Backtracking Counterfactuals

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Counterfactual reasoning—envisioning hypothetical scenarios, or possible worlds, where some circumstances are different from what (f)actually occurred (counter-to-fact)—is ubiquitous in human cognition. Conventionally, counterfactually-altered circumstances have been treated as "small miracles" that locally violate the laws of nature while sharing the same initial conditions. In Pearl's structural causal model (SCM) framework this is made mathematically rigorous via *interventions* that modify the causal laws while the values of exogenous variables are shared. In recent years, however, this purely interventionist account of counterfactuals has increasingly come under scrutiny from both philosophers and psychologists. Instead, they suggest a *backtracking* account of counterfactuals, according to which *the causal laws remain unchanged* in the counterfactual world; differences to the factual world are instead "backtracked" to altered initial conditions (exogenous variables). In the present work, we explore and formalise this alternative mode of counterfactual reasoning within the SCM framework. Despite ample evidence that humans backtrack, the present work constitutes, to the best of our knowledge, the first general account and algorithmisation of backtracking counterfactuals. We discuss our backtracking semantics in the context of related literature and draw connections to recent developments in explainable artificial intelligence (XAI).

Keywords: Causal reasoning, backtracking, counterfactual explanations, explainable AI, XAI

1. Introduction

In a deterministic world, everything that happens is uniquely determined by the laws of nature and the initial (or background) conditions.[1] Counterfactuals invite us to imagine what the world would look like if some events which did occur, had in fact not occurred. As a result, in order to make sense of counterfactuals in a deterministic world one is immediately presented with the following dilemma: for events to have been different, either the laws of nature would have had to be violated, or the background conditions would have had to be different.

Following in the footsteps of Lewis (1973, 1979), the majority of philosophers have opted for the first option: counterfactuals are to be evaluated by imagining "small miracles" that ensure those events which are counter-to-fact to occur by locally violating the laws of nature, thereby disconnecting these events from their causes, and keeping the background conditions fixed. In other words, counterfactuals are *non-backtracking*. Pearl (2009, p.239) objected to Lewis's miracles and the possible-world semantics that comes with it, instead replacing these with structural equations and the well-behaved notion of an intervention that those allow:

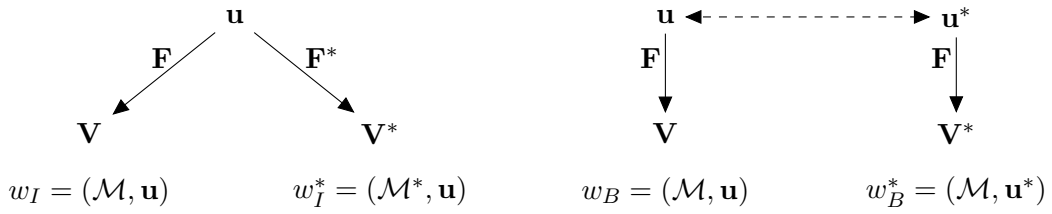
> Lewis's elusive "miracles" are replaced by principled minisurgeries, [...] which represent the minimal change (to a model) necessary for establishing the antecedent.

Still, Pearl's *interventional* counterfactuals are also a commitment to the first option of our dilemma.

Recently this status quo has come under pressure by several philosophers who argue in favour of the second option (Dorr, 2016; Loewer, 2007, 2020; Esfeld, 2021): counter-to-fact events are to be explained by imagining small changes to the background conditions that would result in the

1. Even if the world is fundamentally non-deterministic at the quantum level, the point still holds for deterministic *models* of the world. Causal models in particular can be given a deterministic interpretation, see (2.2).

© 2023 .



(a) Interventional counterfactual (b) Backtracking counterfactual

Figure 1: **Illustration of the main conceptual difference between interventional and backtracking counterfactuals.** (a) For *interventional* counterfactuals, the factual world w_I and counterfactual world w_I^* share the exact same background conditions $\mathbf{u}$. Potential contradictions between the factual outcome $\mathbf{V}$ and the counterfactual outcome $\mathbf{V}^*$ are resolved through changes to the causal laws $\mathbf{F}$ (by means of intervention), giving rise to the modified laws $\mathbf{F}^*$ and submodel $\mathcal{M}^*$. (b) For *backtracking* counterfactuals, on the other hand, the factual world w_B and counterfactual world w_B^* share the same unmodified causal laws $\mathbf{F}$. To allow for the factual and counterfactual outcomes $\mathbf{V}$ and $\mathbf{V}^*$ to differ, the respective background conditions $\mathbf{u}$ and $\mathbf{u}^*$ may differ, too, with a preference for minimal changes, that is, for the two worlds to be as close as possible.

counterfactual events instead of the actual ones, while the laws of nature remain unchanged. In other words, counterfactuals are *backtracking*.[2] Others propose a combination of both options, suggesting that each is appropriate under different circumstances (Fisher, 2017a,b; Woodward, 2021). This dispute is not merely academic: empirical work by psychologists now confirms that, depending on the context, people indeed switch between interpreting counterfactuals according to the first and the second option (Rips, 2010; Gerstenberg et al., 2013; Lucas and Kemp, 2015). Moreover, counterfactual explanations are taking up a prominent role in the work on explainable AI; their interpretation thus has significant repercussions for real-life applications (Wachter et al., 2017).

While non-backtracking, interventional counterfactuals have been given a well-defined semantics within Pearl's framework of structural causal models (SCMs), the same cannot be said for backtracking counterfactuals. In this paper we offer the first general formal semantics for backtracking counterfactuals within SCMs. Our general semantics can be made more specific in various ways, depending on the particular purpose that is at stake. We do not view these backtracking semantics as an alternative to the standard non-backtracking semantics: the value of the latter for causal reasoning is undeniable. Rather, we see both semantics as being fit for different reasoning tasks.

> it is appropriate to use backtracking counterfactuals to answer [...] how the past would have had to have been different had the present been different. [...] backtracking counterfactuals are important in diagnostic reasoning. However, this does not mean that it is misguided to use non-backtracking counterfactuals to answer other sorts of questions such as those having to do with whether Cs cause Es. The two kinds of counterfactuals are just different, with different truth conditions (Woodward, 2021, p. 206)

Structure and Contributions. We start by reviewing the SCM framework and its interventional semantics for counterfactual reasoning (§ 2). In § 3, we then provide a comprehensive account of the alternative, backtracking mode of counterfactual reasoning by explaining the high-level intuition (§ 3.1, see Figure 1 for an overview), introducing a formal semantics (§ 3.2), discussing some of the design choices and desiderata (§ 3.3), and providing some first theoretical insights (§ 3.4). We compare our semantics to previous attempts and discuss related work from cognitive science in § 4. We then highlight the use of backtracking counterfactuals for explainable artificial intelligence (XAI) (§ 5), and conclude with an outlook and suggestions for future work (§ 6).

2. or *observational*, we use both terms interchangeably; likewise for non-backtracking and interventional

2. Preliminaries: Structural Causal Models and Interventional Counterfactuals

The following definitions of SCMs and their counterfactual semantics follow Pearl (2009, § 7.1).

A **causal model** is a triple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ where: (i) $\mathbf{U}$ is a set $\{U_1, ..., U_m\}$ of **exogenous** (background) variables determined by factors outside the model; (ii) $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of **endogenous** variables determined by variables in the model, that is, variables in $\mathbf{U} \cup \mathbf{V}$; and (iii) $\mathbf{F}$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$ s.t. each f_i is a mapping from (the respective domains of) $\mathbf{U}_i \cup \mathbf{PA}_i$ to V_i, where $\mathbf{U}_i \subseteq \mathbf{U}$ and $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$, and the entire set $\mathbf{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. In other words, each f_i in the **structural equations**, or **causal laws**,

$$V_i := f_i(\mathbf{PA}_i, \mathbf{U}_i) \qquad i = 1, \ldots, n, \tag{2.1}$$

assigns a value to V_i that depends on (the values of) a select set of variables in $\mathbf{U} \cup \mathbf{V}$, and the entire set $\mathbf{F}$ has a unique solution $\mathbf{V}(\mathbf{u})$. The latter is ensured, e.g., in acyclic ("recursive") systems.

The **causal diagram** $G(\mathcal{M})$ associated with causal model $\mathcal{M}$ is the directed graph in which each node corresponds to a variable and directed edges point from members of $\mathbf{PA}_i$ and $\mathbf{U}_i$ toward V_i. Since the exogenous (background) variables $\mathbf{U}$ are typically unobserved, it is common to only consider the subset of $G(\mathcal{M})$ corresponding to its projection onto $\mathbf{V}$, where shared exogenous parents between some V_i and V_j are indicated with a bi-directed arrow $V_i \leftrightarrow V_j$

For a subset of endogenous variables $\mathbf{X} \subseteq \mathbf{V}$ and a realisation $\mathbf{x}$ thereof, the **submodel** $\mathcal{M}_{\mathbf{x}}$ of $\mathcal{M}$ is the model $\mathcal{M}_{\mathbf{x}} = (\mathbf{U}, \mathbf{V}, \mathbf{F}_{\mathbf{x}})$ where $\mathbf{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} := \mathbf{x}\}$. The **effect of action** $do(\mathbf{X} = \mathbf{x})$ on $\mathcal{M}$ is given by the submodel $\mathcal{M}_{\mathbf{x}}$. A **causal world** w is a pair $(\mathcal{M}, \mathbf{u})$ where $\mathcal{M}$ is a causal model and $\mathbf{u}$ is a particular realization of the background variables $\mathbf{U}$. The **potential response** of $\mathbf{Y} \subseteq \mathbf{V}$ to action $do(\mathbf{X} = \mathbf{x})$ in world $w = (\mathcal{M}, \mathbf{u})$, denoted $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$, is the solution for $\mathbf{Y}$ of the set of equations $\mathbf{F}_{\mathbf{x}}$, that is, $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{Y}_{\mathcal{M}_{\mathbf{x}}}(\mathbf{u})$. The **counterfactual** sentence "$\mathbf{Y}$ would be $\mathbf{y}$ (in situation $\mathbf{u}$), had $\mathbf{X}$ been $\mathbf{x}$" is then interpreted as the equality $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}$. The part "had $\mathbf{X}$ been $\mathbf{x}$" is called the (counterfactual) **antecedent**.

A **probabilistic causal model** is a distribution over causal worlds, that is, a pair $(\mathcal{M}, P(\mathbf{U}))$ where $\mathcal{M}$ is a causal model and $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$. The function $P(\mathbf{U})$, together with the fact that each endogenous variable is a function of $\mathbf{U}$, defines a **distribution over endogenous variables**: for any $\mathbf{Y} \subseteq \mathbf{V}$ we have[3]

$$P(\mathbf{y}) := P(\mathbf{Y} = \mathbf{y}) = \textstyle\sum_{\mathbf{u}} P(\mathbf{u})\, \mathbf{1}_{\{\mathbf{Y}(\mathbf{u}) = \mathbf{y}\}} \tag{2.2}$$

where $\mathbf{1}$ denotes the indicator function. The **probability of counterfactuals** is defined analogously, through the potential responses induced by different submodels: for any not necessarily disjoint sets of variables $\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$, we have

$$P(\mathbf{Y}_{\mathbf{x}} = \mathbf{y}, \mathbf{Z}_{\mathbf{w}} = \mathbf{z}) = \textstyle\sum_{\mathbf{u}} P(\mathbf{u}) \mathbf{1}_{\{\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}\}} \mathbf{1}_{\{\mathbf{Z}_{\mathbf{w}}(\mathbf{u}) = \mathbf{z}\}}. \tag{2.3}$$

In particular, $P(\mathbf{Y}_{\mathbf{x}} = \mathbf{y}, \mathbf{X} = \mathbf{x}')$ and $P(\mathbf{Y}_{\mathbf{x}} = \mathbf{y}, \mathbf{Y}_{\mathbf{x}'} = \mathbf{y}')$ are well defined using SCM semantics, even though $\mathbf{x} \neq \mathbf{x}'$ may be incompatible and thus cannot be measured simultaneously.[4]

To notationally distinguish the factual variables $\mathbf{V}$ from counterfactual versions, or copies, thereof, the latter are sometimes also denoted with an asterisk $\mathbf{V}^*$ (Balke and Pearl, 1994).

Of special interest are counterfactuals of the form $P(\mathbf{Y}^*_{\mathbf{x}^*} = \mathbf{y}^* \mid \mathbf{z})$ that are conditional on a (f)actual observation $\mathbf{z}$. These can be computed via the following three-step procedure:

3. To focus on the main points, ease the notation, and avoid measure theoretic details, we present all formulae for *discrete* random variables in terms of *sums* of probability *mass* functions; the analogue for *continuous* random variables would involve *integrals* of probability density functions (assuming existence of densities w.r.t. the Lebesgue measure).
4. Some therefore consider statements about such expressions as fundamentally unscientific, see, e.g., Dawid (2000).

1. **Abduction:** Update $P(\mathbf{U})$ by the evidence $\mathbf{z}$ to obtain $P(\mathbf{U} \mid \mathbf{z})$.
2. **Action:** Modify $\mathcal{M}$ by the action $do(\mathbf{x}^*)$ to obtain the submodel $\mathcal{M}_{\mathbf{x}^*}$.
3. **Prediction:** Use the modified model $(\mathcal{M}_{\mathbf{x}^*}, P(\mathbf{U} \mid \mathbf{z}))$ to compute the probability of $\mathbf{Y}^*$.

Example 1 (Interventional counterfactual) *Consider a probabilistic causal model $(\mathcal{M}, P(\mathbf{U}))$ with $\mathbf{V} = \{X, Y, Z\}$, $\mathbf{U} = \{U_X, U_Y, U_Z\}$, causal laws $\mathbf{F}$ given by*

$$X := U_X, \qquad Y := X + U_Y, \qquad Z := X + Y + U_Z, \tag{2.4}$$

and $P(\mathbf{U})$ being the multivariate standard isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_3)$.

Suppose that we make the factual observation $(X = 1, Y = 2, Z = 2)$ and wish to reason about the (interventional) counterfactual "what would have been, had Y been 3", that is, we are interested in $P((X^*, Z^*)_{Y^*=3} \mid X = 1, Y = 2, Z = 2)$, see Figure 2a for a visualisation.

To this end, we follow the aforementioned three-step procedure:

1. ***Abduction:*** *we find the posterior $P(\mathbf{U} \mid X = 1, Y = 2, Z = 2)$ to be a point mass on $(U_X = 1, U_Y = 1, U_Z = -1)$.*
2. ***Action:*** *we modify $\mathcal{M}$ to obtain the submodel $\mathcal{M}_{Y^*=3}$ as*

$$X^* := U_X, \qquad Y^* := 3, \qquad Z^* := X^* + Y^* + U_Z; \tag{2.5}$$

   Note that this step alters the mechanism for the counterfactual antecedent Y^* and removes its dependence on X^* and U_Y, as shown in Figure 2a.
3. ***Prediction:*** *we compute the push-forward of $P(\mathbf{U} \mid X = 1, Y = 2, Z = 2)$ via (2.5) which yields a point mass on $(X^* = 1, Z^* = 3)$. That is, X (being a parent of Y) would have remained unaffected, but Z (being a child of Y) would have increased by one.*[5]

Remark 1 (Forward-tracking) *In acyclic causal models, any manipulation $\mathbf{F}_{\mathbf{x}}$ of the causal laws $\mathbf{F}$ such as $Y^* := 3$ in (2.5) only has "downstream" effects: the change only propagates to descendants of the intervened-upon variables, while any non-descendants remain unaffected. In this sense, interventions and thus also interventional counterfactuals are purely* forward-tracking.

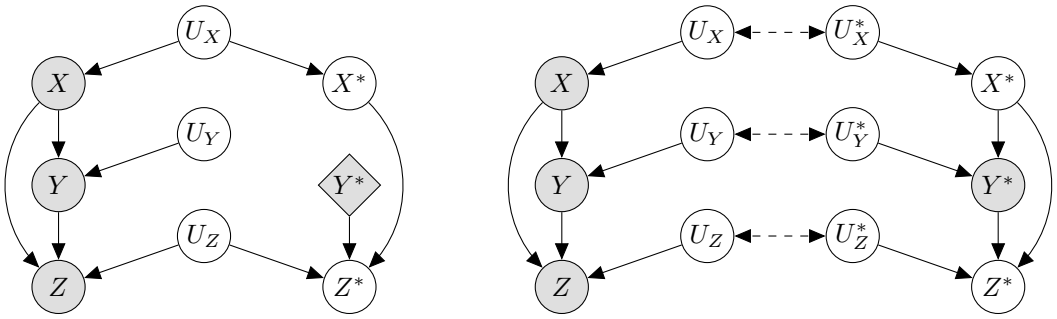
The *interventional* approach to computing counterfactuals in SCMs thus assumes that the background conditions $\mathbf{u}$ are shared between the factual and counterfactual worlds and instead relies on modifying the causal laws (the *action* step) to explain possible discrepancies (such as the difference $Y = 2 \neq Y^* = 3$ in Example 1), as illustrated in Figures 1a and 2a. As Pearl (2009, p.205) puts it:

> [It] interprets the counterfactual phrase "had $\mathbf{X}$ been $\mathbf{x}$" in terms of a *hypothetical modification of the equations in the model*; it simulates an *external action* (or spontaneous change) that modifies the actual course of history and enforces the condition "$\mathbf{X} = \mathbf{x}$" with minimal *change of mechanisms*. This [...] permits $\mathbf{x}$ to differ from the current value of $\mathbf{X}(\mathbf{u})$ without creating logical contradiction; it also suppresses abductive inferences (or *backtracking*) from the counterfactual antecedent $\mathbf{X} = \mathbf{x}$ [*emphasis ours*]

5. $(X^*, Z^*)_{Y^*=3} \mid (X = 1, Y = 2, Z = 2)$ in Example 1 differs from the observational and interventional distributions

$$P(X, Z \mid Y = 3) = \mathcal{N}\left(\begin{pmatrix}1.5\\4.5\end{pmatrix}, \begin{pmatrix}0.5 & 0.5\\0.5 & 0.5\end{pmatrix}\right), \qquad P((X, Z)_{Y=3}) = \mathcal{N}\left(\begin{pmatrix}0\\3\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}\right),$$

both of which preserve uncertainty about (X, Z). (In contrast, the interventional counterfactual is fully determined in Example 1, because the shared background conditions $\mathbf{u}$ could be uniquely inferred from the factual observation.) These three modes of reasoning—observational, interventional, and counterfactual—constitute increasingly difficult tasks, each requiring additional data or assumptions over the previous, and form the three rungs of the so-called "Ladder of Causation" (Pearl and Mackenzie, 2018) or "Pearl Causal Hierarchy" (Bareinboim et al., 2022).



(a) Interventional counterfactual (Example 1) (b) Backtracking counterfactual (Example 2)

Figure 2: **Graphical comparison between the interventional and backtracking approaches to computing the counterfactual $P(X^*, Z^* \mid Y^*, X, Y, Z)$.** Shaded nodes are "observed" (or hypothesised to be known, in the case of counterfactual/starred variables), while white nodes are unobserved/latent. (a) For *interventional* counterfactuals, the counterfactual antecedent Y^* is assumed to arise from a small miracle in the form of a (local) change to the causal laws, illustrated by the diamond node. This intervention removes the dependence of Y^* on its causal parents X^* and U_Y, hence the missing arrows into Y^*. The exogenous variables are assumed to be shared, giving rise to a so-called "twin network" (Balke and Pearl, 1994). As a consequence, non-descendants of Y^* (here, only the parent X^*), remain the same and only descendants of Y^* (here, only the child Z^*) are affected. (b) For *backtracking* counterfactuals, the causal relationship between the counterfactual antecedent Y^* and its parents X^* and U_X^* is preserved. This necessitates the introduction of new counterfactual exogenous variables (U_X^*, U_Y^*, U_Z^*) which should be "close" to the original (U_X, U_Y, U_Z) but may differ from it to absorb, or explain, the difference between Y and Y^*. As a result of such backtracking, not only descendants (Z^*) but also ancestors (X^*) may change.

3. Backtracking Counterfactuals

We now explore and formalise an alternative, non-interventional mode of counterfactual reasoning that does not involve such "change of mechanism" by "external action", and instead relies on so-called *backtracking* all changes to changes in the values of exogenous variables.

3.1. Intuition and Main Idea

The main idea behind backtracking counterfactuals—and its core conceptual difference to interventional counterfactuals—is that the causal laws, not the background conditions, are shared between the factual and counterfactual worlds. Consequently, the exogenous variables need to be allowed to differ to accommodate possibly contradictory facts and counterfacts. This is illustrated in Figure 1b.

The name "backtracking" refers to the act of updating upstream variables to explain counterfacts without breaking the causal laws. Such backtracking is not needed for interventional counterfactuals for which the modified equations take care of this. Let us first explore this idea through an example.

Example 2 (Backtracking counterfactual) *Consider the same model* (2.4) *and factual observation* $(X = 1, Y = 2, Z = 2)$ *from Example 1. Suppose that now we are instead interested in the* backtracking *counterfactual "what would have been, had* Y *instead been* observed *to be* $Y^* = 3$*".*

From abduction on the factual observation, we obtain the same posterior $P(\mathbf{U} \mid X = 1, Y = 2, Z = 2)$ *being a point mass on* $(U_X = 1, U_Y = 1, U_Z = -1)$*. However, in contrast to* (2.5)*, the structural equations in the counterfactual world now take the form:*

$$X^* := U_X^*, \qquad Y^* := X^* + U_Y^* = 3, \qquad Z^* := X^* + Y^* + U_Z^*; \tag{3.1}$$

We highlight the following two key differences between (3.1) *and its interventional counterpart* (2.5)*: (i) the causal laws remain fully in place, as apparent from the dependence of* Y^* *on* X^* *and* U_Y^*

in (3.1); (ii) the exogenous variables are not (necessarily) shared, as apparent from the introduction of new counterfactual background variables (U_X^*, U_Y^*, U_Z^*) in (3.1).

Now, in order for $Y^* := X^* + U_Y^* = 3$ to hold, at least one of X^* or U_Y^* need to differ from their factual values of $X = U_Y = 1$. In other words, with the causal laws in the counterfactual world still being in place, the change from $Y = 2$ to $Y^* = 3$ needs to be explained by a change in X, U_Y, or both. In fact, there is an infinite number of such changed (X^*, U_Y^*) given by the solutions to $X^* + U_Y^* = 3$. Let us explore a few simple options:

(i) We explain the change in Y entirely by a change in U_Y: this can be achieved by keeping X^* and hence also U_X^* equal to their factual values of one and setting $U_Y^* = 2$. Then U_Z^* could remain equal to its factual value of $U_Z = -1$, leading to $Z^* := X^* + Y^* + U_Z^* = 1 + 3 - 1 = 3$ which is identical to the interventional counterfactual computed in Example 1.[6]

(ii) We explain the change in Y entirely by a change in X^*: we keep $U_Y^* = U_Y = 1$ and set $X^* = 2$. As a consequence, for the causal law $X^* := U_X^*$ to still be satisfied, we also need to adjust U_X and set $U_X^* = 2$. Again, we can keep $U_Z^* = U_Z = -1$, leading to $Z^* := X^* + Y^* + U_Z^* = 2 + 3 - 1 = 4$ which differs from the interventional counterfactual computed in Example 1!

(iii) We explain the change in Y by a change in both X^* and U_Y^*: for this we set $U_Y^* = X^* = 1.5$, adjust U_X accordingly to $U_X^* = 1.5$, keep $U_Z^* = U_Z = -1$, and obtain $Z^* := X^* + Y^* + U_Z^* = 1.5 + 3 - 1 = 3.5$, which is yet another outcome for Z^*.

Before returning to Example 2 later, let us summarise some of the points it illustrates already:

First, to explain the change in $Y^* \neq Y$ while keeping the causal laws intact, we had to change at least one of the (endogenous or exogenous) parents of Y, that is, to propagate the change upstream in the causal hierarchy—this is exactly what *backtracking* refers to (cf. Remark 1).

Second, unlike for the interventional counterfactual in Example 1, the backtracking counterfactual in Example 2 does not lead to a unique solution. Since the background variables are allowed to differ, there can be multiple ways of setting $\mathbf{U}^* \neq \mathbf{U}$ that satisfy the causal laws and agree with the counterfactual antecedent. When changing only the exogenous variable U_Y^* corresponding to the counterfactual antecedent Y^* as in case (i), we reached the same conclusion as when interpreting such change in interventional terms as in Example 1. When changing an ancestor that influences a descendant ($X^* \to Z^*$) as in cases (ii) and (iii), however, we reached different conclusions from Example 1. This highlights that backtracking counterfactuals may or may not agree with their interventional counterparts. Note also that it was not necessary to change the exogenous variable U_Z^* of the descendant variable Z of Y from its factual value U_Z, even though this would in principle also be possible.

Recall that in the factual world we have $(U_X, U_Y, U_Z) = (1, 1, -1)$. We then constructed three counterfactual worlds that are consistent with $Y^* = 2$ and the causal laws:[7]

$$(U_X^*, U_Y^*, U_Z^*) = \begin{cases} (1, 2, -1) & \Longrightarrow \quad (X^*, Z^*) = (1, 3) \quad \text{for case (i),} \\ (2, 1, -1) & \Longrightarrow \quad (X^*, Z^*) = (2, 4) \quad \text{for case (ii),} \\ (1.5, 1.5, -1) & \Longrightarrow \quad (X^*, Z^*) = (1.5, 3.5) \quad \text{for case (iii).} \end{cases} \tag{3.2}$$

In order to pick one of these worlds, or form a weighted average of their predictions, we require a notion of preference between worlds, such as a closeness or similarity measure. This will be our starting point for formalising the semantics of backtracking counterfactuals.

6. Except for the detail that, here, $Y^* = 2$ arises from a change in U_Y, instead of through an intervention $do(Y^* = 2)$ as in Example 1. However, the resulting implications for the endogenous variables are the same.
7. In fact, the (infinite) set of all valid counterfactual worlds is given by all $(U_X^*, 3 - U_X^*, U_Z^*)$ in the domain of $\mathbf{U}$, with corresponding predictions $(X^*, Z^*) = (U_X^*, 3 + U_X^* + U_Z^*)$.

3.2. Formal Semantics

In line with the probabilistic causal modelling framework, we now introduce the missing object for formal probabilistic reasoning about backtracking counterfactuals: an a priori similarity measure between factual and counterfactual worlds in the form of a *backtracking conditional* $P_B(\mathbf{U}^* \mid \mathbf{U})$. This is a collection of probability functions $P_B(\mathbf{U}^* \mid \mathbf{U} = \mathbf{u})$, one for each $\mathbf{u}$ in the domain of $\mathbf{U}$, quantifying the likelihood of each counterfactual world $\mathbf{u}^*$ given factual world $\mathbf{u}$,[8] prior to any factual (or counterfactual) observations. This provides us with a very general framework to encode various notions of cross-world similarity, depending on the context and the involved variables' domains. We remain agnostic to the exact choice of $P_B(\mathbf{U}^* \mid \mathbf{U})$ for now and defer further discussion of desiderata, properties, and specific choices thereof to § 3.3.

Together with the prior $P(\mathbf{U})$, the backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$ induces a *joint distribution over factual and counterfactual worlds* given by:

$$P_B(\mathbf{U}^*, \mathbf{U}) = P(\mathbf{U}) P_B(\mathbf{U}^* \mid \mathbf{U}) \tag{3.3}$$

We can now define the (joint) probability of backtracking counterfactuals similarly to (2.3).

Definition 2 (Probability of backtracking counterfactuals) *Let $(\mathcal{M}, P(\mathbf{U}))$ be a probabilistic causal model and $P_B(\mathbf{U}^* \mid \mathbf{U})$ a backtracking conditional, where $\mathbf{U}^*$ denotes the counterfactual version of $\mathbf{U}$ defined over the same domain. For any not necessarily disjoint subsets of variables $\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and any realisations $\mathbf{y}^*, \mathbf{z}$ thereof, the probability of backtracking counterfactuals is given by:*

$$P_B(\mathbf{y}^*, \mathbf{z}) := P_B(\mathbf{Y}^* = \mathbf{y}^*, \mathbf{Z} = \mathbf{z}) = \textstyle\sum_{(\mathbf{u}^*, \mathbf{u})} P_B(\mathbf{u}^*, \mathbf{u}) \, \mathbf{1}_{\{\mathbf{Y}^*(\mathbf{u}^*) = \mathbf{y}\}} \, \mathbf{1}_{\{\mathbf{Z}(\mathbf{u}) = \mathbf{z}\}}. \tag{3.4}$$

Any other quantities of interest can then be derived from (3.4) by standard probabilistic inference (that is, through marginalisation and conditioning). In particular, we can now answer backtracking or observational counterfactuals of the form: "given that we factually observed $\mathbf{Z}$ to be $\mathbf{z}$, what would be the probability that $\mathbf{Y}$ would be $\mathbf{y}^*$, had we *observed* $\mathbf{X}$ to be $\mathbf{x}^*$?". Provided that $P_B(\mathbf{x}^*, \mathbf{z}) > 0$, we obtain the corresponding expression $P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z})$ through the following three-step procedure, which loosely mirrors that for interventional counterfactuals given in § 2:

1. **Cross-World Abduction:** Update $P_B(\mathbf{U}^*, \mathbf{U})$ from (3.3) by the evidence $(\mathbf{x}^*, \mathbf{z},)$ to obtain the joint ("cross-world") posterior $P(\mathbf{U}^*, \mathbf{U} \mid \mathbf{x}^*, \mathbf{z})$ given by

$$P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z}) = \frac{P_B(\mathbf{u}^*, \mathbf{u})}{P_B(\mathbf{x}^*, \mathbf{z})} \, \mathbf{1}_{\{\mathbf{X}^*(\mathbf{u}^*) = \mathbf{x}^*\}} \, \mathbf{1}_{\{\mathbf{Z}(\mathbf{u}) = \mathbf{z}\}} \tag{3.5}$$

   where $P_B(\mathbf{x}^*, \mathbf{z})$ is given by (3.4).[9]
2. **Marginalisation:** Marginalise out $\mathbf{U}$ to obtain the counterfactual posterior $P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z})$:

$$P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) = \textstyle\sum_{\mathbf{u}} P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z}). \tag{3.6}$$

3. **Prediction:** Use the model $(\mathcal{M}, P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z}))$ to predict $\mathbf{Y}^*$:

$$P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z}) = \textstyle\sum_{\mathbf{u}^*} P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) \, \mathbf{1}_{\{\mathbf{Y}^*(\mathbf{u}^*) = \mathbf{y}^*\}} \tag{3.7}$$

8. Since the causal laws, and thus the causal model $\mathcal{M}$, remain unchanged for backtracking counterfactuals, we will also simply refer to instantiations of $\mathbf{U}$ as "worlds" in this context.
9. This step can also be viewed as the projection of $P_B(\mathbf{U}^*, \mathbf{U})$ onto the subspace consistent with $(\mathbf{x}^*, \mathbf{z})$.

Remark 3 *Whereas for interventional counterfactuals the antecedent is always attainable by intervention, backtracking counterfactuals may have no solution if the antecedent is incompatible with the causal laws. For example, if $Y := X$, any backtracking counterfactual with antecedent $y^* \neq x^*$ cannot be realized. Dealing with such "counterlegals" (Fisher, 2017b) appropriately requires a semantics that combines backtracking and interventions. We return to this in § 6.*

To illustrate this procedure, we now carry out the respective calculations for Example 2.

Example 2 (continued) *We now compute $P_B(X^*, Z^* \mid Y^* = 3, X = 1, Y = 2, Z = 2)$. Suppose that we use the backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U} = \mathbf{u}) = \mathcal{N}(\mathbf{u}, diag(\sigma_X^2, \sigma_Y^2, \sigma_Z^2))$ which assigns higher probability to values of $\mathbf{U}^*$ that are close to a given $\mathbf{u}$, with the relative scale of closeness given by $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$ (see § 3.3 for more details).*

1. ***Cross-World Abduction:*** *Ignoring multiplicative constants w.r.t. $(\mathbf{u}^*, \mathbf{u})$, we find*

$$P_B(\mathbf{u}^*, \mathbf{u} \mid Y^* = 3, X = 1, Y = 2, Z = 2) \propto P_B(\mathbf{u}^*, \mathbf{u})\, \mathbf{1}_{\{\mathbf{u}=(1,1,-1)\}}\, \mathbf{1}_{\{u_Y^* = 3 - u_X^*\}} \tag{3.8}$$

2. ***Marginalisation:*** *Only one non-zero term in the sum over $\mathbf{u}$'s in (3.6) remains:*

$$\begin{aligned} P_B(\mathbf{u}^* \mid Y^* = 3, X = 1, Y = 2, Z = 2) &\propto P_B(\mathbf{u}^* \mid \mathbf{u} = (1, 1, -1))\, \mathbf{1}_{\{u_Y^* = 3 - u_X^*\}} \\ &\propto \exp\left\{ -\frac{(u_X^* - 1)^2}{2\sigma_X^2} - \frac{(3 - u_X^* - 1)^2}{2\sigma_Y^2} - \frac{(u_Z^* + 1)^2}{2\sigma_Z^2} \right\} \mathbf{1}_{\{u_Y^* = 3 - u_X^*\}} \end{aligned} \tag{3.9}$$

   Completing the square for u_X^*, we finally obtain the counterfactual posterior:

$$\mathbf{U}^* \mid (Y^*{=}3, X{=}1, Y{=}2, Z{=}2) \sim \mathcal{N}_{(U_X^*, U_Z^*)} \left(\begin{pmatrix} 1 + \frac{r}{1+r} \\ -1 \end{pmatrix}, \begin{pmatrix} \frac{\sigma_X^2}{1+r} & 0 \\ 0 & \sigma_Z^2 \end{pmatrix} \right) \delta(U_Y^* {=} 3 - U_X^*) \tag{3.10}$$

   where $r = \frac{\sigma_X^2}{\sigma_Y^2}$ denotes the ratio of variances from the backtracking conditional.

3. ***Prediction:*** *Using the relation $(X^*, Z^*) = (U_X^*, 3 + U_X^* + U_Z^*)$ derived from the causal laws, we finally obtain the solution as a linear transformation of the Gaussian in (3.10):*

$$(X^*, Z^*) \mid (Y^*{=}3, X{=}1, Y{=}2, Z{=}2) \sim \mathcal{N} \left(\begin{pmatrix} 1 + \frac{r}{1+r} \\ 3 + \frac{r}{1+r} \end{pmatrix}, \begin{pmatrix} \frac{\sigma_X^2}{1+r} & \frac{\sigma_X^2}{1+r} \\ \frac{\sigma_X^2}{1+r} & \sigma_Z^2 + \frac{\sigma_X^2}{1+r} \end{pmatrix} \right) \tag{3.11}$$

This form of (3.11) is consistent with intuition: if changes in U_X^* and U_Y^* are considered equally likely (i.e., for $\sigma_X^2 = \sigma_Y^2$ in the backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$), we have $r = 1$ and case (iii) in (3.2) is the most likely (maximum a posteriori) scenario; if, on the other hand, U_X^* is considered much less likely to change (i.e., $\sigma_X^2 \ll \sigma_Y^2$ and $r \to 0$), (3.11) centers around case (i); conversely, if U_Y^* is much less likely to change (i.e., $\sigma_X^2 \gg \sigma_Y^2$ and $r \to \infty$), (3.11) centers around case (ii).

3.3. On the Choice of the Backtracking Conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$

As we have seen, $P_B(\mathbf{U}^* \mid \mathbf{U})$ plays an important role in answering backtracking counterfactuals: different choices may yield different answers. We therefore now discuss desiderata for, and relevant properties of, the backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$.

First, an intuitive desideratum is that $P_B(\mathbf{U}^* \mid \mathbf{U})$ should assign high probability to values of $\mathbf{U}^*$ that are considered close to a given $\mathbf{u}$, and small probability to those that are considered far.

Property 1 (Preference for Closeness) *We say that* $P_B(\mathbf{U}^* \mid \mathbf{U})$ *has* Preference for Closeness *if*

$$\forall \mathbf{u} : \operatorname{argmax}_{\mathbf{u}^*} P(\mathbf{u}^* \mid \mathbf{u}) = \{\mathbf{u}\}. \tag{3.12}$$

Note that we could also impose such preference for closeness w.r.t. the mean or median, instead of (or in addition to) the mode as in (3.12).

Second, to avoid a-priori asymmetries between the factual and counterfactual world, the following symmetry requirement also seems natural.

Property 2 (Symmetry) *We say that* $P_B(\mathbf{U}^* \mid \mathbf{U})$ *is* symmetric *if for any* $(\mathbf{u}^*, \mathbf{u})$*:*

$$P_B(\mathbf{u}^* \mid \mathbf{u}) = P_B(\mathbf{u} \mid \mathbf{u}^*) \tag{3.13}$$

Note that this implies and is implied by matching marginals, or priors, across worlds:

Lemma 4 (Symmetry is equivalent to matching marginals) $P_B(\mathbf{U}^* \mid \mathbf{U})$ *is symmetric if and only if the marginal of* $\mathbf{U}^*$ *induced by* $P(\mathbf{U})$ *and* $P_B(\mathbf{U}^* \mid \mathbf{U})$ *matches* $P(\mathbf{U})$*, that is:*

$$P_B(\mathbf{U}^*) := \textstyle\sum_{\mathbf{u}} P_B(\mathbf{U}^* \mid \mathbf{u}) P(\mathbf{u}) = P(\mathbf{U}). \tag{3.14}$$

Proof The result follows from (3.13), Bayes rule, and the equality of domains of $\mathbf{U}$ and $\mathbf{U}^*$. ■

Third, since it is often assumed that the exogenous variables are mutually independent (i.e, that $P(\mathbf{U})$ factorises), a similar natural property for the backtracking conditional is that each counterfactual variable U_j^* only depends on its factual counterpart U_j.

Property 3 (Decomposability) *We say that* $P_B(\mathbf{U}^* \mid \mathbf{U})$ *is* decomposable *if for any* $(\mathbf{u}^*, \mathbf{u})$*:*

$$P_B(\mathbf{u}^* \mid \mathbf{u}) = \textstyle\prod_{j=1}^m P_B(u_j^* \mid u_j). \tag{3.15}$$

In line with the philosophical notion of *closest possible worlds*, one option is to construct $P_B(\mathbf{U}^* \mid \mathbf{U})$ based on a distance function $d(\cdot, \cdot)$ defined over the shared domain of $\mathbf{U}$ and $\mathbf{U}^*$:

$$P(\mathbf{u}^* \mid \mathbf{u}) = \tfrac{1}{Z} \exp\{-d(\mathbf{u}^*, \mathbf{u})\} \tag{3.16}$$

where $Z = \sum_{\mathbf{u}^*} \exp\{-d(\mathbf{u}^*, \mathbf{u})\}$ is a normalisation constant.

Lemma 5 *Any* $P(\mathbf{u}^* \mid \mathbf{u})$ *of the form* (3.16) *satisfies Properties 1 and 2. Further, if* $d(\mathbf{u}^*, \mathbf{u})$ *can be written as* $\sum_j d(u_j^*, u_j)$*, then* $P(\mathbf{u}^* \mid \mathbf{u})$ *also satisfies Property 3.*

Proof By definition, any distance satisfies $d(\mathbf{u}^*, \mathbf{u}) \geq 0$ with equality iff $\mathbf{u}^* = \mathbf{u}$ (implying Property 1) and $d(\mathbf{u}^*, \mathbf{u}) = d(\mathbf{u}, \mathbf{u}^*)$ (implying Property 2). Property 3 follows from substituting the sum of component-wise distances into (3.16) and writing it as a product of exponentials. ■

For real valued $\mathbf{U} \in \mathbb{R}^m$, a natural choice is the (squared) Mahalanobis distance

$$d(\mathbf{u}^*, \mathbf{u}) = \tfrac{1}{2}(\mathbf{u} - \mathbf{u}^*)^\top \mathbf{\Sigma}^{-1} (\mathbf{u} - \mathbf{u}^*) \tag{3.17}$$

for some positive-definite, symmetric (covariance) matrix $\mathbf{\Sigma}$, giving rise to the multivariate *Gaussian backtracking conditional* $P(\mathbf{U}^* | \mathbf{U} = \mathbf{u}) = \mathcal{N}(\mathbf{u}, \mathbf{\Sigma})$ (for which mode, median and mean coincide). If $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1^2, ..., \sigma_m^2)$ as used in Example 2, this satisfies Properties 1, 2, and 3.

3.4. Theoretical Insights

Our first theoretical insight is that exogenous non-ancestors of factual and counterfactual observations remain unaffected, in that their posterior is equal to their prior.

Proposition 6 *Let $\mathcal{M}$ be an acyclic causal model. Suppose that $P(\mathbf{U})$ factorises and that $P_B(\mathbf{U}^* \mid \mathbf{U})$ is decomposable. For any $\mathbf{X}, \mathbf{Z} \subseteq \mathbf{V}$ with (endogenous and exogenous) ancestors $\mathbf{AN}_\mathbf{X}, \mathbf{AN}_\mathbf{Z}$:*

$$\mathbf{U}_J \subseteq \mathbf{U} \setminus (\mathbf{AN}_\mathbf{X} \cup \mathbf{AN}_\mathbf{Z}) \implies \forall(\mathbf{x}^*, \mathbf{z}) : P(\mathbf{U}^*_J, \mathbf{U}_J \mid \mathbf{x}^*, \mathbf{z}) = \textstyle\prod_{j \in J} P(U^*_j, U_j) \tag{3.18}$$

Proof The result follows from d-separation between $(\mathbf{U}^*_J, \mathbf{U}_J)$ and $(\mathbf{X}^*, \mathbf{Z})$. ■

Our second insight concerns the distinguishability of SCMs based on backtracking counterfactuals. As is well-known, identifying probabilities of *interventional* counterfactuals requires full knowledge of the laws $\mathbf{F}$, which is hard to obtain in practice (Bareinboim et al., 2022). Interestingly, probabilities of *observational* counterfactuals are not as demanding since they only depend on the solution or "reduced form" $\mathbf{V}(\mathbf{U})$ (see, e.g., Schölkopf and von Kügelgen, 2022, § 10 for details).

Proposition 7 *Let $(\mathcal{M}_1, P(\mathbf{U}))$ and $(\mathcal{M}_2, P(\mathbf{U}))$ be causal models over the same variables $\mathbf{V}$ and $\mathbf{U}$ whose laws $\mathbf{F}_1$ and $\mathbf{F}_2$ have identical solutions $\mathbf{V}_1(\mathbf{u}) = \mathbf{V}_2(\mathbf{u})$ for all values $\mathbf{u}$. Then for any choice of $P_B(\mathbf{U}^* \mid \mathbf{U})$, both models will imply identical $P_B(\mathbf{y}^*, \mathbf{z})$ for all choices $\mathbf{y}^*$ and $\mathbf{z}$.*

Proof This is an immediate consequence of (3.4) only involving (3.3) and the solution function. ■

In particular, this result implies that causal structure is not discernible purely based on backtracking counterfactuals, as demonstrated by the following example.

Example 3 *Consider the following three causal models over $\mathbf{V} = \{X, Y\}$ and $\mathbf{U} = \{U\}$ with laws $\mathbf{F}_{X \to Y} = \{X := U, Y := X\}$, $\mathbf{F}_{Y \to X} = \{Y := U, X := Y\}$, $\mathbf{F}_{X \leftrightarrow Y} = \{X := U, Y := U\}$, and some shared $P(\mathbf{U})$. Then all share the same reduced form $(X, Y) = (U, U)$ and hence the same backtracking counterfactuals, despite differing in their causal diagrams.*

4. Related Work

Most closely related to our formalisation is the Extended Structural Model proposed by Lucas and Kemp (2015, Eq. 1) to model recent empirical findings on the context-dependency of humans' use of backtracking. It employs a decomposable backtracking conditional for Boolean variables,

$$P_B(U^*_j \mid U_j) = s\delta(U_j) + (1 - s)P(U^*_j), \tag{4.1}$$

where the "stability" hyperparameter $s \in [0, 1]$ interpolates between independent worlds ($s = 0$) and perfectly shared exogenous variables ($s = 1$). This account is thus restrictive in that it forces the counterfactual variables to either copy the factual ones or to completely ignore them.

Other existing accounts are less formal: they mostly consider non-probabilistic Boolean conditionals of the form "Given world w, if $\mathbf{x}^*$ were true then $\mathbf{y}^*$ would be true" and typically involve minimising the number of (endogenous and) exogenous variables that change across worlds. Hiddleston (2005)'s semantics focuses on minimising the number of exogenous non-descendants of the antecedent that change values, which corresponds to using a variant of the Manhattan distance in (3.16). Interestingly, Pearl's interventional semantics (§ 2) can be interpreted in a similar vein, instead minimising the number of intervened-upon variables. In fact, he uses this formulation to argue that this is essentially equivalent to Lewis's possible world semantics (Pearl, 2009, p.241).

Fisher (2017b) combines both approaches by applying Hiddleston's condition if there exists a backtracking solution and resorting to a minimal number of interventions only if necessary. Lee (2017) also considers both semantics, but does not combine them into a single one.

Empirical research by cognitive scientists confirms that depending on the context, humans interpret counterfactuals as either backtracking or non-backtracking: "participants are more likely to backtrack when explicitly asked to consider a counterfactual's *causes*. However, when directly asked about the *effects* of a counterfactual state, most people don't backtrack" (Gerstenberg et al., 2013). Moreover, people use the broader context, exact wording, and level of determinism of the involved mechanism "to infer how the antecedent is most likely to have come about" and use this to decide whether to backtrack or not (Rips and Edwards, 2013). It has also been suggested that backtracking is preferred if and only if doing so makes a counterfactual claim true (Han et al., 2014).

5. Connections and Applications to Explainable AI (XAI)

We believe that backtracking counterfactuals may hold great promise for XAI, which is concerned with offering explanations for the decisions of a machine learning model implementing a function $Y = f(\mathbf{X})$, where $\mathbf{X}$ are *input features* and Y is the *output or label*. Given that we observe some input $\mathbf{x}$ and corresponding output y, the general aim is to find a feature subset $\mathbf{Z} \subseteq \mathbf{X}$ that "explains" $y = f(\mathbf{x})$. A particularly promising approach has been to look for so-called nearest *counterfactual explanations*, meaning that we look for both $\mathbf{z}$ *and* $\mathbf{z}^*$ such that changing $\mathbf{Z}$ from $\mathbf{z}$ to $\mathbf{z}^*$ would have resulted in $y^* \neq y$, and $\mathbf{z}$ and $\mathbf{z}^*$ are close according to some distance function (Wachter et al., 2017). Although the general idea is rather intuitive, there has been much discussion on how exactly to implement it. The core problem lies in the choice of values $\mathbf{w}^*$ for the remaining features $\mathbf{W}^* = \mathbf{X}^* \setminus \mathbf{Z}^*$ that should accompany $\mathbf{z}^*$ so that $f(\mathbf{z}^*, \mathbf{w}^*) = y^*$.

This discussion between AI-researchers can in fact be seen as the operational analogue of the philosophical discussion on the semantics of counterfactuals that we started out with in § 1, and which is far from resolved. Hence, we are faced with the same dilemma as before: to backtrack, or not to backtrack? Interestingly, the proposal of Wachter et al. (2017) does neither, keeping all other variables fixed at their factual values $\mathbf{w}^* = \mathbf{w}$. This proposal has recently come under criticism for its failure to take into account causal dependencies. On the one hand, Beckers (2022) criticises it for failing to follow Pearl (and the entire causal modelling tradition) in maintaining an *interventionist* reading of counterfactuals. On the other, Mahajan et al. (2019) criticise it for ignoring violations of the causal laws, and offer an alternative proposal that—without making it explicit—can be viewed as building on a kind of *backtracking* semantics instead, though one that involves distances between endogenous variables and requires complete knowledge of the causal laws.

When counterfactual explanations are offered in the context of *algorithmic recourse* (Ustun et al., 2019), meaning that one is looking for actionable changes to the input features that result in a more favourable outcome, the first criticism is undoubtedly on the right track (Karimi et al., 2020, 2021). For example, if an agent is told to increase their income by $\$5,000$ in order to be granted a loan, then this explanation only succeeds in correctly predicting the decision if it has correctly taken into account the forward-tracking consequences of this change on any features that are downstream of income, such as savings. If these changes are implemented externally (thereby overriding the observational distribution), they should indeed be modelled as interventions.

Yet, in most cases the use of interventional counterfactuals is unrealistic, since they cannot be identified solely from the observational distribution that ML-methods are trained on. Moreover, it is not at all clear that interventional counterfactuals are the right approach in the context of expla-

nations that aim to *contest*, or to simply *understand* or *diagnose*, the outcome that was reached. In line with the second criticism (and the recent trend in the philosophy literature), we claim that backtracking counterfactuals deserve attention as a promising alternative for these contexts, and we have here developed the tools to do so. Importantly, as Proposition 7 suggests, backtracking counterfactuals should be easier to identify, as they depend on the causal dependencies between endogenous variables only through the reduced form. By leaving the causal laws intact, they also remain closer to the observational distribution, or data manifold, which is a desirable constraint for counterfactual explanations (Wexler et al., 2019; Joshi et al., 2019; Poyiadzi et al., 2020; Sharma et al., 2020).

An implementation of this idea could look roughly as follows. Let $(\mathcal{M}, P(\mathbf{U}))$ be a probabilistic causal model with endogenous variables $\mathbf{V} = \mathbf{X} \cup \{Y\}$, $P(\mathbf{U})$ having full support, and laws such that $Y = f(\mathbf{X})$ holds with probability one. For a particular choice of $P_B(\mathbf{U}^* \mid \mathbf{U})$, we then say that "$\mathbf{x}$ rather than $\mathbf{x}^*$ explains why $f(\mathbf{x}) = y$ rather than $y^* \neq y$" if such a change to y^* would be most likely to have come about through $\mathbf{x}^*$, that is, if $\mathbf{x}^* \in \operatorname{argmax}_{\mathbf{x}^*} P_B(\mathbf{x}^* \mid y^*, \mathbf{x}, y)$. Note that this is exactly the *diagnostic* kind of reasoning that only backtracking allows, thus establishing a clear link between counterfactual explanations and maximum a posteriori backtracking counterfactuals. In practice, interpretable explanations should be concise, highlighting just a few features for which relatively small changes make a big difference. To implement this, we could set a probability threshold α and a maximal number k of "explanatory" features, and then look for an optimal $\mathbf{Z}^* = \mathbf{z}^*$ satisfying $|\mathbf{Z}| \leq k$ and $P_B(\mathbf{z}^* \mid y^*, \mathbf{x}, y) > \alpha$. Lastly, note that a variant of the original proposal can be retrieved within our proposed backtracking framework by considering $\operatorname{argmax}_{\mathbf{z}^*} P_B(\mathbf{z}^* \mid \mathbf{w}^* = \mathbf{w}, y^*, \mathbf{x}, y)$ for all choices $\mathbf{Z}$ and $\mathbf{z}^* \neq \mathbf{z}$ (where as before $\mathbf{X} = \mathbf{W} \cup \mathbf{Z}$).

6. Conclusion and Future Work

We have here presented the first formal account of backtracking counterfactuals within probabilistic causal models (§ 3). Doing so in full generality required the introduction of a new object, the backtracking conditional, which quantifies a notion of similarity between worlds (§ 3.2). This involves a design choice that can take into account the specific context and the modeller's background assumptions. In fact, many previous accounts of backtracking can be seen as specific choices of such a conditional (§ 4). We have laid out some sensible desiderata and means to achieve them in § 3.3.

As stressed throughout, we do not view backtracking as a replacement of the interventionist account (§ 2) but rather as complementary to it. Proposition 7 and Example 3 make clear that backtracking counterfactuals do not allow for discerning causal structure and are therefore of limited use for interventional reasoning. At the same time, they are helpful for the kind of diagnostic reasoning that occurs in settings where interventions are either inconsistent with human judgement (§ 4), require knowledge that is unavailable (§ 5), or are perhaps not even meaningful. The latter shows up in the context of causal fairness analysis (Kusner et al., 2017; Kilbertus et al., 2017; Plecko and Bareinboim, 2022), as it often focuses on such attributes (like race or gender), leading some to dispute their status as valid causes (Holland, 2008; Hu and Kohler-Hausmann, 2020).

By endowing backtracking with a formal semantics within the same general SCM framework also used for interventional counterfactuals, we hope that the present work may pave the way for a unified framework of backtracking *and* non-backtracking counterfactual reasoning. The general idea could be to allow modifying both the causal laws and the background conditions, and to weigh off the changes to both of them, with a preference towards solutions with minimal joint changes. This would not only imply the property of being guaranteed a solution (recall Remark 3), but may also allow for more accurate models of human counterfactual reasoning.

References

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 230–237. AAAI Press / The MIT Press, 1994.

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*, pages 507–556. 2022.

Sander Beckers. Causal explanations and XAI. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177, pages 90–109. PMLR, 2022.

Philip Dawid. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424, 2000.

Cian Dorr. Against counterfactual miracles. *The Philosophical Review*, 125(2):241–286, 2016.

Michael Esfeld. Super-humeanism and free will. *Synthese*, 198(7):6245–6258, 2021.

Tyrus Fisher. Causal counterfactuals are not interventionist counterfactuals. *Synthese*, 194(12): 4935–4957, 2017a.

Tyrus Fisher. Counterlegal dependence and causation's arrows: Causal models for backtrackers and counterlegals. *Synthese*, 194(12):4983–5003, 2017b.

Tobias Gerstenberg, Christos Bechlivanidis, and David A Lagnado. Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

Jung-Ho Han, William Jimenez-Leal, and Steve Sloman. Conditions for backtracking with counterfactual conditionals. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

Eric Hiddleston. A causal theory of counterfactuals. *Noûs*, 39(4):632–657, 2005.

Paul W Holland. Causation and race. *White logic, white methods: Racism and methodology*, pages 93–109, 2008.

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513, 2020.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 33:265–277, 2020.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Kok Yong Lee. Hiddleston's causal modeling semantics and the distinction between forward-tracking and backtracking counterfactuals. *Studies in Logic*, 10(1), 2017.

David Lewis. *Counterfactuals*. Oxford: Blackwell Publishers and Cambridge, MA: Harvard University Press, 1973.

David Lewis. Counterfactual dependence and time's arrow. *Noûs*, pages 455–476, 1979.

Barry Loewer. Counterfactuals and the second law. In Huw Price and Richard Corry, editors, *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press, 2007.

Barry Loewer. The consequence argument meets the mentaculus, 2020. Working papers, Rutgers University.

Christopher G Lucas and Charles Kemp. An improved probabilistic account of counterfactual reasoning. *Psychological review*, 122(4):700, 2015.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

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

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Lance J Rips. Two causal theories of counterfactual conditionals. *Cognitive science*, 34(2):175–221, 2010.

Lance J Rips and Brian J Edwards. Inference and explanation in counterfactual reasoning. *Cognitive Science*, 37(6):1107–1135, 2013.

Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning. *arXiv preprint arXiv:2204.00607*, 2022.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

James Woodward. *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press, 2021.