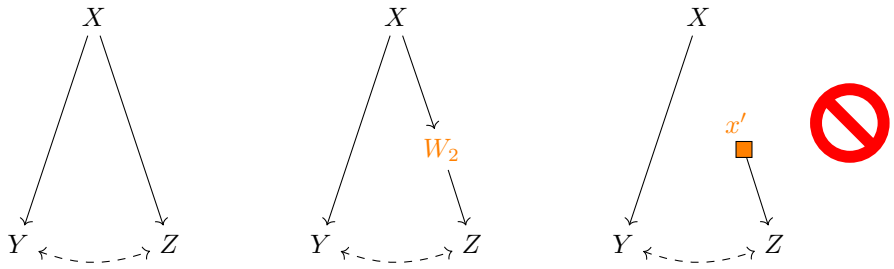
Now consider the diagram $\mathcal{G}_2$ shown in Fig. 17(b-Left). $\mathcal{G}_2$ is a valid projection of $\mathcal{G}_1$ obtained by marginalizing out variable A, and is thus also a valid causal diagram of the environment.

Suppose that there exists a counterfactual mediator W_2 as shown in 17(b-Middle). This means that the agent can also perform $\text{CTF-RAND}(X \to Z)^{(i)}$ in the same environment. However, since we are referring to the same environment, this means that the agent is able to perform $\text{CTF-RAND}(X \to Z)^{(i)}$ w.r.t the diagram $\mathcal{G}_1$, where Z is not a child node of X! This would translate to even greater experimental power w.r.t graph $\mathcal{G}_1$, where an agent is able to perform counterfactual randomization of X w.r.t further descendants like Z and draw i.i.d samples from queries like $P(A_x, Z_{x'})$ by simultaneously performing both counterfactual randomizations (i.e. by randomizing W_1, W_2 simultaneously).

However, this scenario is not possible. Essentially, this would require an "expanded" causal diagram (Def. D.1) like shown in Fig. 18, where W_2 is a counterfactual mediator of X w.r.t Z that comes after another variable A. If A satisfies positivity w.r.t X, i.e., if $P^{\mathcal{M}}(x, a) > 0, \forall x, a$ in $\mathcal{L}_1$, then W_2 cannot be a counterfactual mediator since it cannot be uniquely mapped back to X.



(a) : (Left) Graph $\mathcal{G}_1$; (Middle) enhanced diagram showing counterfactual mediator of X; (Right) agent performing CTF-RAND$(X \to A)^{(i)}$ using W_1



(b) : (Left) Graph $\mathcal{G}_2$; (Middle) enhanced diagram showing counterfactual mediator of X; (Right) agent performing CTF-RAND$(X \to Z)^{(i)}$ using W_2

Figure 17: Given a causal diagram $\mathcal{G}_1$ of the environment, $\mathcal{G}_2$ is a valid projection of $\mathcal{G}_1$ (marginalizing A). If A satisfies positivity w.r.t X, then there *cannot* be a counterfactual mediator W_2 as shown in (b-Middle). Which means an agent *cannot* perform CTF-RAND$(X \to Z)^{(i)}$ as shown in (b-Right).

Lemma D.7. *Given a causal diagram $\mathcal{G}$ of a true SCM $\mathcal{M}$ with a variable X and $A \in Desc(X)_{\mathcal{G}}$ where $P(x, a) > 0, \forall x, a$. There cannot be a variable W in an "expanded" SCM $\mathcal{M}^+$ of the environment (Def. D.1) s.t.*

- *$W \in Desc(A)_{\mathcal{G}^+}$, where $\mathcal{G}^+$ is the "expanded" causal diagram of $\mathcal{M}^+$; and*
- *W is invertible to X, i.e. exists f_W^{-1} s.t. $f_W^{-1}(W) = X$.* □

Proof. If A satisfies positivity w.r.t X, then a given value w cannot be mapped back to a unique x, even if we marginalize out A from the SCM.

Note that, by Assumption D.3, a counterfactual mediator has only one parent in the "expanded" causal diagram (Def. D.1). I.e., if it were a descendant of A, its perception of X is fully mediated by A.

If $f'_W(X, \mathbf{U})$ is invertible from W to X, then so is $f_W \circ f_A(X, \mathbf{U})$. It is evident that $f_W'^{-1}$ is well defined iff $f_A^{-1} \circ f_W^{-1}$ is well defined.

f_A^{-1} is not defined. The positivity condition entails that a given value a could have been generated by any value x (when unit is unknown).

Since f'_W is not invertible, W cannot be a counterfactual mediator. ■

Lemma D.7 leads to some important conclusions.

Remark D.8. There cannot be an "expanded" causal diagram (such as in Fig. 18), with a counterfactual mediator that bypasses a child-node and directly fixes a descendent-node's perception of X. I.e., an agent cannot perform CTF-RAND$(X \to D)^{(i)}$ for some $D \in Desc(X) \setminus Ch(X) \cup \{X\}$. ■

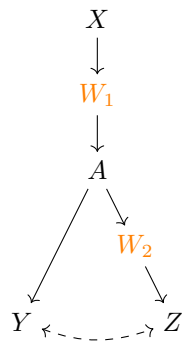


Figure 18: "Expanded" diagram that is needed to sample directly from $P(A_x, Z_{x'})$. This is not possible, per Lemma D.7.

Remark D.9. Conversely, given a graph like $\mathcal{G}_2$ in Fig. 17(b), if we are told that the agent can perform the action CTF-RAND$(X \to Z)^{(i)}$, then $\mathcal{G}_2$ cannot be a projection of $\mathcal{G}_1$ (Fig. 17(a)) for the same environment. ■

The upshot of this discussion is that, in general (i.e. without making further assumptions), counterfactual randomization can only be done via counterfactual mediators (Def. D.2) of a decision variable X, and it can only be performed on the children-nodes of X in the general case.

E Details on Examples

E.1 Example 1 (Mediation analysis) - further discussion

For the interested reader who wants to track the formal definition of a counterfactual mediator (Def. D.2), we provide below a discussion of how these structural assumptions may be verified in a field experiment setting.

This discussion is not strictly needed for following the main contributions in Secs. 2 and 3.

Verification of structural assumptions: In order for W to be a counterfactual mediator of X w.r.t Y, it needs to satisfy 3 conditions in Def. D.2.

Condition [i] says that each w (say, a specific RGB range of pixels) belongs to an equivalence class that maps back uniquely to a car color x. This can be verified using the RCT data. **Condition [ii]** is satisfies since they can do a targeted randomization of W. **Condition [iii]** stipulates that f_Y is not affected by any artefacts introduced by the color-editing tool: this can be verified, for instance, by swapping a car's color from x to x' and then swap it back from x' to x, to ensure that the model's decision Y does not change, thus verifying that the mechanism f_Y only cares about what the color features W reveal about X, and not about any image artefacts that may be introduced by editing.

E.2 Example 2 (Causal fairness)

E.2.1 Simulations

The details for the simulations shown in Fig. 5(c) are as follows. We first parameterize the space of SCMs that are compatible with the causal diagram in Fig. 5(a) using *canonical parameters* (Zhang et al., 2022, Def. 1, Thm. 1).

These parameters essentially discretize the domain of the latent confounder between Y and Z where each value of the confounder represents a joint mapping in $(X \to Y) \times (X \to Z)$. I.e., $2^2 \times 2^2 = 16$ values. We then set up a constrained optimization program to draw samples from the space of all SCMs (i.e., from the space of all trained models in Example 2) which satisfy either

- $\mathcal{L}_2$ fairness measure $\mu_{int1} + \mu_{int2}$ being penalized; or
- $\mathcal{L}_3$ fairness measure μ_{ctf} being penalized

Drawing 1000 samples from each gives us the chart in Fig. 5(c).

E.2.2 SCM specification

Just to give intuition for how this disparity can arise, we present below an instantiation of an SCM using canonical parameters, showing how $\mathcal{L}_2$ measures misleadingly suggest no discrimination, whereas the $\mathcal{L}_3$ measure actually detects unfairness.

For simplicity, we assume X is binary (0 indicates Race A, 1 indicates Race B). Y, Z are binary outcomes indicating, respectively, the CV passing through the 1st stage of screening for college admission, and for receiving a financial scholarship.

U_X is the random assignment of applicant race in the CV, in the absence of intervention.

The mechanisms f_Y and f_Z represent the decisions of the trained classifiers used by the college data science team, which have been trained using data from previous years' decisions of committees for CV screening and financial aid.

SCM $\mathcal{M}^\star$

$$X \leftarrow U_X \sim Bernoulli(0.5)$$

$$Y \leftarrow \begin{cases} 1, & \text{if Y-type = always-approve} \\ X, & \text{if Y-type = approve iff } x_1 \\ 1-X, & \text{if Y-type = approve iff } x_0 \\ 0, & \text{if Y-type = always-reject} \end{cases}$$

$$Z \leftarrow \begin{cases} 1, & \text{if Z-type = always-approve} \\ X, & \text{if Z-type = approve iff } x_1 \\ 1-X, & \text{if Z-type = approve iff } x_0 \\ 0, & \text{if Z-type = always-reject} \end{cases}$$

Y-type	**Z-type**	$P(U_{YZ})$
Always-approve	Always-approve	0.040
Always-approve	Approve iff x_1	0.175
Always-approve	Approve iff x_0	0.160
Always-approve	Always-reject	0.010
Approve iff x_1	Always-approve	0.040
Approve iff x_1	Approve iff x_1	0.055
Approve iff x_1	Approve iff x_0	0.170
Approve iff x_1	Always-reject	0.010
Approve iff x_0	Always-approve	0.040
Approve iff x_0	Approve iff x_1	0.140
Approve iff x_0	Approve iff x_0	0.025
Approve iff x_0	Always-reject	0.025
Always-reject	Always-approve	0.050
Always-reject	Approve iff x_1	0.010
Always-reject	Approve iff x_0	0.025
Always-reject	Always-reject	0.025

The CV bodies of fake applicants used in the holdout set are divided into "canonical types" such that each type elicits an approval/reject response from the models f_Y and f_Z. Factors influencing this decision could be the prejudice of the committees, the accomplishments listed on the CV etc. (since those preferences went into building the model). U_{YZ} represents of the distribution of these CV types. There are 16 such types, based on the 4 types each per model, as shown above.

For instance, row 1 of the probability table indicates a CV body such that both f_Y and f_Z would approve such a candidate regardless of the perceived race. Row 2 indicates a CV body such that f_Y would always approve such a candidate, but f_Z is biased to only approve such a candidate if they belonged to Race B.

We want to track whether, given a candidate of Race B who passed the CV screening but was rejected for financial aid, this candidate would still be denied financial aid had they been of Race A. In

particular, they care about the fairness metric μ_{ctf}, defined as follows.

$$\mu_{ctf} := P(Y_{x_1} = 1, Z_{x_1} = 0) - P(Y_{x_1} = 1, Z_{x_0} = 0) = P(y_x, z'_x) - P(y_x, z'_{x'}) = 10\% \quad (38)$$

The actual value of this measure can be computed from the probabilities in the SCM above. In practice, μ_{ctf} can be directly estimated using the counterfactual randomization procedure illustrated above in Fig. 5(b).

If the college data scientists instead follow the standard procedure of using only $\mathcal{L}_2$-data from a Fisherian RCT, they can only estimate $\mathcal{L}_2$ fairness metrics, such as μ_{int1}, μ_{int2} defined below.

$$\mu_{int1} := P(y_x).P(z'_x) - P(y_x).P(z'_{x'}) = 0 \quad (39)$$

$$\mu_{int2} := P(y, z'; do(x)) - P(y, z'; do(x')) = 0 \quad (40)$$

The $\mathcal{L}_2$-metrics μ_{int1}, μ_{int2} show no issues with fairness, and thus fail to capture the insight obtained from the $\mathcal{L}_3$-metric $\mu_{ctf} = 10\%$. This counterfactual insight helps the college to quantitatively characterize the financial hurdles faced by different racial groups in accessing college education, and to prevent unfair disparities.

E.3 Example 3 (Counterfactual bandit policies)

The SCM used in this hypothetical scenario to generate data is as follows:

$$U_1 \sim Bernoulli(0.5)$$
$$U_2 \sim Bernoulli(0.5)$$
$$U_3 \sim Bernoulli(0.5)$$

$$X \leftarrow U_1 \oplus U_2 \qquad (\oplus \text{ is the XOR function})$$
$$D \leftarrow X \oplus U_3$$

Since Y is a function of X, the average outcome is shown below for different realizations of the latents

i. Avg. Y, when $U_3 = 0$

	$U_3 = 0$			
	$U_1 = 0$		$U_1 = 1$	
	$U_2 = 0$	$U_2 = 1$	$U_2 = 0$	$U_2 = 1$
$do(x_0)$	**0.6**	0.9	0.8	**0.5**
$do(x_1)$	0.9	**0.6**	**0.5**	0.8

Natural choice of X marked **bold**

ii. Avg. Y, when $U_3 = 1$

	$U_3 = 1$			
	$U_1 = 0$		$U_1 = 1$	
	$U_2 = 0$	$U_2 = 1$	$U_2 = 0$	$U_2 = 1$
$do(x_0)$	**0.8**	0.7	0.6	**0.7**
$do(x_1)$	0.7	**0.8**	**0.7**	0.6

iii. Avg. Y, with U_3 marginalized (consolidating i. and ii.)

	$U_1 = 0$		$U_1 = 1$	
	$U_2 = 0$	$U_2 = 1$	$U_2 = 0$	$U_2 = 1$
$do(x_0)$	**0.7**	0.8	0.7	**0.6**
$do(x_1)$	0.8	**0.7**	**0.6**	0.7

Let us call the social media user Alice, for ease of reference. U_1, U_2, U_3 are latent attributes affecting Alice's decisions each evening. In particular, U_1 indicates whether she is tired, U_2 indicates whether she had a busy day and is distracted, U_3 indicates whether she is hungry, on any given evening.

If Alice is either tired but mentally relaxed ($X = 1 \oplus 0$), or if she is physically energetic but distracted ($X = 0 \oplus 1$), Alice decides to take a walk and use social media via mobile app. If Alice is neither tired nor distracted, she prefers to continue working on her desktop and uses social media via desktop app during breaks ($X = 0 \oplus 0$). If she is both tired and distracted, she also decides to use the social media app on her desktop because she has no energy to take a walk ($X = 1 \oplus 1$).

There are so many possible factors affecting her decisions, Alice is unaware that these are the specific unconscious causes of her natural choices. However, the social media company's unscrupulous data scientists surveil U_1, U_2, U_3 (perhaps by tracking Alice's wearable health monitor and calendar) and predict her natural choice. The company then uses behavioural insights to ping Alice with the precise notifications and content to maximize her time spent on the platform for each realization of U_1, U_2, U_3.

D is the type of ads Alice sees when she logs in to the social media app for the day.

The detailed causal diagram is shown in Fig. 19.

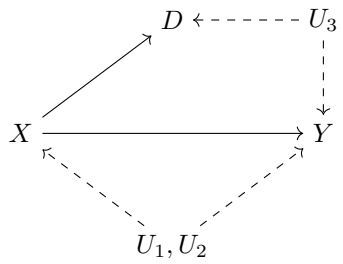


Figure 19: Causal diagram for Example 3. X : app usage type; D : advertisement-type received; Y : compliance with app usage time-limit; U_1 : agent tiredness; U_2 : agent busyness earlier in the day; U_3 : indicator of whether the agent is hungry.

$\mathcal{L}_1$-regime: The observational data is contained in Table (iii) in the SCM above, where the bold values correspond to Alice's natural choices. Given that all combinations of latents happen with equal probability, it is easy to see that the expected reward in the observational regime is $E[Y] = (0.25)(0.7 + 0.7 + 0.6 + 0.6) = 0.65$.

$\mathcal{L}_2$-regime: Applying the interventions $do(x_0), do(x_1)$, we can compute the expected outcome from the SCM as shown in Table 2. This is simply the average of all the values in Table (iii) of the SCM above.

	$E[Y; do(x)]$
$do(x_0)$	0.7
$do(x_1)$	0.7

Table 2: Expected outcome $E[Y_x]$ computed under the interventional regime.

An interventional strategy of randomizing ones actions (or fixing a constant action) outperforms the observational $\mathcal{L}_1$ regime of allowing one's actions to be determined by natural inclination.

$\mathcal{L}_3$-regime - ETT: By counterfactual randomization Alice can sample from the $\mathcal{L}_3$ distribution $P(Y_x, X)$. She records her natural choice $X = x'$ on a particular evening (what she would have normally done) and randomizes the choice of X that she actually undertakes, during the explore phase. Using this distribution, she then performs the following action, for the natural $X = x'$ that she observes in the exploit phase:

$$do(X = \arg\max_x E[Y_x \mid x'])$$

We can compute this from Table (iii) of the SCM. Alice simply chooses to do the opposite of what she naturally feels like doing (corresponding to the the non-bold cells of the Table). This "ETT" $\mathcal{L}_3$

strategy yields an expected outcome of

$$\sum_{x'} P(x').\max_x E[Y_x \mid x'] = (0.5)[0.7 + 0.8] = 0.75,$$

outperforming both $\mathcal{L}_1$ and $\mathcal{L}_2$ strategies.

Of course, the explore-exploit phases are combined adaptively in a bandit algorithm like Thompson Sampling.

$\mathcal{L}_3$-regime - Optimal: Finally, Alice leverages her ability to perform *path-specific* randomization to sample from the distribution $P(Y_x, X, D_{x'})$. She then adapts a bandit algorithm to performs the following actions in the exploit phase:

$$\text{READ}(X) = x'$$

$$\text{CTF-WRITE}(x'' \to D), \text{ where } x'' = \arg\max_{x''} \left(\max_x E[Y_x \mid x', D_{x''}] \right); \text{READ}(D) = d$$

$$\text{CTF-WRITE}(x \to Y), \text{ where } x = \arg\max_x E[Y_x \mid x', d_{x''}]$$

,

where CTF-WRITE is simply the deterministic equivalent of CTF-RAND.

In words, during the explore phase, Alice gathers data on which arm x optimizes $\mathbb{E}[Y_x \mid x', d_{x''}]$, for all $x', , x'', d$. Then, during the exploit phase, Alice first observes $D_{x''} = d$ and $X = x'$, and then performs the action x which maximizes her outcome Y_x. Performing another optimization over the x'' gives her the best global optimum of $E[Y_x \mid x', d_{x''}]$.

Again, the explore-exploit phases are not separated in a bandit algorithm like Thompson Sampling, but incorporated adaptively.

Computing this from the SCM, suppose Alice chooses to record $D_{x_0} = 0 \oplus U_3 = U_3$.

- When $D_{x_0} = U_3 = 0$, Alice sees according to Table (i) of the SCM that the optimal strategy is to choose the opposite of what she naturally feels like doing (the values not in bold), giving the expected outcome $E[Y_x \mid x', D_{x_0} = 0]$, where $x \neq x'$, as $(0.5)[0.9+0.8] = 0.85$
- When $D_{x_0} = U_3 = 1$, Alice sees according to Table (ii) of the SCM that the optimal strategy is to go with her natural inclination (the values in bold), giving the expected outcome $E[Y_{x'} \mid x', D_{x_0} = 0] = (0.5)[0.8 + 0.7] = 0.75$
- Overall, since both values of D_{x_0} are equally likely, this strategy yields an expected outcome of $0.5[0.85 + 0.75] = 0.8$, which outperforms $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$-ETT strategies.

This walk-through can be repeated identically from the SCM had Alice chosen to measure Dx_1 instead.

E.3.1 SIMULATIONS

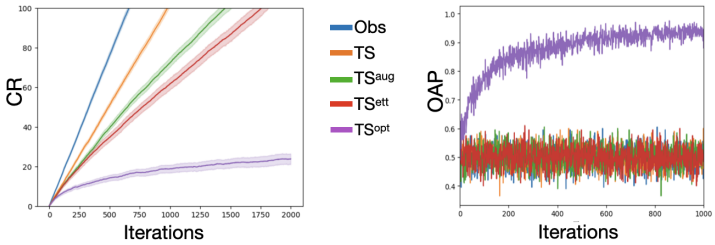


Figure 20: Example 3: Cumulative Regret (CR) and Optimal Arm Probability (OAP) for all strategies tested via Thompson Sampling.

The simulation compares the performance of four algorithms

- TS is the conventional Thompson Sampling algorithm that optimizes the $\mathcal{L}_2$ learning objective $\mathbb{E}[Y; do(x)]$;
- TS^{aug} is a contextual Thompson Sampling algorithm that treats $\{X = x', D_{x''} = d\}$ as merely some extra context variables in each round, and ignoring the $\mathcal{L}_3$ significance of these variables;
- TS^{ett} is given in Algorithm 5, implementing the ETT baseline strategy described earlier;
- TS^{opt} is given in Algorithm 4, implementing the $\mathcal{L}_3$-optimal strategy described earlier.

Importantly, TS^{opt} doesn't treat $\{X = x', D_{x''} = d\}$ merely as extra context variables. Rather, the counterfactual significance of these variables is leveraged via the consistency property

$$E[Y_x \mid x, d_x] = E[Y \mid x, d] \tag{41}$$

This means that for several arms being explored, the r.h.s allows us to hot-start the Thompson Sampling using offline ($\mathcal{L}_1$) data, as implemented in Line 18 of Algorithm 4. This allows for a dramatically faster convergence of the purple vs. green plot in Fig. 20.

Simulations were run for 2000 iterations, 200 epochs (Confidence Interval = 95%).

Algorithm 4 TS^{opt}: Thompson Sampling OPTIMAL (Bernoulli-Beta case)

Input: No. of timesteps, T; Observational data, $P(\mathbf{v})$
for $x'' \in$ Domain(X) **do**
  for $x' \in$ Domain(X) **do**
    $\alpha_D[x''][x'] \leftarrow 1$
    $\beta_D[x''][x'] \leftarrow 1$ {Initializing D-priors}
  end for
end for
for $i \in$ Domain(X) **do**
  for $j \in$ Domain(X) **do**
    for $d \in$ Domain(D) **do**
      for $k \in$ Domain(X) **do**
        $\alpha_Y[x_i][x_j][d][x_k] \leftarrow 1$
        $\beta_Y[x_i][x_j][d][x_k] \leftarrow 1$ {Initializing Y-priors}
      end for
    end for
  end for
end for
$t = 1$
while $t <= T$ **do**
  Perform READ(X) $= x'$, for unit
  for $j \in$ Domain(X) **do**
    $\mu_i^D \sim Beta(\alpha_D[x''][x_j], \beta_D[x''][x_j])$
  end for
  Perform CTF-WRITE($x' \to D$) for $x' = x_j; j := \arg\max_{j'} \mu_{j'}^D$
  Perform READ(D) $= d$, for unit {Get value of $D_{x''}$}
  for $k \in$ Domain(X) **do**
    if $x_k = x' = x''$ **then**
      $\mu_k^Y \leftarrow E[Y \mid x'', d]$ {Hot-start using obs. data}
    else
      $\mu_k^Y \sim Beta(\alpha_Y[x''][x'][d][x_k], \beta_Y[x''][x'][d][x_k])$
    end if
  end for
  Perform CTF-WRITE($x \to Y$) for $x = x_k; k := \arg\max_{k'} \mu_{k'}^Y$
  Perform READ(Y) $= y$, for unit {Get value of Y_x}
  $\alpha_D[x''][x'] \leftarrow \alpha_D[x''][x'] + y$
  $\beta_D[x''][x'] \leftarrow \beta_D[x''][x'] + 1 - y$ {Update D-priors}
  if $\neg(x = x' = x'')$ **then**
    $\alpha_Y[x''][x'][d][x] \leftarrow \alpha_Y[x''][x'][d][x] + y$
    $\beta_Y[x''][x'][d][x] \leftarrow \beta_Y[x''][x'][d][x] + 1 - y$ {Update Y-priors}
  end if
  $t \leftarrow t + 1$
end while

Algorithm 5 TS^{ett}: Thompson Sampling ETT (Bernoulli-Beta case)

Input: No. of timesteps, T; Observational data, $P(\mathbf{v})$
for $z \in$ Domain(Z) **do**
  for $x' \in$ Domain(X) **do**
    $\alpha[z][x'] \leftarrow 1$
    $\beta[z][x'] \leftarrow 1$ {Initializing priors}
  end for
end for
$t = 1$
while $t <= T$ **do**
  Perform $\textsc{Read}(Z) = z$, for unit
  Perform $\textsc{Read}(X) = x'$, for unit
  for $i \in$ Domain(X) **do**
    if $x_i = x'$ **then**
      $\mu_i \leftarrow E[Y \mid x', z]$ {Hot-start using obs. data}
    else
      $\mu_i \sim Beta(\alpha[z][x_i], \beta[z][x_i])$
    end if
  end for
  Perform $\textsc{ctf-Write}(x \rightarrow Y)$ where $x = x_i$ s.t. $i := \arg\max_{i'} \mu_{i'}$
  Perform $\textsc{Read}(Y) = y$, for unit {Get value of Y_x}
  if $x \neq x'$ **then**
    $\alpha[z][x'] \leftarrow \alpha[z][x'] + y$
    $\beta[z][x'] \leftarrow \beta[z][x'] + 1 - y$ {Update priors}
  end if
  $t \leftarrow t + 1$
end while

F OPTIMALITY RESULT AND PROOF

In this appendix, we identify a strategy that is provably optimal for decision-making in multi-arm bandit (MAB) problems. To focus the discussion, we define a generic *MAB template* (Fig. 21) that is generally representative of a broad class of bandit problems in the literature (the discussion can also be extended to other settings such as sequential or Markov decision processes in future work). X is the decision variable, Z is a context variable, Y is the reward, and D is a descendant of X confounded with Y.

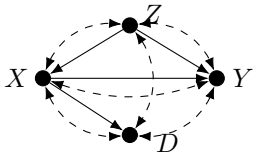


Figure 21: MAB template.

Definition F.1 (Decision strategy)**.** Given a decision problem following the MAB template (Fig. 21), a *decision strategy* π is a mapping from a set of variables $\mathbf{W}_\star$ (possibly counterfactual) to a set of actions $\mathcal{A}$ involving decision variable X. The expected reward of following this strategy is notated $\mu_\pi := \mathbb{E}[Y_\mathcal{A} \mid \mathbf{W}_\star]$, where $Y_\mathcal{A}$ is the potential response of Y under the actions $\mathcal{A}$.[5] ■

Example. The $\mathcal{L}_1$ strategy of simply observing the natural behavior of some behavioral agent is $\pi^{\text{obs}} : \{\} \mapsto \{\}$, which incurs the observational reward of $\mu_{\pi^{\text{obs}}} = \mathbb{E}[Y]$.

Example. As discussed in Sec. 4.2, the typical approach in the RL literature is the $\mathcal{L}_2$ strategy $\pi^{\text{int}} : \{z\} \mapsto \{\textsc{Write}(X : x^\star)\}$, where $x^\star := \arg\max_x \mathbb{E}[Y_x \mid z]$. In words, this strategy involves observing context $Z = z$ for a each round, and then performing the intervention $do(x)$ that maximizes the $\mathcal{L}_2$ quantity $\mathbb{E}[Y_x \mid z]$, also known as the *conditional average treatment effect*, or CATE.

Example. As discussed in Secs. 2, 4.2, a valid counterfactual strategy would be $\pi^{\text{ett}} : \{x', z\} \mapsto \{\textsc{ctf-Write}(x^\star) \to Y\}$, where $x^\star := \arg\max_x \mathbb{E}[Y_x \mid x', z]$.[6] In words, this strategy involves observing context $Z = z$ and the unit's natural inclination $X = x'$ for each each round, and then performing the intervention $do(x)$ that maximizes the $\mathcal{L}_3$ quantity $\mathbb{E}[Y_x \mid x', z]$, related to the *effect of the treatment on the treated*, or ETT.

In Example 3, we introduced a superior counterfactual strategy

$$\pi^{\text{opt}} : \{X, Z, D_{x''}\} \mapsto \{\textsc{ctf-Write}(x \to Y), \textsc{ctf-Write}(x'' \to D)\}, \tag{42}$$

where $x, x'' := \arg\max_{x,x''} \mathbb{E}[Y_x \mid Z, X, D_{x''}]$.

With minor abuse of notation, this is the strategy that (1) observes $Z = z, X = x'$ for a round; (2) maps from $\{z, x'\} \mapsto x''$, to perform the counterfactual intervention $\textsc{ctf-Write}(x'' \to D)$ to observe $D_{x''} = d$; and (3) maps from $\{z, x', d_{x''}\} \mapsto x$, to perform the counterfactual intervention $\textsc{ctf-Write}(x \to Y)$ that maximizes $\mathbb{E}[Y_x \mid z, x', d_{x''}]$.

For each mapping in (2), (3) yields a local optimum, in expectation over $X, Z, D_{x''}$. Optimizing over all choices of x'' in (2) yields a global optimum. Translating this to practice, we provide a general algorithm (Algorithm 6) that adapts any standard MAB solver to implement the optimal $\mathcal{L}_3$-strategy π^{opt}. We provide examples using Thompson Sampling in the Appendix E.3.1 (Algorithms 4,5).

The natural question is whether we can keep going higher up in Layer 3 of the PCH. Could we construct higher order strategies that map from $\{X, Z, D_{x''}, D_{x'''}\}$ by drawing samples from more refined counterfactuals? Sadly no, because a distribution like $P(Y_x, X, D_{x''}, D_{x'''})$ is not realizable (Def. 3.4), and the machinery we developed in Sec. 3 gives us the tools to reason about this.

Theorem F.2 (Optimality)**.** *Given a decision problem following the MAB template (Fig. 21), π^{opt} is an optimal realizable strategy. I.e., $\mu_{\pi^{opt}} \geq \mu_\pi, \forall \pi \in \Pi$, the space of realizable strategies.* ■

The significance of this result is that it averts the need to apply Thm. 3.5 and Cor. 3.7 to search intractably over the space of all possible $\mathcal{L}_3$-distributions for which ones are realizable. Of course, π^{opt} need not be *uniquely* optimal.

Corollary F.3 ($\mathcal{L}_3$-dominance)**.** *Given an MAB decision problem with causal diagram described by the MAB template (Fig. 21), the optimal $\mathcal{L}_3$-strategy π^{opt} dominates the $\mathcal{L}_1$-strategy π^{obs} and the optimal $\mathcal{L}_2$-strategy π^{int}. I.e., $\mu_{\pi^{opt}} \geq \mu_{\pi^{obs}}$ and $\mu_{\pi^{opt}} \geq \mu_{\pi^{int}}$.* ■

[5] We use *strategy* interchangeably with *policy* when the context is clear. However, it should be noted that a *policy* is usually defined w.r.t. a certain policy space, mapping from a fixed domain to actions on X. Here, we consider different domains to map from.

[6] $\textsc{ctf-Write}$ is simply the deterministic equivalent of $\textsc{ctf-Rand}$ (Def. 2.3).

For decades, the Fisherian RCT methodology used to enact π^{int} was deemed to be the "gold standard" for decision-making. We show that the $\mathcal{L}_3$-strategy π^{opt} is at least as good (and often better) than $\mathcal{L}_2$-strategies. This means that if MAB solvers UCB, EXP3 etc. were deployed to enact an $\mathcal{L}_2$-strategy in an environment where π^{opt} is better, the agent would incur linear cumulative regret, since the learning approach comes no closer to discovering the optimal strategy as the number of trials increases.

Algorithm 6 MAB-OPT

Input: MAB problem following Fig. 21; MAB solver (e.g. UCB, EXP3, TS); No. of rounds T; Obs. data $P(\mathbf{v})$
for each z, x' **do**
  Initialize D-arms x''
end for
for each z, x', x'', d **do**
  Initialize Y-arms x
  If $x = x' = x''$, hot-start using $P(\mathbf{v})$
end for
for $t \in [T]$ **do**
  Observe x', z
  Draw D-arm x'' using MAB solver
  Perform CTF-WRITE$(x'' \to D)$ and get $D_{x''} = d$
  Draw Y-arm x using MAB solver
  Perform CTF-WRITE$(x \to Y)$ and get $Y_x = y$
  Update D-arms and Y-arms according to MAB solver rules using y
end for

F.1 PROOFS FOR SECTION THEOREM F.2 AND COROLLARY F.3

Remark F.4. In order for an agent to enact a non-trivial decision strategy $\pi : \{\mathbf{W}_\star = \mathbf{w}\} \mapsto \mathcal{A}$, we observe that (1) the distribution $P(Y_\mathcal{A}, \mathbf{W}_\star)$ must be realizable (Def. 3.4); and (2) the agent must be able to observe $\mathbf{W}_\star$ before performing actions $\mathcal{A}$. We call this a *realizable decision strategy*, and notate the space of all realizable strategies in a MAB problem as Π. ■

Corollary F.3 : This result follows immediately from Theorem F.2, by simply recognizing that $\pi^{int}, \pi^{obs} \in \Pi$, the space of realizable strategies (the agent is presumed capable of performing the actions RAND(X), WRITE$(X : x)$).

Therefore, $\mu_{\pi^{ctf}}$ cannot be less than $\mu_{\pi^{int}}, \mu_{\pi^{obs}}$, by Theorem F.2. ■

Theorem F.2 :

From Lemma F.5, all strategies involve mappings, where each mapping maps to one of the following 5 possible action sets: (1) $\{\}$ (no action); (2) WRITE$(X : x)$, for some x; (3) only CTF-WRITE$(x \to Y)$ for some x; (4) only CTF-WRITE$(x'' \to D)$ for some x''; or (5) both CTF-WRITE$(x \to Y)$, CTF-WRITE$(x'' \to D)$ for some x, x''.

Define Π_5 to be the space of strategies where every mapping of each strategy in Π_5 is mapping to a pair of actions CTF-WRITE$(x \to Y)$, CTF-WRITE$(x'' \to D)$ for some x, x''. I.e., all mappings only involve possibility (5) under these strategies, for all any unit encountered in the decision problem.

Let π_5 be an optimal strategy in this space. I.e., $\pi_5 \in \arg\max_{\pi \in \Pi_5} \mu_\pi$.

By Lemma F.7, π_5 is also an optimal strategy in the space of all possible strategies. This means we only need to consider strategies whose mappings are mappings to a pair of CTF-WRITE procedures.

Let $\mathbf{W}_\star$ be the context used by π_5. If $\mathbf{W}_\star$ does not already contain the natural variables X, Z, we can always define π_5' that use context $\mathbf{W}_\star' = \mathbf{W}_\star \cup \{X, Z\}$ s.t. $\mu_{\pi_5'} = \mu_{\pi_5}$, where π_5' simply ignores the extra context variable in the mapping.

Such a move would not affect the realizability of π_5' because CTF-WRITE does not override the natural value of X, and both X, Z can be observed before decision-making.

Combinatorially, there are only 3 possibilities for picking each mapping in π_5'.

1. Mapping from $\{x', z\}$ to a pair of actions $\{\text{CTF-WRITE}(x \to Y), \text{CTF-WRITE}(x'' \to D)\}$
2. Mapping from $\{x', z\}$ to some $\text{CTF-WRITE}(x \to Y)$, observing $Y_x = y$, and mapping from $\{x', z, y_x\}$ to $\text{CTF-WRITE}(x'' \to D)$; or
3. Mapping from $\{x', z\}$ to some $\text{CTF-WRITE}(x'' \to D)$, observing $D_{x''} = d$, and mapping from $\{x', z, d_{x''}\}$ to $\text{CTF-WRITE}(x \to Y)$

We can use similar arguments to Lemma F.7, where we restricted our attention to the space of strategies Π_5 which could mimic all other optimal strategies.

Possibility 2 can be mimicked by some mapping following possibility 1 which maps to a joint pair of actions. The two are equivalent in terms of outcome, because conditioning on y_x to choose x'' does not affect the outcome Y. So we can restrict our attention to possibilities 1 and 2.

Each mapping of possibility 1 can be mimicked by possibility 3, where the extra step of conditioning on $d_{x''}$ just ignores the extra information about $d_{x''}$. Thus, we can replace all mappings in the optimal strategy π_5' with mappings of possibility 3, to get a strategy π_5'' that also performs optimally.

Since there are two mappings in π_5'', they must be the mappings which maximize the outcome.

This is precisely the definition of the strategy π^{opt} given in Equation 42 and in the description immediately following it.

■

Lemma F.5. *Any decision strategy π for a decision problem having causal structure same as the MAB template is s.t. each mapping of the strategy maps from domain of the context to one of the five following possible sets of actions: (1) $\{\}$ (no action); (2) $\text{WRITE}(X : x)$, for some x; (3) only $\text{CTF-WRITE}(x \to Y)$ for some x; (4) only $\text{CTF-WRITE}(x'' \to D)$ for some x''; or (5) both $\text{CTF-WRITE}(x \to Y)$, $\text{CTF-WRITE}(x'' \to D)$ for some x, x''.*

Proof. Since the physical action space only involves doing nothing, WRITE or CTF-WRITE. Any other combination would be equivalent to one of the 5 above. E.g., $\text{WRITE}(X : x)$ and $\text{CTF-WRITE}(x'' \to D)$ is the equivalent to the pair $\text{CTF-WRITE}(x \to Y)$ and $\text{CTF-WRITE}(x'' \to D)$ (see Remark D.5).

We ignore randomized actions for simplicity. From standard results in learning theory, there is an optimum to be found at a simplex corner so we need only search over the space of hard interventions. ■

Lemma F.6. *The context $\mathbf{W}_\star$ used in the strategy $\pi : \{\mathbf{W}_\star = \mathbf{w}\} \mapsto \mathcal{A}$ can only possibly contain a subset of $X, Z, D, D_{x''}$ for some x'', and at most one potential response of D.*

Proof. There are only 4 variables to consider: X, Y, Z, D.

By the definition of a realizable strategy (Remark F.4), we need $P(Y_{\mathcal{A}}, \mathbf{W}_\star)$ to be realizable. By Cor. 3.7 there cannot be two potential responses of the same variable in a realizable distribution. This rules out any other potential response of Y, and ensures only one each of X, Z, D.

Since the only possible actions are interventions involving X, which do not affect Z and X (natural variable), these are the only potential responses that could appear involving these variables.

Likewise, with D, D (natural value) and $D_{x''}$ are the only possible potential responses that could appear, and at most one of them can. ■

Lemma F.7. *If π_5 is an optimal strategy in Π_5, the set of all strategies which map to a pair of* CTF-WRITE *procedures, then π_5 is also an optimal strategy in the set of all strategies possible in the MAB decision problem.*

Proof. Let Π_1 be the space of all strategies possible in the problem. Note that $\Pi_5 \subseteq \Pi_1$. Let $\pi_1 \notin \Pi_5$ be an optimal strategy. I.e. $\pi_1 \in \arg\max_{\pi \in \Pi_1} \mu_\pi$. If no such π_1 the Lemma stands proved.

Let $\mathbf{W}_\star$ be the context used by π_1. If $\mathbf{W}_\star$ does not already contain the natural variable X, we can always define π_1' that uses context $\mathbf{W}_\star' = \mathbf{W}_\star \cup \{X\}$ s.t. $\mu_{\pi_1'} = \mu_{\pi_1}$, where π_1' simply ignores the extra context variable in the mapping. For now, it doesn't matter whether such π_1' is realizable or not. Just that it is also an optimal strategy.

Each mapping in the strategy π_1' maps from the domain of $\mathbf{W}_\star'$ to one of the five possible action sets mentioned in Lemma F.5. E.g., for some $\mathbf{W}_\star' = \mathbf{w}$, the strategy π_1' maps this to $\mathbf{w} \mapsto \{\}$ or $\mathbf{w} \mapsto$ WRITE(X).

Consider a mapping in π_1' from the domain of $\mathbf{W}_\star' = \{x', ...\}$ to possibility (1), empty set of actions (recall, the context includes natural X). Such a mapping can be mimicked by an equivalent mapping $\mathbf{W}_\star' = \{x', ...\} \mapsto \{\text{CTF-WRITE}(x' \to Y), \text{CTF-WRITE}(x' \to D)\}$. By the consistency property if $X(\mathbf{u}) = x'$, then $Y_{x'}(\mathbf{u}) = Y(\mathbf{u})$ and $D_{x'}(\mathbf{u}) = D(\mathbf{u})$.

Thus, we can replace all the mappings in π_1' that involve a mapping to the empty set of actions, with an equivalent pair of CTF-WRITE using the natural value of X observed in the context. Call this new strategy π_2. π_2 is as good as π_1' because the mappings are all equivalent. Thus, π_2 is also optimal in Π_1. Again, it doesn't matter that π_2 may not be realizable, just that it is optimal.

Next, consider a mapping in π_2 from the domain of $\mathbf{W}_\star' = \{x', ...\}$ to possibility (2), some action $\text{WRITE}(X : x)$. Such a mapping can be mimicked by an equivalent mapping $\mathbf{W}_\star' = \{x', ...\} \mapsto \{\text{CTF-WRITE}(x \to Y), \text{CTF-WRITE}(x \to D)\}$. The evaluation of $f_Y(x, Z, \mathbf{u})$ in both scenarios is the same, with the only difference being that f_X is overwritten, which doesn't affect the outcome Y for each $\mathbf{u}$. I.e., the outcome Y would be the same for every unit under both strategies.

Thus, we can replace all the mappings in π_2 that involve a mapping to some action $\text{WRITE}(X : x)$, with an equivalent pair of CTF-WRITE. Call this new strategy π_3. π_3 is as good as π_2 because the mappings are all equivalent in terms of outcome. Thus, π_3 is also optimal in Π_1.

Next, consider a mapping in π_3 from the domain of $\mathbf{W}_\star' = \{x', ...\}$ to possibility (3), some action $\text{CTF-WRITE}(x \to Y)$. Such a mapping can be mimicked by an equivalent mapping $\mathbf{W}_\star' = \{x', ...\} \mapsto \{\text{CTF-WRITE}(x \to Y), \text{CTF-WRITE}(x' \to D)\}$ for natural value x'. By the consistency property, if $X(\mathbf{u}) = x'$ then $D_{x'}(\mathbf{u}) = D(\mathbf{u})$.

Thus, we can replace all the mappings in π_3 that involve a mapping to some action $\text{CTF-WRITE}(x \to Y)$, with an equivalent pair of CTF-WRITE. Call this new strategy π_4. π_4 is as good as π_3 because the mappings are all equivalent in terms of outcome. Thus, π_4 is also optimal in Π_1.

Next, consider a mapping in π_4 from the domain of $\mathbf{W}_\star' = \{x', ...\}$ to possibility (4), some action $\text{CTF-WRITE}(x \to D)$. Such a mapping can be mimicked by an equivalent mapping $\mathbf{W}_\star' = \{x', ...\} \mapsto \{\text{CTF-WRITE}(x' \to Y), \text{CTF-WRITE}(x \to D)\}$ for natural value x'. By the consistency property, if $X(\mathbf{u}) = x'$ then $Y_{x'}(\mathbf{u}) = Y(\mathbf{u})$.

Thus, we can replace all the mappings in π_4 that involve a mapping to some action $\text{CTF-WRITE}(x \to D)$, with an equivalent pair of CTF-WRITE. Call this new strategy π_5'. π_5' is as good as π_4 because the mappings are all equivalent in terms of outcome. Thus, π_5' is also optimal in Π_1.

However, note that the only possible mappings in π_5' are possibility (5) involving a pair of CTF-WRITE actions. Which means $\pi_5' \in \Pi$.

Thus, we show that all optimal strategies in Π_5 are also optimal in the overall space of strategies. ■