The difference between structural counterfactuals and potential outcomes

Lucas de Lara

The difference between structural counterfactuals and potential outcomes

Lucas De Lara*

Institut de Mathématiques de Toulouse, Université Paul Sabatier

Abstract

Most of the literature on causality considers the structural framework of Pearl and the potential-outcome framework of Neyman and Rubin to be formally equivalent, and therefore interchangeably uses the do-notation and the potential-outcome subscript notation to write counterfactual outcomes. In this paper, we superimpose the two causal model to prove that structural counterfactual outcomes and potential outcomes do not coincide in general—not even in law. More precisely, we express the law of the potential outcomes in terms of the latent structural causal model under the fundamental assumptions of causal inference. This enables us to precisely identify when counterfactual inference is or is not equivalent between approaches, and to clarify the meaning of each kind of counterfactuals.

Keywords: causality, structural causal models, potential outcomes

1 Introduction

Understanding causation between phenomena rather than mere association is a fundamental scientific challenge. Over the last decades, two mathematical frameworks using a terminology based on random variables have become the gold standards to address this problem.

On the one hand, the notorious *structural account* of Pearl [2009] rests on the knowledge of a *structural causal model* (SCM) which specifies all cause-effect equations between observed random variables (often depicted by a graph). The interest of such equations comes from the possibility of carrying out *do-interventions*: forcing a variable to take a given value while keeping the rest of the mechanism untouched. More concretely, let T and Y be observed variables of the model such that we would like to understand the downstream effect of T onto Y. Replacing the formula generating T by $T = t$ for a given possible value $t \in \mathcal{T}$ and propagating this change through the other equations defines the altered variable $Y_{T=t}$, representing Y *had* T *been equal to* t.

On the other hand, the widely-used *potential-outcome account* of Rubin [1974] mathematically formalizes causal inference in clinical trials. Letting T denote a *treatment status* (e.g., taking a drug or not) and Y an *outcome* of interest (e.g., recovering or not), this framework postulates the existence of *potential outcomes* $(Y_t)_{t\in\mathcal{T}}$ representing what the outcome would be *were* T *equal to* t for any $t \in \mathcal{T}$. The *fundamental problem of causal inference* [Holland, 1986] refers to the fact that in practice we cannot observe simultaneously all the potential outcomes, rendering unidentifiable the causal effect of T onto Y. Nevertheless, causal inference can

*E-mail: lucas.de_lara@math.univ-toulouse.fr

still be achieved thanks to a mix of untestable assumptions and statistical tools: adjusting on a set of available covariates X containing all possible *confounders* between the treatment and the potential outcomes notoriously permits to identify the law of counterfactual outcomes.

Each of these causal theories enables to carry out *counterfactual reasoning*, that is answering contrary-to-fact questions such as "Had they taken the drug, would have they recovered?": by applying do-interventions on an SCM, one can compute the outcomes $(Y_{T=t})_{t\in\mathcal{T}}$ for all possible treatment statuses; using the Neyman-Rubin causal model, one can infer the law of the potential outcomes $(Y_t)_{t\in\mathcal{T}}$. Both approaches involve variables describing counterfactual outcomes, more precisely outcomes *had the variable T taken a certain value*. This naturally raises the question: are these outcome variables equal (almost surely or in law) across frameworks? We believe the literature on causal inference to be strongly misleading on this matter. A plethora of scientific books and survey papers interchangeably use Pearl's do-notation and the potential-outcome subscript notation to write outcomes after interventions, suggesting that the corresponding definitions of counterfactuals are identical and differ only from theirs perspectives [Barocas et al., 2019, Colnet et al., 2020, Imbens, 2020, Makhlouf et al., 2020, Neal, 2020]. To justify this, they often refer to Pearl, who argued that "the two frameworks can be used interchangeably and symbiotically".[1] However, to our knowledge, works on equivalences between the two causal frameworks focus on translating conditional-independence restrictions into graphical assumptions instead of actually proving whether counterfactual outcomes are *algebraically* equal across models, or implicitly address specific cases. Notably, both [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013]—acclaimed references unifying both causal frameworks—consider *ex nihilo* the algebraic equivalence between the two notations.

In this paper, we prove that structural counterfactual outcomes and potential outcomes do not coincide in general—not even in law—and discuss the implications of this result. More specifically, the rest of the article is organized as follows. In Section 2, we introduce the basic knowledge on structural causal models and the potential-outcome framework. In Section 3, we superimpose the two causal frameworks to derive a mathematical analysis of their similarities and differences. As our main result, we express under classical assumptions the law of the potential outcomes $(Y_t)_{t\in\mathcal{T}}$ using the causal equations of the SCM. This critically implies that—in contrary to what the mainstream literature often suggests—the two definitions of counterfactual outcomes are not always equal in law. In Section 4, we precisely analyze when counterfactual inference in one framework does (not) yield the same conclusions as in the other, and clarify how such results relate to the formal equivalence between causal frameworks accepted by the causal-inference community. This contribution notably highlights that $(Y_t)_{t\in\mathcal{T}}$ represents *ceteris paribus* counterfactuals under the standard causal-inference setting while $(Y_{t=t})_{t\in\mathcal{T}}$ always represents *mutatis mutandis* counterfactuals, which has concrete consequences on the computation of causal effects. In doing this work, we aim at clarifying the role of each framework in the past, current, and future causal-inference research.

2 Preliminaries

This section provides the necessary background on structural causal models and potential outcomes. It is meant to keep the paper self-contained and can be skipped by a reader familiar with these frameworks.

Let us fix some notations before proceeding. Throughout, we consider a probability space $(\Omega, \Sigma, \mathbb{P})$ with Ω a sample space, Σ a σ-algebra, and $\mathbb{P} : \Sigma \to [0, 1]$ a probability measure. We write $\mathcal{L}(W)$ and $\mathbb{E}[W]$ for respectively the law and expectation under $\mathbb{P}$ of a random variable or random vector W. Two variables W_1 and W_2 are $\mathbb{P}$-*almost surely equal*, denoted by $W_1 \overset{\mathbb{P}-a.s.}{=} W_2$, if $\mathbb{P}(W_1 = W_2) = 1$; they are *equal in law* (under

[1] http://causality.cs.ucla.edu/blog/index.php/2012/12/03/judea-pearl-on-potential-outcomes/

$\mathbb{P}$), denoted by $W_1 \overset{\mathcal{L}}{=} W_2$, if $\mathcal{L}(W_1) = \mathcal{L}(W_2)$. Additionally, $W_1 \perp\!\!\!\perp W_2$ means that W_1 and W_2 are independent (under $\mathbb{P}$). Besides, for any tuple $w := (w_i)_{i\in\mathcal{I}}$ indexed by a finite index set $\mathcal{I}$ and any subset $I \subseteq \mathcal{I}$ we write $w_I := (w_i)_{i\in I}$. Similarly, we define $\mathcal{W}_I := \prod_{i\in I} \mathcal{W}_i$ for any collection of spaces $(\mathcal{W}_i)_{i\in\mathcal{I}}$.

2.1 Pearl's causal framework

Pearl's causal modeling mathematically formalizes associations that standard probability calculus cannot describe through the notions of structural causal models and do-interventions [Pearl, 2009]. This section recalls the basics on this topic, borrowing the introduction proposed by Bongers et al. [2021].

2.1.1 Structural causal models

A *structural causal model* (SCM) represents the causal relationships between the studied variables. It is the cornerstone of Pearl's causal framework.

Definition 1 (Structural causal model)**.** *Let $\mathcal{I}$ and $\mathcal{J}$ be two disjoint finite index sets, and write $\mathcal{V} := \prod_{i\in\mathcal{I}} \mathcal{V}_i \subseteq \mathbb{R}^{|\mathcal{I}|}$, $\mathcal{U} := \prod_{i\in\mathcal{J}} \mathcal{U}_i \subseteq \mathbb{R}^{|\mathcal{J}|}$ for two measurable product spaces. A* structural causal model *$\mathcal{M}$ is a couple $\langle U, G\rangle$ where:*

1. *$U : \Omega \to \mathcal{U}$ is a vector of mutually independent random variables, sometimes called the* random noise*;*
2. *$G = \{G_i\}_{i\in\mathcal{I}}$ is a collection of measurable $\mathbb{R}$-valued functions, where for every $i \in \mathcal{I}$ there exist two subsets of indices $\mathrm{Endo}(i) \subseteq \mathcal{I}$ and $\mathrm{Exo}(i) \subseteq \mathcal{J}$, respectively called the* endogenous *and* exogenous parents *of i, such that G_i is from $\mathcal{V}_{\mathrm{Endo}(i)} \times \mathcal{U}_{\mathrm{Exo}(i)}$ to $\mathcal{V}_i$.*[2]

A random vector $V : \Omega \to \mathcal{V}$ is a solution of $\mathcal{M}$ if for every $i \in \mathcal{I}$,

$$V_i \overset{\mathbb{P}-a.s.}{=} G_i(V_{Endo(i)}, U_{Exo(i)}). \tag{1}$$

The collection of equations defined by (1) *and characterized by G and U are called the* structural equations*.*

Such a model explains how some *endogenous* variables V, representing observed data, are generated from *exogenous* variables U, describing background factors. The structural equations quantify the causal dependencies between all these variables and are frequently illustrated by the directed graph with nodes $\mathcal{I} \cup \mathcal{J}$, and such that a directed edge points from node k to node l if and only if $k \in \mathrm{Endo}(l) \cup \mathrm{Exo}(l)$ (we say in this case that k is a parent of l). For convenience, we make the common assumption that the studied models are *acyclic*, which means that their associated graphs do not contain any cycles.

Assumption 2 (Acyclicity)**.** *The structural causal model $\mathcal{M}$ induces a directed* acyclic *graph.*

Not only acyclicity simplifies the interpretation of causal dependencies, but it entails *unique solvability* of the SCM: according to [Bongers et al., 2021, Proposition 3.4], Equation (1) admits a unique solution up to $\mathbb{P}$-negligible sets. We will abusively refer to such a solution as *the* solution of the SCM.

The purpose of causal structures is to capture the assumption that features are not independently manipulable. As we detail next, they enable to understand the downstream effect of fixing some variables to certain values onto nonintervened variables.

[2] This definition tolerates that distinct endogenous variables share the same exogenous parents, that is $\mathrm{Exo}(i) \cap \mathrm{Exo}(i') \neq \emptyset$ for some $i \neq i'$. Therefore, the $(U_{\mathrm{Exo}(i)})_{i\in\mathcal{I}}$ are not necessarily mutually independent.

2.1.2 Do-intervention

A *do-intervention* is an operation forcing a set of endogenous variables to take predefined values while keeping all the rest of the causal mechanism equal.

Definition 3 (Do-intervention)**.** *Let $\mathcal{M} = \langle U, G\rangle$ be an SCM, $I \subseteq \mathcal{I}$ a subset of endogenous variables, and $v_I \in \mathcal{V}_I$ a value. The action* $\mathrm{do}(I, v_I)$ *defines the modified model $\mathcal{M}_{\mathrm{do}(I,v_I)} = \langle U, \tilde{G}\rangle$ where $\tilde{G}$ is given by*

$$\tilde{G}_i := \begin{cases} v_i \text{ if } i \in I, \\ G_i \text{ if } i \in \mathcal{I} \setminus I. \end{cases}$$

Do-interventions preserve acyclicity, and therefore unique solvability. As a consequence, if V is the solution of an acyclic $\mathcal{M}$, one can define (up to $\mathbb{P}$-negligible sets) its post-intervention counterpart $V_{\mathrm{do}(I,v_I)}$ solution to $\mathcal{M}_{\mathrm{do}(I,v_I)}$. It describes an alternative world where every V_i for $i \in I$ is set to value v_i. In the sequel, we simply write $\mathrm{do}(V_I = v_I)$ for the operation $\mathrm{do}(I, v_I)$, and use the subscript $V_I = v_I$ to indicate results of this operation. Crucially, intervening does not amount to conditioning in general, that is $\mathcal{L}(V \mid V_I = v_I) \neq \mathcal{L}(V_{V_I = v_I})$.

The next proposition provides a general expression of the solution before and after intervention, and will play a key role throughout this paper.

Proposition 4 (Do-calculus on variables)**.** *Let $\mathcal{M} = \langle U, G\rangle$ be an SCM satisfying Assumption 2 with solution V, and consider a partition $\{I, J\}$ of $\mathcal{I}$. There exists a deterministic measurable function F_J such that*

$$V_J \overset{\mathbb{P}-a.s.}{=} F_J(V_{\mathrm{Endo}(J)\setminus J}, U_{\mathrm{Exo}(J)}).$$

Moreover, for any intervention $\mathrm{do}(V_I = v_I)$ *the solution $\tilde{V}$ of $\mathcal{M}_{V_I = v_I}$ verifies*

$$\tilde{V_J} \overset{\mathbb{P}-a.s.}{=} F_J(v_{\mathrm{Endo}(J)\setminus J}, U_{\mathrm{Exo}(J)}),$$
$$\tilde{V_I} \overset{\mathbb{P}-a.s.}{=} v_I.$$

Importantly, this is the same deterministic function F_J that generates V_J and its post-intervention counterpart $\tilde{V_J}$, the only change being the assignment $V_I = v_I$. Slightly abusing notations, we will sometimes artificially extend the input variables of F_J to write $V_J \overset{\mathbb{P}-a.s.}{=} F_J(V_I, U_{\mathrm{Exo}(J)})$ and $\tilde{V_J} \overset{\mathbb{P}-a.s.}{=} F_J(v_I, U_{\mathrm{Exo}(J)})$.

2.1.3 Counterfactual inference

Do-calculus provides a natural framework to address counterfactual queries. Let for instance $V := (T, X, Y)$ be the solution to an acyclical SCM $\mathcal{M} := \langle U, G\rangle$. We aim at answering the counterfactual question: *had T been equal to t, what would have been the value of Y for a unit factually described by $X = x$?* Pearl [2009] answers this question using the so-called *three-step procedure*:

1. **Abduction:** Deduce the posterior distribution of U given the reference $\{X = x\}$;
2. **Action:** Carry out do-calculus on $\mathcal{M}$ to obtain the intervened causal mechanism $G_{T=t}$ of $\mathcal{M}_{T=t}$;
3. **Prediction:** Pass the posterior distribution $\mathcal{L}(U \mid X = x)$ through $G_{T=t}$ to generate in particular the distribution $\mathcal{L}(Y_{T=t} \mid X = x)$ of counterfactual outcomes.

More generally, an SCM enables one to compute counterfactual distributions for any choices of events of reference, variables to alter by do-intervention, and outcomes of interest.

2.2 Neyman-Rubin causal framework

The potential-outcome framework, also known as *Neyman-Rubin causal modeling* [Rubin, 1974], was designed to understand the causal effect of a treatment onto an outcome of interest, for instance when one aims at assessing the contribution of a drug to recovering from some disease in clinical trials. In this section, we introduce this widely-used framework in the specific case of a binary treatment.

2.2.1 Potential outcomes

Let $T : \Omega \to \{0, 1\}$ represent a binary *treatment status*, typically such that $T = 0$ indicates the absence of treatment and $T = 1$ indicates a treatment. More generally, it can encode any distinction between some groups (e.g., men and women). Assuming *no interference between units*, this framework postulates two *potential outcomes* $Y_0 : \Omega \to \mathbb{R}$ and $Y_1 : \Omega \to \mathbb{R}$, one for each treatment status. These potential outcomes as well as the treatment may depend on some covariates $X : \Omega \to \mathbb{R}^d$ (such as the patient's weight, height, or historical data in clinical trials). Critically, we cannot observe simultaneously Y_0 and Y_1 for a single unit ω: a problem referred as the *fundamental problem of causal inference* [Holland, 1986]. We only have access to the realized *outcome variable* $Y : \Omega \to \mathbb{R}$ which is supposed to be *consistent* with (Y_0, Y_1), that is satisfying $Y = (1 - T) \cdot Y_0 + T \cdot Y_1$. Concretely, if $T(\omega) = 1$ for some $\omega \in \Omega$, then $Y(\omega) = Y_1(\omega)$, and $Y_0(\omega)$ becomes unidentifiable by mere observations. In this case, $Y_1(\omega)$ is called the *factual* outcome while $Y_0(\omega)$ is called the *counterfactual* outcome.

Understanding the causal relationship between the treatment and the outcome in this framework consists in answering counterfactual questions such as "What would have been the value of Y had T been equal to 1 instead of 0 (for a unit described by $X = x$)?". In practice, this amounts to estimating the discrepancy between $\mathcal{L}(Y_1)$ and $\mathcal{L}(Y_0)$, or $\mathcal{L}(Y_1 \mid X = x)$ and $\mathcal{L}(Y_0 \mid X = x)$. People commonly focus on computing the *average treatment effect* $\mathbb{E}[Y_1 - Y_0]$ or the *conditional average treatment effect* $\mathbb{E}[Y_1 - Y_0 \mid X = x]$. The main challenge lies in the fact that *association is not causation* in general. In particular, the observable quantity $\mathbb{E}[Y \mid T = t]$ does not necessarily coincide with the unobservable quantity $\mathbb{E}[Y_t]$ for $t \in \{0, 1\}$. Typically, if some medical treatment is more likely to be taken by weaker patients, we may observe a lower rate of recovery among the treated group compared to the nontreated group due to the health condition even though the medicine does increase recovery all other things being kept equal: we would observe $\mathbb{E}[Y \mid T = 1] < \mathbb{E}[Y \mid T = 0]$ while $\mathbb{E}[Y_1] > \mathbb{E}[Y_0]$ (a phenomenon referred as *Simpson's paradox* [Blyth, 1972]). In this case, the health condition is called a *confounder*: a variable associated with both the distribution of the treatment and the outcome. However, causal inference from observational data is still possible, as explained next.

2.2.2 Estimation of causal effects

A treatment effect is *identifiable* if it can be expressed with observational quantities only, that is in terms of X, T and Y. Identifiability requires two fundamental assumptions. The first one goes by many names through the literature: *conditional ignorability*, *conditional exchangeability*, *conditional exogeneity*, and *conditional unconfoundedness* (among others). Originally formulated by Rosenbaum and Rubin [1983], it states that the potential outcomes are independent of the treatment conditional to the covariates, that is $(Y_0, Y_1) \perp\!\!\!\perp T \mid X$. Said differently, it prevents the existence of unmodeled confounders between the treatment and the potential outcomes. Note that this assumption is untestable, as it would require to observe simultaneously the two potential outcomes. The second key hypothesis is *positivity*, which ensures that all individuals can be exposed to both treatment statuses, that is $0 < \mathbb{P}(T = 1 \mid X) < 1$. It readily follows from positivity and conditional ignorability that $\mathcal{L}(Y \mid X, T = t)$ is well defined and coincides with $\mathcal{L}(Y_t \mid X)$ for $t \in \{0, 1\}$, meaning

that observable outcomes have a causal interpretation. Several statistical methods coexist to estimate the (conditional) average causal effect, all building upon this implication (see for instance [Imbens, 2004, Yao et al., 2021]). We do not detail them for concision and clarity since it is not the topic of this paper. We only point out that, similarly to structural causal models, the potential-outcome framework enables one to carry out counterfactual inference.

3 Main result

In this section, we show that a Neyman-Rubin causal model and a structural causal model compatible with a same distribution of observations produce counterfactual outcomes that are generally not almost-surely equal, nor equal in law.

3.1 Setting

Let $N, d, p \geq 1$ be integers, and define three random variables $T : \Omega \to \mathcal{T} := \{0, 1, \ldots, N\}$, $X : \Omega \to \mathbb{R}^d$, and $Y : \Omega \to \mathbb{R}^p$. In order to study the consistency of counterfactual statements between the Neyman-Rubin causal framework and Pearl's causal framework, we consider a superimposed construction where the observations described by (T, X, Y) are concurrently governed by a potential-outcome model and a structural causal model.

On the one hand, we assume that Y is the outcome of interest, T the treatment status, and X some covariates in a potential-outcome framework. This amounts to postulating N random vectors $(Y_t)_{t\in\mathcal{T}}$ satisfying the *consistency rule*:

$$Y \overset{\mathbb{P}-a.s.}{=} \sum_{t\in\mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t.$$

Note that we address a more general framework than in Section 2, considering a nonbinary treatment and a multivariate outcome. In this setting, the two fundamental assumptions for causal inference can be written as follows.

Assumption 5 (Positivity)**.**

$$0 < \mathbb{P}(T = t \mid X) < 1, \ \textit{for all } t \in \mathcal{T}.$$

Assumption 6 (Conditional ignorability)**.**

$$(Y_t)_{t\in\mathcal{T}} \perp\!\!\!\perp T \mid X.$$

On the other hand, we assume that these variables are generated by a latent, unknown structural causal model: the random vector $V := (T, X, Y)$ is the solution to an acyclical SCM $\mathcal{M} = \langle U, G\rangle$ where U_T, U_X and U_Y denote the exogenous parents of respectively T, X, and Y. Moreover, we suppose that $\mathcal{M}$ satisfies:

Assumption 7 (Outcome)**.**

$$Y_{\mathrm{Endo}(T)} = Y_{\mathrm{Endo}(X)} = \emptyset \ \textit{and}\ U_Y \perp\!\!\!\perp (U_T, U_X).$$

The first item of Assumption 7 is a graphical condition that formally defines the variable Y as the *outcome*; it changes in response to X and T but not the contrary. Through Proposition 4, it permits to write

$$\begin{aligned} T &\overset{\mathbb{P}-a.s.}{=} F_T(X, U_T), \\ X &\overset{\mathbb{P}-a.s.}{=} F_X(T, U_X), \\ Y &\overset{\mathbb{P}-a.s.}{=} F_Y(T, X, U_Y), \end{aligned}$$

where F_X, F_T and F_Y are deterministic measurable functions derived from G. The artificial cycle in these formulas (i.e., X and T are both functions of each other) merely serves to consider all configurations of causal links between T and X (see Figure 1); strictly, $\mathcal{M}$ satisfies Assumption 2. The following lemma clarifies the role of second item in Assumption 7.

Lemma 8 (Random noise)**.** *Let (T, X, Y) be the solution of an SCM $\mathcal{M}$ satisfying Assumptions 2 and 7 where U_T, U_X and U_Y denote the exogenous parents of respectively T, X and Y. Then $U_Y \perp\!\!\!\perp (T, X)$.*

It guarantees that all potential confounders between T and Y—except T itself—are included in X. All in all, Assumption 7 simply means through Lemma 8 that the randomness of the outcome $Y \overset{\mathbb{P}-a.s.}{=} F_Y(T, X, U_Y)$ can be divided into three sources: the direct effect of the treatment status T, the direct effect of the covariates X, and any other possible effects U_Y independent to T and X.

To conclude this setup, recall that Proposition 4 also enables to define for every $t \in \mathcal{T}$ the post-intervention outcome under $\mathrm{do}(T = t)$ as

$$Y_{T=t} \overset{\mathbb{P}-a.s.}{=} F_Y(t, X_{T=t}, U_Y),$$

where the altered covariates are $X_{T=t} \overset{\mathbb{P}-a.s.}{=} F_X(t, U_X)$. We have now set the stage to present our main results.

3.2 Three levels of difference

We argue that counterfactual outcomes are not equivalent between the two models at three levels: at the definition level, at the variable level, at the distributional level.

3.2.1 The definition level

Before addressing mathematical equalities, we would like to underline a more conceptual distinction: the two considered causal models differ fundamentally in their constructions of counterfactual outcomes. As noted by Pearl [2010], the potential outcomes $(Y_t)_{t\in\mathcal{T}}$ are "undefined *primitives*" of the Neyman-Rubin causal model, not related to any formal of measurable quantities, while the intervened outcomes $(Y_{T=t})_{t\in\mathcal{T}}$ are "*derivatives*" of the structural causal model by application of do-calculus. Said differently, the firsts are inputs *defining* the causal model, whereas the seconds are post-intervention outputs *defined by* the causal model.

However, Pearl uses the same notation for both constructions, suggesting a mathematical equality. Are they truly equivalent in the sense that $Y_t \overset{\mathbb{P}-a.s.}{=} Y_{T=t}$, or at least $Y_t \overset{\mathcal{L}}{=} Y_{T=t}$? This is what we address next.

3.2.2 The variable level

The aforementioned input/output difference is in fact more than just conceptual. Because an input can be arbitrarily chosen whereas an output is a necessary consequence, it feels that we could easily find settings where they are not equal. Of course, potential outcomes are not completely arbitrary: they must follow the consistency rule, that is $Y \overset{\mathbb{P}-a.s.}{=} \sum_{t\in\mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t$. But this property does not fully characterize the outcomes in the sense that there is no unique choice of $(Y_t)_{t\in\mathcal{T}}$ satisfying the consistency rule. More precisely, while necessarily $Y_t = Y$ on $\{T = t\}$ for $t \in \mathcal{T}$, there is no restriction on Y_t over $\Omega \setminus \{T = t\}$; it could take any value over it without violating the consistency rule. Consequently, it is mathematically impossible to relate Y_t—well-identifiable on the event $\{T = t\}$ only—to $Y_{T=t}$—defined (almost) everywhere through the structural causal model $\mathcal{M}$. Without further assumptions, we only have identification of the *observed outcomes*, namely $Y_t = Y_{T=t} = Y$ on $\{T = t\}$, as a direct consequence of the proposition below.

Proposition 9 (Consistency rule for structural counterfactuals)**.** *Let (T, X, Y) be the solution of an SCM $\mathcal{M}$ satisfying Assumptions 2.*[3] *Then, $(Y_{T=t})_{t\in\mathcal{T}}$ verifies the consistency rule,*

$$Y \overset{\mathbb{P}-a.s.}{=} \sum_{t\in\mathcal{T}} \mathbf{1}_{\{T=t\}} Y_{T=t}.$$

As such structural counterfactuals can be defined as potential outcomes, since the only requirement to be admissible potential outcomes is to follow the consistency rule. However, this does not signify that any pair or each causal model compatible with a same dataset produces potential outcomes and structural counterfactuals that coincide. In the next examples, we exhibit structural causal models and potential outcomes such that counterfactual outcomes are not almost-surely equal between frameworks.

Example 1 (Nonequality almost-surely)**.** *Consider any acyclic structural causal model with solution (T, X, Y) such that Y is unidimensional and $0 < \mathbb{P}(T = t)$ for all $t \in \mathcal{T}$. We can construct via do-calculus the structural counterfactual outcomes $(Y_{T=t})_{t\in\mathcal{T}}$. Now, define the potential outcomes $(Y_t)_{t\in\mathcal{T}}$ as follows. For any $t \in \mathcal{T}$,*

$$Y_t = \mathbf{1}_{\{T=t\}} Y_{T=t} + \mathbf{1}_{\{T\neq t\}} (Y_{T=t} + 1).$$

The tuple $(Y_t)_{t\in\mathcal{T}}$ satisfies the consistency rule according to Proposition 9, but is clearly not almost-surely equal to $(Y_{T=t})_{t\in\mathcal{T}}$.

Note that this example focuses on the most general setting, where the potential outcomes verify only consistency. The next example proves that counterfactual outcomes are not necessarily equal when the fundamental assumptions of causal inference hold.

Example 2 (Nonequality almost-surely under causal-inference assumptions)**.** *Consider the following structural causal model:*

$$\begin{aligned} T &\overset{\mathbb{P}-a.s.}{=} U_T, \\ X &\overset{\mathbb{P}-a.s.}{=} U_X, \\ Y &\overset{\mathbb{P}-a.s.}{=} T + U_Y, \end{aligned}$$

where U_T follows a Bernoulli distribution with parameter $1/2$, U_X is any random variable, and U_Y follows a centered Gaussian distribution with unit variance, such that U_T, U_X and U_Y are mutually independent. According to the rules of do-calculus, there exists a set Ω^* satisfying $\mathbb{P}(\Omega^*) = 1$ such that for any $\omega \in \Omega^*$, $Y_{T=0}(\omega) = U_Y(\omega)$ and $Y_{T=1}(\omega) = 1 + U_Y(\omega)$.

Next, set $U'_Y := -U_Y$ and define the potential outcomes as follows:

$$\begin{aligned} Y_0 &:= (1-T)\cdot U_Y + T \cdot U'_Y = (1-T)\cdot U_Y - T \cdot U_Y, \\ Y_1 &:= (1-T)\cdot (1+U'_Y) + T \cdot (1+U_Y) = (1-T)\cdot (1-U_Y) + T \cdot (1+U_Y). \end{aligned}$$

They clearly satisfy the consistency rule. Let us show that Assumptions 5 and 6 hold. Firstly, since $T \perp\!\!\!\perp X$, we have $\mathbb{P}(T = 1 \mid X) = \mathbb{P}(T = 1) = 1/2$ which entails positivity. Secondly, $\mathcal{L}((Y_0, Y_1) \mid X = x, T = 1) = \mathcal{L}((Y_0, Y_1) \mid T = 1)$ because $U_Y \perp\!\!\!\perp (T, X)$. Additionally, $\mathcal{L}((Y_0, Y_1) \mid T = 1) = \mathcal{L}((-U_Y, 1 + U_Y)) = \mathcal{L}((U_Y, 1 - U_Y))$ since $U_Y \overset{\mathcal{L}}{=} -U_Y$, and $\mathcal{L}((U_Y, 1 - U_Y)) = \mathcal{L}((Y_0, Y_1) \mid T = 0) = \mathcal{L}((Y_0, Y_1) \mid X = x, T = 0)$ since $U_Y \perp\!\!\!\perp (T, X)$. Therefore, $\mathcal{L}((Y_0, Y_1) \mid X = x, T = 1) = \mathcal{L}((Y_0, Y_1) \mid X = x, T = 0)$, meaning that conditional ignorability holds.

[3] We point out that Assumption 7 is not required.

We now turn to proving that (Y_0, Y_1) is not almost-surely equal to $(Y_{T=0}, Y_{T=1})$. On the event $\Omega^* \cap \{T = 0\}$, we have $Y_{T=1} = 1 + U_Y$ while $Y_1 = 1 + U'_Y = 1 - U_Y$. From $\mathbb{P}(U_Y = 0) = 0$ and $\mathbb{P}(\Omega^* \cap \{T = 0\}) = 1/2$ it follows that $\mathbb{P}(Y_{T=1} \neq Y_1) > 0$, meaning that the two counterfactual outcomes are not almost surely equal. We can prove a similar result for $Y_{T=0}$ and Y_0.

Remark also that $Y_t \overset{\mathcal{L}}{=} Y_{T=t}$ for every $t \in \{0, 1\}$, since $U_Y \overset{\mathcal{L}}{=} U'_Y$. In this scenario, counterfactual outcomes are not almost surely equal but equal in law.

Nevertheless, one could argue that this level of difference is not significant, since people doing counterfactual inference do not work with the variables themselves but their laws. They typically aim at computing quantities such as $\mathbb{E}[Y_1 - Y_0]$ or $\mathcal{L}(Y_{T=1} \mid X = x, T = 0)$, which only depend on the (conditional) distributions of counterfactual outcomes. Therefore, as long as these distributions are equal across causal models, the two approaches will produce the same answers to counterfactual questions, even if the variables are not almost-surely equal. Note that there is equality in law in Example 2 (which verifies positivity and conditional ignorability), but not necessarily in Example 1. In what follows, we show that equality in law does not always hold under the fundamental assumptions of causal inference.

3.2.3 The distributional level

We now address the most critical level: the one that concerns counterfactual inference. All answers to counterfactual questions in the structural framework and in the potential-outcome framework are respectively characterized by the joint probability distributions $\mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$ and $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}}))$. Should these distributions be equal, counterfactual reasoning would be equivalent between models. To prove that such an equality does not hold in general, we provide a *structural* identification of the law of potential outcomes.

Counterfactual inference in the potential-outcome framework can be achieved when both positivity and conditional ignorability hold. Although these two fundamental assumptions of causal inference cannot fully solve the identification of the potential outcomes, as they do not constrain the variables almost surely, they permit to completely identify the law of the potential outcomes in terms of observable quantities: they entail that $\mathcal{L}(Y_t \mid X = x) = \mathcal{L}(Y \mid X = x, T = t)$ for any $t \in \mathcal{T}$. As our main mathematical result, we propose a different kind of identification under the same assumptions. The theorem below identifies the law of the potential outcomes through the latent SCM $\mathcal{M}$, thereby enabling us to compare $(Y_t)_{t\in\mathcal{T}}$ with $(Y_{T=t})_{t\in\mathcal{T}}$.

Theorem 10 (Structural law of potential outcomes)**.** *Let $(Y_t)_{t\in\mathcal{T}}$ be random variables such that*

$$Y \overset{\mathbb{P}-a.s.}{=} \sum_{t\in\mathcal{T}} \mathbf{1}_{\{T=t\}} Y_t,$$

and suppose that T, X, and $(Y_t)_{t\in\mathcal{T}}$ verify Assumption 5 along with Assumption 6. Additionally, assume that $V := (T, X, Y)$ is the solution to an SCM $\mathcal{M} = \langle U, G\rangle$ satisfying Assumptions 2 and 7 where U_Y denotes the exogenous parents of Y. This notably entails that there exists a deterministic function F_Y such that $Y \overset{\mathbb{P}-a.s.}{=} F_Y(T, X, U_Y)$ where $U_Y \perp\!\!\!\perp (T, X)$. Then,

$$(T, X, (Y_t)_{t\in\mathcal{T}}) \overset{\mathcal{L}}{=} (T, X, (F_Y(t, X, U_Y))_{t\in\mathcal{T}}),$$

This means in particular that under the assumptions of Theorem 10, we concurrently have

$$(Y_t)_{t\in\mathcal{T}} \overset{\mathcal{L}}{=} (F_Y(t, X, U_Y))_{t\in\mathcal{T}},$$
$$(Y_{T=t})_{t\in\mathcal{T}} \overset{\mathbb{P}-a.s.}{=} (F_Y(t, X_{T=t}, U_Y))_{t\in\mathcal{T}}.$$

Therefore, $(Y_t)_{t\in\mathcal{T}}$ and $(Y_{T=t})_{t\in\mathcal{T}}$ are not necessarily equal in law since $\mathcal{L}(X) \neq \mathcal{L}(X_{T=t})$ in general. This difference also occurs at the individual level, that is conditional to X: $\mathcal{L}(Y_t \mid X = x) \neq \mathcal{L}(Y_{T=t} \mid X = x)$, since $\mathcal{L}(X_{T=t} \mid X = x) \neq \delta_x$ in general (here δ_x denotes the Dirac measure at x). As a consequence, in contrast to what many papers suggest, *the potential-outcome subscript notation and the do notation are not equivalent for counterfactual inference.* We provide a concrete example in the next section.

Nevertheless, one can derive potential-outcome counterfactuals from the latent SCM by intervening on *both* T and X instead of T only. According to the rules of do-calculus, $Y_{T=t,X=x} \overset{\mathbb{P}-a.s.}{=} F_Y(t, x, U_Y)$ whose law coincides with $\mathcal{L}(Y_t \mid X = x, T = t)$ under the assumptions of Theorem 10. We use this result in Remark 11.

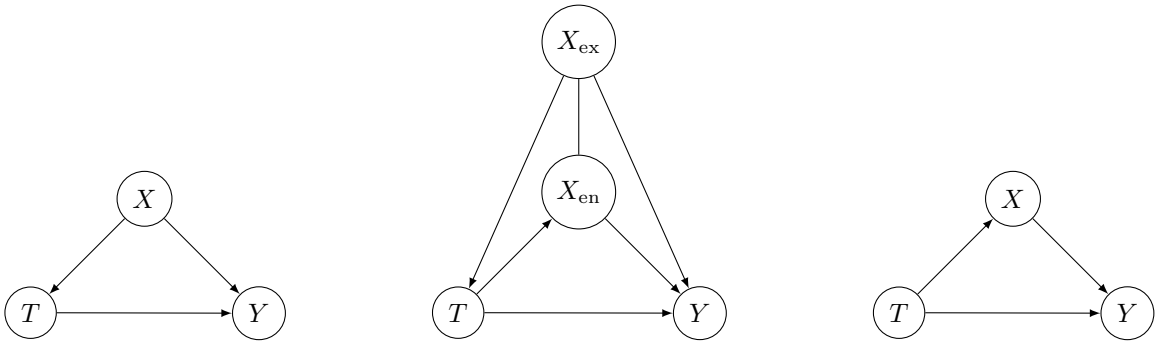
4 Discussion

In this section, we discuss the consequences of the three levels of difference explained in Section 3. Firstly, Section 4.1 specifies when counterfactual inference is equivalent or not between approaches. Secondly, Section 4.2 focuses on the importance of using distinct notations for potential outcomes and structural counterfactuals. Finally, Section 4.3 clarifies the apparent dissonance between our main results and the formal equivalence between frameworks mentioned in the literature.

4.1 When potential outcomes and structural counterfactuals are (not) equivalent

In this subsection, we study the assumptions and implications of Theorem 10 to specify when counterfactual inference in one framework gives (or not) the same results as in the other.

4.1.1 Nature of the treatment



(a) Controllable treatment (b) Partially controllable treatment (c) Noncontrollable treatment

Figure 1: The three possible configurations of the treatment under Assumption 7. $X_{\text{ex}} := X_{\text{Endo}(T)}$ denotes the parents of T in X while X_{en} are the remaining covariates. Exogenous variables are not represented. A single node can represent several variables. In (a), T does not impact X; in (b), T may impact some X-variables and some X-variables may impact T; in (c), X does not impact T.

Theorem 10 implies under the fundamental assumptions for causal inference that the law of potential outcomes coincides with the one of structural outcomes if: (1) X is not altered by do-interventions on T, or (2) Y is not impacted by X (as in Example 2). Let us focus on scenario (1), which is the most relevant in causal inference. The covariates are not altered by the treatment if T is not a parent of X in $\mathcal{M}$, as illustrated

in Figure 1a. Notably, this configuration encompasses various typical causal-inference scenarios: in clinical trials, the covariates X may influence the treatment allocation T but never the contrary. Therefore, both the Neyman-Rubin causal model and Pearl's causal model produce the same counterfactuals in common situations.

However, it is also mathematically possible for T to be a parent of several (or even all) covariates in $\mathcal{M}$. Figures 1b and 1c illustrate the possible causal graphs. In these situations, Theorem 10 does not guarantee equality in law between potential outcomes and structural counterfactuals. In fact, it follows from the expressions of $\mathcal{L}((Y_t)_{t\in\mathcal{T}})$ and $\mathcal{L}((Y_{T=t})_{t\in\mathcal{T}})$ that equality in law will generally not hold in such configurations. Consequently, confusion between the two causal approaches can lead to misleading results: the Neyman-Rubin causal model considers counterfactual outcomes at fixed X, whereas Pearl's causal model alters the covariates into $X_{T=t}$. These cases are empirically relevant, since people also rely on causal inference outside the scope of clinical trials, in settings where the treatment cannot be completely manipulated and thereby impacts the covariates. For example, T drives X but not the contrary (as in Figure 1c) in emblematic causal problems such as the Berkeley's admission paradox where T represents the sex and X the course choice [Bickel et al., 1975]. This is more generally true in the whole causal-fairness literature, where the variable to alter typically encodes the sex, the race, or the age of individuals (see for instance [Kusner et al., 2017, Barocas et al., 2019, Nilforoshan et al., 2022] for machine-learning-related research). Section 4.1.2 below exemplifies all these points by studying an SCM corresponding to Figure 1c.

All in all, under the fundamental assumptions of causal inference, equivalence of counterfactuals across causal frameworks depends on the relationships between the treatment and the covariates, as described in Figure 1. What distinguishes the different configurations is the nature of the so-called treatment. In particular, if the treatment can be assigned *a posteriori* to units (as in Figure 1a), then the two notions of counterfactuals coincide. If the treatment is an intrinsic feature of units (as in Figure 1c), such as individuals' race or sex, then structural counterfactuals and potential-outcomes counterfactuals are generally not equal.

4.1.2 Illustration: an immutable treatment and two different kinds of counterfactuals

Counterfactual reasoning can be defined as thinking about outcomes in hypothetical worlds where some circumstances changes from what factually happened while others are kept equal. Crucially, there is not a single way of reasoning counterfactually. Theorem 10 clearly shows that the potential-outcome framework compares worlds sharing the same observed features X but differing in T, while the structural account compares worlds sharing the same exogenous parents U_X but differing in T. Said differently, potential-outcome counterfactuals are *ceteris paribus* counterfactuals (i.e., all other things being kept equal) with respect to the covariates, whereas structural counterfactuals are *mutatis mutandis* counterfactuals (i.e., after changing what must be changed) with respect to the covariates. We emphasize that both definitions are perfectly legitimate, but convey distinct meanings and thereby correspond to different causal effects. Therefore, *they should not be employed for the same purpose*. Let us illustrate their implications on a concrete case.

The following fairness-inspired example generalizes and circumstantiates the discussion from [Kusner et al., 2017, Appendix S1]. The treatment status T indicates the gender, $T = 0$ standing for women and $T = 1$ standing for men; the covariate X quantifies the level of work experience, a higher score encoding a richer experience; the outcome Y evaluates a candidate's application for some position, a better score giving a higher

probability of acceptance. Suppose that these three variables are ruled by the following SCM fitting Figure 1c:

$$\begin{aligned} T &\overset{\mathbb{P}-a.s.}{=} U_T, \\ X &\overset{\mathbb{P}-a.s.}{=} \alpha T + U_X, \\ Y &\overset{\mathbb{P}-a.s.}{=} X + \beta T + U_Y, \end{aligned}$$

where α and β are deterministic parameters quantifying the causal influence of T onto respectively X and Y, and U_X represents the hidden merit or effort of an individual. Typically, a positive parameter α describes the societal inequalities leading women to have a lower level of work experience than men with equal merit U_X. Moreover, we suppose that Assumption 5 is true, and that $U_Y \perp\!\!\!\perp (U_T, U_X)$ so that Assumption 7 holds. Finally, we set two potential outcomes (Y_0, Y_1) verifying the consistency rule and Assumption 6. We consider the problem of assessing the causal effect of T onto Y conditional to $X = x$. In the potential-outcome approach this amounts to computing the conditional average treatment effect:

$$\begin{aligned} \mathrm{CATE}(x) &:= \mathbb{E}[Y_1 - Y_0 \mid X = x] \\ &= \mathbb{E}[(X + \beta + U_Y) - (X + U_Y) \mid X = x] \\ &= \beta, \end{aligned}$$

where we used Theorem 10. Observe that this first quantity measures only the *direct effect* of the treatment: it completely ignores the dependence of Y on T through X, as it involves only β. This is due to the fact that the CATE keeps the covariate X fixed, comparing two *distinct* individuals with identical profiles but different genders. In contrast, Pearl's approach assesses the following structural counterfactual effect:

$$\begin{aligned} \mathrm{SCE}(x) &:= \mathbb{E}[Y_{T=1} - Y_{T=0} \mid X = x] \\ &= \mathbb{E}[(X_{T=1} + \beta + U_Y) - (X_{T=0} + U_Y) \mid X = x] \\ &= \mathbb{E}[X_{T=1} - X_{T=0} \mid X = x] + \beta \\ &= \mathbb{E}[(\alpha + U_X) - U_X \mid X = x] + \beta \\ &= \alpha + \beta. \end{aligned}$$

Remark that this second quantity measures the *total effect* of the treatment: it takes into account the whole path of influence of T onto Y, involving both α and β. This comes from the fact that the SCE fixes the random seed U and not the covariates, comparing a *same* individual in two alternative realities where the gender is switched. Most importantly, CATE $\neq$ SCE if $\alpha \neq 0$, and consequently $\mathcal{L}((T, X, Y_{T=0}, Y_{T=1})) \neq \mathcal{L}((T, X, Y_0, Y_1))$.

From a fairness perspective, the CATE says that if $\beta = 0$, that is if T is not a *direct* cause of Y, then the application process if fair; whether it is unfair towards men or women when $\beta \neq 0$ depends on the sign of β. In contrast, the SCE says that if $\beta = -\alpha$, that is if the decision rule Y compensates the discrepancy of work experiences X across genders T, then the application process is fair. Each analysis points out a different notion of fairness: considering the SCE as a fairness criterion suggests that recruiters should correct societal inequalities by preferring women with potentially lower work experience but higher merit whereas relying on the CATE suggests it is only explicitly including the gender in the decision-rule pipeline that is unfair. Critically, if $\alpha \neq 0$, practitioners mixing potential outcomes with structural counterfactuals could reach contradictory conclusions on fairness.

Remark 11 (Computing direct effects from an SCM)**.** *One can still compute the* CATE*, that is the* direct *effect, using do-interventions on the SCM. Under the same assumptions as above,*

$$\begin{aligned}\mathbb{E}[Y_{T=1,X=x} - Y_{T=0,X=x}] &= \mathbb{E}[(x + \beta + U_Y) - (x + U_Y)],\\ &= \beta,\\ &= \text{CATE}(x).\end{aligned}$$

This is due to the fact that, more generally, $\mathcal{L}(Y_{T=t,X=x}) = \mathcal{L}(Y_t \mid X = x)$ *under the fundamental assumptions of causal inference. In contrast, one cannot always compute the* SCE*, that is the* indirect *effect from potential outcomes. Besides, we point out that even though the* CATE *can be theoretically derived from both a potential-outcome model and an SCM, the practical methods to estimate them from data differ between approaches. In the Neyman-Rubin causal framework, one only needs to estimate* $\mathbb{E}[Y \mid X = x, T = t]$ *(as it equals* $\mathbb{E}[Y_t \mid X = x]$*); in Pearl's causal framework, one needs to learn the full SCM beforehand (which is a notoriously difficult task), and then to apply the three-step procedure with* $\text{do}(T = t, X = x)$*.*

To sum-up, each approach has a different signification, and therefore corresponds to a specific way of reasoning counterfactually. This signifies that the difference between frameworks does not amount to practical considerations only. Analysts and researchers should also justify the chosen model depending on the kind of causal effects they want to compute.

4.1.3 Nonequivalence in applicability

Up until now, we have discussed the differences between frameworks from a theoretical viewpoint, not focusing on the practical aspects. Summing-up: to apply Pearl's causal framework, one must postulate a plausible structural causal model or infer it from data to then carry out do-calculus; to apply the Neyman-Rubin causal framework, one must find a set of covariates believed to satisfy conditional ignorability and then use statistical methods (e.g., matching, stratification, re-weighting, regression) to estimate observable quantities with causal meanings. As such, the two approaches are different in *how* they are applied; but there also exists a critical difference in *when* they can be applied due to the positivity assumption. Let us detail this point, as it also concerns a distinction between counterfactuals across frameworks.

Causal inference in the potential-outcome framework requires two fundamental assumptions: conditional ignorability (Assumption 6) and positivity (Assumption 5). While the second is testable in contrast to the first, it raises practical issues. It basically states that the distributions $\mathcal{L}(X \mid T = t)$ for $t \in \mathcal{T}$ share the same support, which is violated as soon as the groups represented by T bear unique properties. Consider for example that we study individuals where T encodes their genders, and that the covariates X specify their jobs (among other attributes). Positivity would not hold if gender-locked positions existed, and consequently we could not identify the counterfactual outcome *had she been a man* of every woman occupying a women-only job. In contrast, a structural causal model always allows such a computation. Therefore, there exist problems where the two causal models cannot be simultaneously applied to carry out counterfactual inference in practice. Not only the two kinds of counterfactuals should not be used for the same purpose (as they provably have different significations), but they cannot always be used for the same tasks.

4.2 On the importance of notations

All in all, it is inappropriate to exchange the do-notation and the potential-outcome notation, since they refer to variables with different distributions in common scenarios, such as almost every problem tackled by the

causal-fairness literature. Going further, we argue that distinct notations should always be used—even when the laws are identical—for two reasons: a practical reason and a conceptual one.

First, papers bridging the two causal approaches with a common notation often address favorable settings where the treatment is controllable, as in [Colnet et al., 2020, Section 5], which has no consequence on the consistency of counterfactual reasoning. While this may seem harmless, this could create confusion since the authors never explicitly state that they focus on controllable treatments, and even less justify why the laws of counterfactuals are equals. Therefore, people could wrongly believe that this equivalence holds in general. Notably, Makhlouf et al. [2020] study the application of both the structural account of counterfactuals and potential outcomes in fairness settings where the treatment encodes sex or race, and suggest that the appropriate choice of framework is mostly a matter of practical considerations. This could lead to contradictory results, as previously exemplified.

Second, a notation is the essence of mathematical object. Therefore, a same notation can be used for two objects just in case: (1) the objects have the same definition, (2) the objects are mathematically equal. The first two levels of difference we studied in Section 3 show that none of these conditions hold in general for structural counterfactuals and potential outcomes, even when their laws are equal (recall Example 2); the third level shows that none of these conditions hold in general for law-dependent quantities.

4.3 On the formal equivalence between frameworks

Before concluding, we emphasize that our work does not contradict the formal equivalence between causal frameworks addressed notably by Pearl [2009] and Richardson and Robins [2013]. We think, however, that some people possibly drew incorrect conclusions from this equivalence, in particular the systematic identification of $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}}))$ to $\mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$. In what follows, we try to explain where the confusion comes from.

4.3.1 Treating potential outcomes as structural counterfactuals is a choice

Let us start with a crucial reminder: in all generality, potential outcomes are merely variables $(Y_t)_{t\in\mathcal{T}}$ satisfying the consistency rule for a treatment of interest T and an outcome Y. In particular, even though the potential-outcome framework is impractical without the two fundamental assumptions of causal inference, one can perfectly consider outcomes that do not satisfy conditional ignorability in theory.[4] In this sense, structural counterfactuals can always be defined as potential outcomes according to Proposition 9 (which only requires acyclicity).

Interestingly, if one chooses to define potential outcomes as structural counterfactuals, then the Neyman-Rubin causal model and Pearl's causal model become two different languages to talk about the same objects. In the Neyman-Rubin causal model, assumptions for causal inference are generally framed as conditional-independence restrictions (e.g., conditional ignorability); in Pearl's causal framework, assumptions on causal relationships are generally framed in terms of graphical conditions (e.g., backdoor criterion). Both [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013] focus on unifying these two mathematical languages by providing rules for translating assumptions and theorems from one viewpoint to the other. This ensures what people often refer as the *logical* or *formal equivalence* between frameworks. It notably allows analysts to work symbiotically with both models, as long as equivalent assumptions are made across them.

However, we underline that defining potential outcomes as structural counterfactuals is a *choice*—it does not rest on any proof. When Pearl writes $Y_t := Y_{T=t}$ in [Pearl, 2009, Equation 3.51], claiming that the oper-

[4] We dot not focus on positivity since it does not constrain the potential outcomes.

ation $\text{do}(T = t)$ on the SCM gives a physical meaning to the vague "had T been t" of the potential outcome, this is nothing more than an arbitrary choice. Naively looking at the definitions of the Neyman-Rubin causal framework and of Pearl's causal framework, there is nothing that mathematically constrains potential outcomes to coincide with the structural counterfactuals of a given SCM (as demonstrated in Examples 1 and 2, and Section 4.1.2).[5] What [Pearl, 2009, Chapter 7] and [Richardson and Robins, 2013] prove is not the algebraic equality between potential outcomes and structural counterfactual. Instead, they show that *if* potential outcomes are chosen to be structural counterfactuals, then one can translate the assumptions made on potential outcomes into assumptions on the underlying causal graph and *vice versa*. However, these references did not make their definition choices clear enough. As a consequence, it seems that the formal equivalence between viewpoints has sometimes been understood as the systematic interchangeability between $(Y_t)_{t\in\mathcal{T}}$ and $(Y_{T=t})_{t\in\mathcal{T}}$, or between $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}}))$ and $\mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$.

4.3.2 The meaning of treating potential outcomes as distinct to structural counterfactuals

If we consider the two models to be different causal mechanisms rather than just different perspectives, then we can make nonequivalent assumptions on potential outcomes and structural counterfactuals. Supposing distinct axioms across models means giving distinct interpretations to their respective counterfactuals, which mathematically translates into $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}})) \neq \mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$. Theorem 10 and its implications precisely illustrate this aspect: assuming the conditional ignorability of the potential outcomes but not making an equivalent hypothesis on the SCM defining the structural counterfactuals entails that counterfactuals do not have the same law in general hence not the same interpretation. As illustrated in Section 4.1.2, the conditional-ignorability assumption defines potential outcomes as counterfactuals switching the treatment but keeping all other variables equal, whereas Pearl's do-intervention on the treatment defines structural counterfactuals altering the remaining variables. In a scenario where the treatment is controllable, these definitions happen to coincide. As such, the interest of defining potential outcomes as distinct to structural counterfactuals is notably to estimate the *ceteris paribus* causal effects of immutable treatments by leveraging statistical methods on observable distributions rather than by learning a fully specified plausible SCM and to then apply the three-step procedure with $\text{do}(T = t, X = x)$ (as mentioned in Remark 11).

An important corollary of Section 4.1 is that one *cannot always* make an equivalent assumption to conditional ignorability on the structural counterfactuals. From the SCM perspective, conditional ignorability is simply $(Y_{T=t})_{t\in\mathcal{T}} \perp\!\!\!\perp T \mid X$, which under Assumption 7 occurs when the treatment is fully controllable (that is fitting Figure 1a). Therefore, in the many aforementioned settings where the treatment is factually not controllable (that is fitting Figure 1c) potential outcomes verifying conditional ignorability must be defined as distinct to the *true* structural counterfactuals.[6] Critically, this means that even though it is generally implicitly done, *not* defining potential outcomes as structural counterfactuals is actually something common in the scientific literature. Notably, [Li et al., 2017, Glymour and Spiegelman, 2017, Khademi et al., 2019, Khademi and Honavar, 2020, Makhlouf et al., 2020, Qureshi et al., 2020] rely on (or refer to) the potential-outcome framework equipped the fundamental assumption of causal inference to understand the influence of

[5] From a more philosophical angle, [Markus, 2021, Section 2.1] made a similar remark to argue that the two frameworks were *weakly* equivalent rather than *strongly* equivalent.

[6] More generally, the logical equivalence between framework ensures that there always exists an SCM that induces the law of given potential outcomes, but does not guarantee that this SCM correctly describe the observations (T, X, Y). The discussion that followed Theorem 10 is based on the principle that the considered SCM is the true one, not necessarily one that fits the potential-outcome assumptions.

a noncontrollable treatment like sex, race, and biological factors.[7] As explained, $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}}))$ generally differs from $\mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$ in such studies. Therefore, if we consider this corpus of the causal-inference literature to be admissible, then we must accept that unifying potential outcomes and structural counterfactuals is not an obligation.

4.3.3 What is the best paradigm?

The results from Section 3 show that leveraging a structural causal model as the axiomatic characterization of potential outcomes is a choice. This outlines two paradigms for defining and applying potential outcomes: in synergy with the latent SCM generating the observations; as hypothetical outcomes verifying conditional ignorability. Settling the debate on whether there is a most relevant or inappropriate choice lies outside the scope of this paper. On the one hand, Pearl has advocated for long to *always* use the potential-outcome framework in symbiosis with an SCM, as the latter enables to formulate its conditional-independence conditions deemed nebulous in the intelligible language of causal graphs. On the other hand, deviating from this rule to work with SCM-free potential outcomes allows to compute the direct causal effects of noncontrollable treatments through statistical methods, as done in the aforementioned studies. Each approach can be legitimate; *what crucially matters is having a clear understanding of the produced counterfactuals*. Modeling potential outcomes through an SCM formally defines their "had T been t" as the operation $\text{do}(T = t)$ which changes the remaining variables accordingly. Dressing potential outcomes with conditional ignorability defines their "had T been t" as switching T into t while keeping the remaining variables unchanged. We hope this discussion to make clear the (sometimes implicit) signification of potential outcomes in the causal-inference literature.

To summarize, we think that the scope and implications of the formal equivalence between causal frameworks can be misleading. In particular, it does not mean that $\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}})) = \mathcal{L}((T, X, (Y_{T=t})_{t\in\mathcal{T}}))$ in general; it means that such an equality in law holds *if equivalent assumptions are made on the counterfactual outcomes across models*. However, using the potential-outcome framework to estimate direct causal effects requires assuming conditional ignorability, which cannot always be assumed on the structural counterfactuals from the *true* latent SCM due to the possibly immutable nature of the treatment. This why we recommend to present the Neyman-Rubin causal framework and Pearl's causal framework as distinct ways of reasoning counterfactually, that coincide under specific assumptions and choices.

5 Conclusion

In this paper, we superimposed Pearl's causal framework and the Neyman-Rubin causal framework without presuppositions to show that structural counterfactual outcomes and potential outcomes do not coincide at three levels: they are defined differently, they are not equal in general, they do not have the same law in general. To prove the third level of difference, we expressed the law of potential outcomes in terms of the latent structural causal model under classical causal-inference assumptions. On the basis of this result, we gave a detailed interpretation of counterfactuals in each causal framework, specifying when they entailed different conclusions. More specifically, counterfactual inference with potential outcomes under conditional ignorability yields *ceteris paribus* counterfactuals, whereas counterfactual inference with a do-intervention on a structural causal model yields *mutatis mutandis* counterfactuals. If the cause of interest is immutable, these constructions are generally

[7] Other articles, for instance [Ridgeway, 2006] and [Gaebler et al., 2022], explicitly focus on sex and race as *perceived* by a decider. Such a perception could depend on the covariates, hence not be immutable.

not equivalent. For these reasons, we call the community to not interchangeably use the do-notation and the potential-outcome notation, unless the justification is explicitly made.

We emphasize that this work is not an argument in favor of using one causal model rather than the other, or against the formal equivalence between frameworks. It is meant to shed light on the different mathematical choices that analysts can make when working with counterfactual outcomes, and to precise their implications in order to prevent incorrect or ambiguous conclusions in causal studies. In doing this paper, we hope to clarify the similarities and differences between the two major causal approaches.

A Proofs

Proof of Proposition 4 Since $\mathcal{M}$ is acyclical, there exists a topological ordering on the indices in $\mathcal{I}$, and therefore on the subset J. This means in particular that there exist some $j \in J$ such that G_j takes only variables in V_I as endogenous inputs. Starting from these indices, and recursively substituting along the topological ordering produces a measurable F_J such that

$$V_J \overset{\mathbb{P}-a.s.}{=} F_J(V_{\mathrm{Endo}(J)\setminus J}, U_{\mathrm{Exo}(J)}).$$

Note that $\mathrm{Endo}(J) \setminus J \subseteq I$. Carrying out the same substitution on the intervened model $\mathcal{M}_{V_I=v_I}$ with solution $\tilde{V}$ gives

$$\tilde{V}_J \overset{\mathbb{P}-a.s.}{=} F_J(v_{\mathrm{Endo}(J)\setminus J}, U_{\mathrm{Exo}(J)}),$$

while by definition $\tilde{V}_I \overset{\mathbb{P}-a.s.}{=} v_I$. ■

Proof of Lemma 8 By assumption, the random vector $V := (T, X, Y)$ is the solution to an acylical structural causal model where U_T, U_X and U_Y denote the exogenous parents of respectively T, X and Y. Since additionally $Y_{\mathrm{Endo}(T)} = Y_{\mathrm{Endo}(X)} = \emptyset$, Proposition 4 ensures the existence of a measurable function $F_{T,X}$ such that $(T, X) \overset{\mathbb{P}-a.s.}{=} F_{T,X}(U_T, U_X)$. Therefore, if $U_Y \perp\!\!\!\perp (U_T, U_X)$, then $U_Y \perp\!\!\!\perp (T, X)$. ■

Proof of Proposition 9 Let $t \in \mathcal{T}$. By assumption, the random vector $V := (T, X, Y)$ is the solution to an acylical structural causal model. We write U_X and U_Y the exogenous parents of respectively X and Y. Therefore, by partitioning V into T and (X, Y), Proposition 4 guarantees the existence of a measurable function $F_{X,Y}$ such that

$$(X, Y) \overset{\mathbb{P}-a.s.}{=} F_{X,Y}(T, U_X, U_Y)$$
$$(X_{T=t}, Y_{T=t}) \overset{\mathbb{P}-a.s.}{=} F_{X,Y}(t, U_X, U_Y).$$

Therefore, selecting the coordinates corresponding to Y furnishes a measurable function $\tilde{F}_Y$ such that

$$Y \overset{\mathbb{P}-a.s.}{=} \tilde{F}_Y(T, U_X, U_Y)$$
$$Y_{T=t} \overset{\mathbb{P}-a.s.}{=} \tilde{F}_Y(t, U_X, U_Y).$$

These identities hold on a measurable set $\Omega^* \subseteq \Omega$ such that $\mathbb{P}(\Omega^*) = 1$. To conclude, simply observe that for any $\omega \in \Omega^*$ such that $T(\omega) = t$, we have

$$Y(\omega) = \tilde{F}_Y(t, U_X(\omega), U_Y(\omega)) = Y_{T=t}(\omega).$$

■

Proof of Theorem 10 Let us compute the conditional joint distribution $\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x, T = t)$ which is well-defined for all $x \in X(\Omega)$ and $t \in \mathcal{T}$ by positivity. The consistency rule entails that

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x, T = t) = \mathcal{L}((Y_0, \ldots, Y_{t-1}, Y, Y_{t+1}, \ldots, Y_N) \mid X = x, T = t).$$

Moreover, according to $\mathcal{M}$ the observed outcome can be written as $Y \overset{\mathbb{P}-a.s.}{=} F_Y(T, X, U_Y)$, leading to

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x, T = t) = \mathcal{L}((Y_0, \ldots, Y_{t-1}, F_Y(t, x, U_Y), Y_{t+1}, \ldots, Y_N) \mid X = x, T = t).$$

Next, recall that Assumption 7 entails through Lemma 8 that $U_Y \perp\!\!\!\perp (T, X)$. Therefore, it follows from $(Y_{t''})_{t''\in\mathcal{T}} \perp\!\!\!\perp T \mid X$ that for any $t' \neq t$ the above equality is equivalent to

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x) = \mathcal{L}((Y_0, \ldots, Y_{t-1}, F_Y(t, x, U_Y), Y_{t+1}, \ldots, Y_N) \mid X = x, T = t').$$

Then, using once the again the consistency rule we obtain

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x) = \mathcal{L}((Y_0, \ldots, F_Y(t, x, U_Y), \ldots, Y_{t'-1}, Y, Y_{t'+1}, \ldots) \mid X = x, T = t'),$$

and the expression of Y through F_Y yields

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x) = \mathcal{L}((Y_0, \ldots, F_Y(t, x, U_Y), \ldots, F_Y(t', x, U_Y), \ldots) \mid X = x, T = t').$$

We repeat this step by conditioning on all possible values of T to finally obtain

$$\mathcal{L}((Y_{t''})_{t''\in\mathcal{T}} \mid X = x) = \mathcal{L}((F_Y(t'', x, U_Y)_{t''\in\mathcal{T}}) \mid X = x) = \mathcal{L}((F_Y(t'', x, U_Y)_{t''\in\mathcal{T}})).$$

It then follows from $U_Y \perp\!\!\!\perp (T, X)$ that for any $t \in \mathcal{T}$,

$$\mathcal{L}((t, x, (Y_{t''})_{t''\in\mathcal{T}}) \mid X = x, T = t) = \mathcal{L}((t, x, (F_Y(t'', x, U_Y)_{t''\in\mathcal{T}})) \mid X = x, T = t).$$

Therefore, marginalizing on (T, X) yields

$$\mathcal{L}((T, X, (Y_t)_{t\in\mathcal{T}})) = \mathcal{L}((T, X, (F_Y(t, X, U_Y))_{t\in\mathcal{T}})).$$

■

References

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404, 1975.

C. R. Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

J. Gaebler, W. Cai, G. Basse, R. Shroff, S. Goel, and J. Hill. A causal framework for observational studies of discrimination. *Statistics and public policy*, 9(1):26–48, 2022.

M. M. Glymour and D. Spiegelman. Evaluating public health interventions: 5. causal inference in public health research—do sex, race, and biological factors cause health outcomes? *American journal of public health*, 107 (1):81–85, 2017.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

G. W. Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

A. Khademi and V. Honavar. Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13839–13840, 2020.

A. Khademi, S. Lee, D. Foley, and V. Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pages 2907–2914, 2019.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc., 2017.

J. Li, J. Liu, L. Liu, T. D. Le, S. Ma, and Y. Han. Discrimination detection by causal effect estimation. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1087–1094. IEEE, 2017.

K. Makhlouf, S. Zhioua, and C. Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

K. A. Markus. Causal effects and counterfactual conditionals: contrasting Rubin, Lewis and Pearl. *Economics & Philosophy*, 37(3):441–461, 2021.

B. Neal. *Introduction to causal inference from a machine learning perspective*. bradyneal.com, 2020. `https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf`.

H. Nilforoshan, J. D. Gaebler, R. Shroff, and S. Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. Brief report: On the consistency rule in causal inference:" axiom, definition, assumption, or theorem?". *Epidemiology*, pages 872–875, 2010.

B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, and D. Pedreschi. Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, 54:425–437, 2020.

T. S. Richardson and J. M. Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences*, 128(30):2013, 2013. Working Paper.

G. Ridgeway. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of quantitative criminology*, 22:1–29, 2006.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.