Backtracking Counterfactuals

XAI Seminar Series at Imperial College London

Julius von Kügelgen, Abdirisak Mohamed, Sander Beckers

Max Planck Institute for Intelligent Systems & University of Cambridge

3 May 2023

Outline

Counterfactuals are Ubiquitous

Why care about counterfactuals?

- Essential for defining causation: *"if the first object had not been, the second never had existed"* (Hume, 1748)
- Explanations for *why* something happened
  (*Why was my loan application rejected?*)
- Planning and reasoning about hypotheticals
  (*Would I have got the loan, had I had 5k more in savings?*)
- Assigning credit and blame
  (*Was it the aspirin that cured my headache?*)

Making Sense of Counterfactuals

Counterfactuals: *What would the world look like* ($\mathbf{V}^*$) *if some events* ($\mathbf{V}$) *which did occur had, in fact, not occurred?*

In a deterministic world, everything that happens is determined by

- the laws of nature $\mathbf{F}$; and
- the initial / background conditions $\mathbf{u}$.

Dilemma: either

(A) the laws $\mathbf{F}$ would have had to be violated; or

(B) the background conditions $\mathbf{u}$ would have had to be different.

$\rightarrow$ different counterfactual semantics

Interventional vs Backtracking Semantics

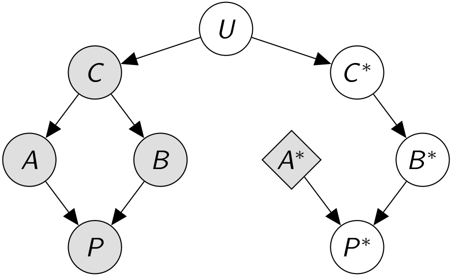
	(A) **Interventional**	(B) **Backtracking**
Shared	initial state $\mathbf{u}$	laws $\mathbf{F}$
Changing	laws $\mathbf{F} \to \mathbf{F}^*$	initial state $\mathbf{u} \to \mathbf{u}^*$
Illustration	$\begin{tikzcd} & \mathbf{u} \arrow[dl, "\mathbf{F}"'] \arrow[dr, "\mathbf{F}^*"] & \\ \mathbf{V} & & \mathbf{V}^* \end{tikzcd}$	$\begin{tikzcd} \mathbf{u} \arrow[r, leftrightarrow, dashed] \arrow[d, "\mathbf{F}"'] & \mathbf{u}^* \arrow[d, "\mathbf{F}"'] \\ \mathbf{V} & \mathbf{V}^* \end{tikzcd}$
Formalisation	Lewis (1979, 1973): small miracles & possible worlds; Pearl (2009): structural equations & minisurgeries	**This Work**

Firing Squad Example (Pearl, 2009, § 7.1.2)

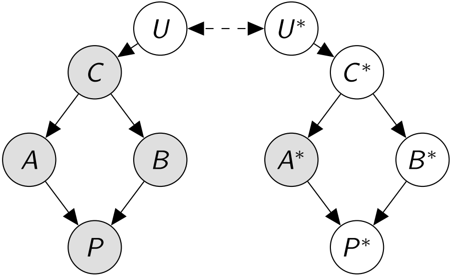
The captain C of two riflemen A and B is waiting for a court order U on whether a prisoner P should be executed (all Boolean).

$$C := U, \qquad A := C, \qquad B := C, \qquad P := A \vee B$$

Suppose $C = A = B = P = 1$. **Q:** *What if rifleman A had not shot?*



Interventional: P^* Dead



Backtracking: P^* Alive

Why did A not shoot? — Disobedience (left) vs no court order (right)

Outline

1. Motivation & Overview
2. Background: SCMs & Interventional Counterfactuals
3. Backtracking Counterfactuals
4. Connections to XAI
5. Discussion

Structural Causal Models (SCMs; Pearl, 2009)

A causal model is a triple $\mathcal{M} = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ where:

- $\mathbf{U}$ is a set $\{U_1, ..., U_m\}$ of exogenous (background) variables
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of endogenous (observable) variables
- $\mathbf{F}$ is a set $\{f_1, f_2, \ldots, f_n\}$ of structural equations, or causal laws

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

where $\mathbf{U}_i \subseteq \mathbf{U}$ and $\mathbf{PA}_i \subseteq \mathbf{V} \setminus \{V_i\}$ s.t. $\mathbf{F}$ has a unique solution $\mathbf{V}(\mathbf{u})$.[1]

A causal world w is a pair $(\mathcal{M}, \mathbf{u})$

A probabilistic causal model is a distribution over causal worlds $(\mathcal{M}, P(\mathbf{U}))$

[1] Ensured, e.g., in acyclic ("recursive") systems.

Interventional Counterfactuals in SCMs

The potential response $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ of $\mathbf{Y}$ under action $do(\mathbf{X} = \mathbf{x})$ in world $w = (\mathcal{M}, \mathbf{u})$ is the solution for $\mathbf{Y}$ of the modified set of equations

$$\mathbf{F}_{\mathbf{x}} = \{f_i : V_i \notin \mathbf{X}\} \cup \{\mathbf{X} := \mathbf{x}\}.$$

"$\mathbf{Y}$ *would be* $\mathbf{y}$ *(in situation* $\mathbf{u}$*), had* $\mathbf{X}$ *been* $\mathbf{x}$" is interpreted as $\mathbf{Y}_{\mathbf{x}}(\mathbf{u}) = \mathbf{y}$.
(here, "*had* $\mathbf{X}$ *been* $\mathbf{x}$" is called the counterfactual antecedent)

The probability of counterfactuals for any $\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{W} \subseteq \mathbf{V}$ is given by

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

Twin Network Representation & Example

Observation: $(X, Y, Z) = (1, 2, 2)$.
Question: What if Y had been 3?

$$\boxed{\begin{aligned} X &:= U_X, & (1)\\ Y &:= X + U_Y, & (2)\\ Z &:= X + Y + U_Z, & (3)\end{aligned}}$$

1. Abduction: from Eqs. (1)–(3) infer

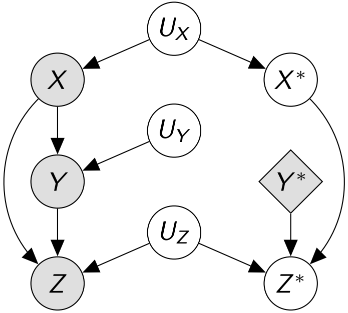
$$(U_X, U_Y, U_Z) = (1, 1, -1)$$

2. Action: replace Eq. (2) by

$$Y := 3$$

3. Prediction: use modified SCM,

$$Z := X + Y + U_Z = 1 + 3 - 1 = 3$$



Summary of Interventionist Semantics

"[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 $\mathbf{X} = \mathbf{x}$"

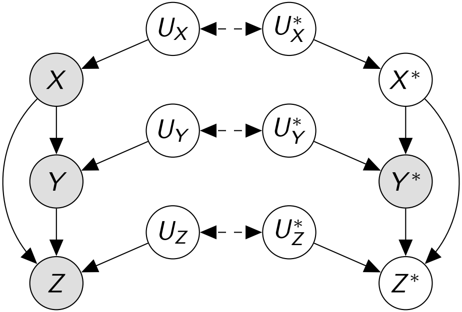
—Pearl (2009, p.205)

Outline

Intuition and Main Idea

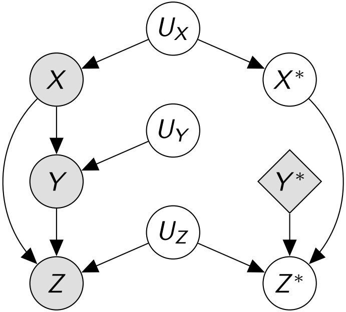
The causal laws, not the background conditions, are shared across worlds

$\rightarrow$ backtrack all changes to changes in exogenous variables



Backtracking

$$Y^* := X^* + U_Y^* = 3$$



Interventional

$$Y^* := 3$$

Non-Uniqueness of Backtracking

Many worlds $(\mathcal{M}, \mathbf{u}^*)$ consistent with counterfactual antecedent $(Y^* = 3)$:

$$(U_X^*, U_Y^*, U_Z^*) = \begin{cases} (1, 2, -1) & \implies \quad (X^*, Z^*) = (1, 3) \\ (2, 1, -1) & \implies \quad (X^*, Z^*) = (2, 4) \\ (1.5, 1.5, -1) & \implies \quad (X^*, Z^*) = (1.5, 3.5) \\ & \dots \\ (U_X^*, 3 - U_X^*, U_Z^*) & \implies \quad (X^*, Z^*) = (U_X^*, 3 + U_X^* + U_Z^*) \end{cases}$$

Q: How to pick one or form a weighted average of their predictions?

$\rightarrow$ need a similarity measure across worlds:
the backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$.

Probabilistic Backtracking

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

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

The joint probability of backtracking counterfactuals is given by:

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

for any (not necessarily disjoint) $\mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ and realizations $\mathbf{y}^*, \mathbf{z}$ thereof.[2]

[2] Other quantities are then derived via marginalisation & conditioning.

3-Step Procedure for Backtracking

Q: 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 $\mathbf{x}^*$?[3]

$$P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z})$$

1. Cross-World Abduction: Update $P_B(\mathbf{U}^*, \mathbf{U})$ by the evidence $(\mathbf{x}^*, \mathbf{z},)$ to obtain the joint ("cross-world") posterior $P(\mathbf{U}^*, \mathbf{U} \mid \mathbf{x}^*, \mathbf{z})$
2. Marginalisation: Marginalise out $\mathbf{U}$ to obtain the counterfactual posterior $P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) = \sum_{\mathbf{u}} P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z})$.
3. Prediction: Use the model $(\mathcal{M}, P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z}))$ to predict $\mathbf{Y}^*$:

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

[3] Provided that $P_B(\mathbf{x}^*, \mathbf{z}) > 0$

3-Step Procedure for Backtracking

Q: 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 $\mathbf{x}^*$?[3]

$$P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z})$$

1. Cross-World Abduction: Update $P_B(\mathbf{U}^*, \mathbf{U})$ by the evidence $(\mathbf{x}^*, \mathbf{z},)$ to obtain the joint ("cross-world") posterior $P(\mathbf{U}^*, \mathbf{U} \mid \mathbf{x}^*, \mathbf{z})$
2. Marginalisation: Marginalise out $\mathbf{U}$ to obtain the counterfactual posterior $P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) = \sum_{\mathbf{u}} P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z})$.
3. Prediction: Use the model $(\mathcal{M}, P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z}))$ to predict $\mathbf{Y}^*$:

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

[3] Provided that $P_B(\mathbf{x}^*, \mathbf{z}) > 0$

3-Step Procedure for Backtracking

Q: 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 $\mathbf{x}^*$?[3]

$$P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z})$$

1. Cross-World Abduction: Update $P_B(\mathbf{U}^*, \mathbf{U})$ by the evidence $(\mathbf{x}^*, \mathbf{z},)$ to obtain the joint ("cross-world") posterior $P(\mathbf{U}^*, \mathbf{U} \mid \mathbf{x}^*, \mathbf{z})$
2. Marginalisation: Marginalise out $\mathbf{U}$ to obtain the counterfactual posterior $P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) = \sum_{\mathbf{u}} P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z})$.
3. Prediction: Use the model $(\mathcal{M}, P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z}))$ to predict $\mathbf{Y}^*$:

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

[3] Provided that $P_B(\mathbf{x}^*, \mathbf{z}) > 0$

3-Step Procedure for Backtracking

Q: 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 $\mathbf{x}^*$?[3]

$$P_B(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{z})$$

1. Cross-World Abduction: Update $P_B(\mathbf{U}^*, \mathbf{U})$ by the evidence $(\mathbf{x}^*, \mathbf{z},)$ to obtain the joint ("cross-world") posterior $P(\mathbf{U}^*, \mathbf{U} \mid \mathbf{x}^*, \mathbf{z})$
2. Marginalisation: Marginalise out $\mathbf{U}$ to obtain the counterfactual posterior $P_B(\mathbf{u}^* \mid \mathbf{x}^*, \mathbf{z}) = \sum_{\mathbf{u}} P_B(\mathbf{u}^*, \mathbf{u} \mid \mathbf{x}^*, \mathbf{z})$.
3. Prediction: Use the model $(\mathcal{M}, P_B(\mathbf{U}^* \mid \mathbf{x}^*, \mathbf{z}))$ to predict $\mathbf{Y}^*$:

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

[3] Provided that $P_B(\mathbf{x}^*, \mathbf{z}) > 0$

Choice of Backtracking Conditional

Desiderata/Properties:

1. Preference for Closeness: $\forall \mathbf{u} : \arg\max_{\mathbf{u}^*} P_B(\mathbf{u}^* \mid \mathbf{u}) = \{\mathbf{u}\}$.
2. Symmetry:[4] $\forall (\mathbf{u}^*, \mathbf{u}) : P_B(\mathbf{u}^* \mid \mathbf{u}) = P_B(\mathbf{u} \mid \mathbf{u}^*)$
3. Decomposability: $P_B(\mathbf{u}^* \mid \mathbf{u}) = \prod_{j=1}^m P_B(u_j^* \mid u_j)$.

Example

Using some distance function $d(\cdot, \cdot)$ over $\mathcal{U} \times \mathcal{U}$,

$$P_B(\mathbf{u}^* \mid \mathbf{u}) = \frac{1}{Z} \exp\{-d(\mathbf{u}^*, \mathbf{u})\}$$

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

$\rightarrow$ connection to *distance-based counterfactual explanations*

[4] equivalently, matching marginals: $P_B(\mathbf{U}^*) := \sum_{\mathbf{u}} P_B(\mathbf{U}^* \mid \mathbf{u}) P(\mathbf{u}) = P(\mathbf{U})$

Choice of Backtracking Conditional

Desiderata/Properties:

1. Preference for Closeness: $\forall \mathbf{u} : \arg\max_{\mathbf{u}^*} P_B(\mathbf{u}^* \mid \mathbf{u}) = \{\mathbf{u}\}$.
2. Symmetry:[4] $\forall (\mathbf{u}^*, \mathbf{u}) : P_B(\mathbf{u}^* \mid \mathbf{u}) = P_B(\mathbf{u} \mid \mathbf{u}^*)$
3. Decomposability: $P_B(\mathbf{u}^* \mid \mathbf{u}) = \prod_{j=1}^{m} P_B(u_j^* \mid u_j)$.

Example

Using some distance function $d(\cdot, \cdot)$ over $\mathcal{U} \times \mathcal{U}$,

$$P_B(\mathbf{u}^* \mid \mathbf{u}) = \frac{1}{Z} \exp\{-d(\mathbf{u}^*, \mathbf{u})\}$$

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

$\rightarrow$ connection to *distance-based counterfactual explanations*

[4] equivalently, matching marginals: $P_B(\mathbf{U}^*) := \sum_{\mathbf{u}} P_B(\mathbf{U}^* \mid \mathbf{u}) P(\mathbf{u}) = P(\mathbf{U})$

Theoretical Insights

Proposition (Informal)

Exogenous non-ancestors of factual and counterfactual observations remain unaffected: their posterior is equal to their prior.

Proposition (Informal)

Backtracking counterfactuals only depend on the reduced form/solution function (since the causal laws are kept fixed): different SCMs with the same $\mathbf{V}(\mathbf{u})$, agree on all backtracking counterfactuals.

Corollary (Informal)

Backtracking counterfactuals cannot discern causal structure.[a]

[a] E.g., $X := U$, $Y := X$ ($X \to Y$) vs $X := U =: Y$ ($X \leftarrow U \to Y$) have same solution.

Theoretical Insights

Proposition (Informal)

Exogenous non-ancestors of factual and counterfactual observations remain unaffected: their posterior is equal to their prior.

Proposition (Informal)

Backtracking counterfactuals only depend on the reduced form/solution function (since the causal laws are kept fixed): different SCMs with the same $\mathbf{V}(\mathbf{u})$, agree on all backtracking counterfactuals.

Corollary (Informal)

Backtracking counterfactuals cannot discern causal structure.[a]

[a] E.g., $X := U$, $Y := X$ $(X \to Y)$ vs $X := U =: Y$ $(X \leftarrow U \to Y)$ have same solution.

Theoretical Insights

Proposition (Informal)

Exogenous non-ancestors of factual and counterfactual observations remain unaffected: their posterior is equal to their prior.

Proposition (Informal)

Backtracking counterfactuals only depend on the reduced form/solution function (since the causal laws are kept fixed): different SCMs with the same $\mathbf{V}(\mathbf{u})$, agree on all backtracking counterfactuals.

Corollary (Informal)

Backtracking counterfactuals cannot discern causal structure.[a]

[a] E.g., $X := U$, $Y := X$ ($X \to Y$) vs $X := U =: Y$ ($X \leftarrow U \to Y$) have same solution.

Outline

Counterfactual Explanations in AI

Setting: model $Y = f(\mathbf{X})$ with input features $\mathbf{X}$ and targets/labels Y.

Goal: find (sparse) feature subset $\mathbf{Z} \subseteq \mathbf{X}$ that "explains" a given $y = f(\mathbf{x})$.

Nearest counterfactual explanations: look for $\mathbf{Z} \subseteq \mathbf{X}$ and $\mathbf{z}^*$ s.t. changing $\mathbf{z} \to \mathbf{z}^*$ results in $y^* \neq y$ and $d(\mathbf{z}, \mathbf{z}^*)$ is small (Wachter et al., 2017).

Key question: how to treat the remaining features $\mathbf{W} = \mathbf{X} \setminus \mathbf{Z}$? That is, how to choose the corresponding value $\mathbf{w}^*$ such that $f(\mathbf{z}^*, \mathbf{w}^*) = y^*$?

Analogous to philosophical debate about counterfactual semantics: *To backtrack or not to backtrack?*

Counterfactual Explanations in AI

Setting: model $Y = f(\mathbf{X})$ with input features $\mathbf{X}$ and targets/labels Y.

Goal: find (sparse) feature subset $\mathbf{Z} \subseteq \mathbf{X}$ that "explains" a given $y = f(\mathbf{x})$.

Nearest counterfactual explanations: look for $\mathbf{Z} \subseteq \mathbf{X}$ and $\mathbf{z}^*$ s.t. changing $\mathbf{z} \to \mathbf{z}^*$ results in $y^* \neq y$ and $d(\mathbf{z}, \mathbf{z}^*)$ is small (Wachter et al., 2017).

Key question: how to treat the remaining features $\mathbf{W} = \mathbf{X} \setminus \mathbf{Z}$? That is, how to choose the corresponding value $\mathbf{w}^*$ such that $f(\mathbf{z}^*, \mathbf{w}^*) = y^*$?

Analogous to philosophical debate about counterfactual semantics: *To backtrack or not to backtrack?*

Counterfactual Explanations in AI

Setting: model $Y = f(\mathbf{X})$ with input features $\mathbf{X}$ and targets/labels Y.

Goal: find (sparse) feature subset $\mathbf{Z} \subseteq \mathbf{X}$ that "explains" a given $y = f(\mathbf{x})$.

Nearest counterfactual explanations: look for $\mathbf{Z} \subseteq \mathbf{X}$ and $\mathbf{z}^*$ s.t. changing $\mathbf{z} \to \mathbf{z}^*$ results in $y^* \neq y$ and $d(\mathbf{z}, \mathbf{z}^*)$ is small (Wachter et al., 2017).

Key question: how to treat the remaining features $\mathbf{W} = \mathbf{X} \setminus \mathbf{Z}$? That is, how to choose the corresponding value $\mathbf{w}^*$ such that $f(\mathbf{z}^*, \mathbf{w}^*) = y^*$?

Analogous to philosophical debate about counterfactual semantics:
To backtrack or not to backtrack?

To Backtrack or Not To Backtrack?

Neither: keep other features fixed, $\mathbf{w}^* = \mathbf{w}$ (Wachter et al., 2017).

– implicitly assuming independent features

Interventional: forward-track changes to downstream (descendant) features (Beckers, 2022; Karimi* et al., 2022).

+ appropriate, e.g., for algorithmic recourse (Ustun et al., 2019)
– requires access to full causal model
– may not be best to contest or diagnose the outcome that was reached

Backtracking: avoid violations of the causal laws (Mahajan et al., 2019).

+ explanations remain on (observational) data manifold (Joshi et al., 2019; Poyiadzi et al., 2020; Sharma et al., 2020; Wexler et al., 2019).

Backtracking Counterfactuals for XAI

Given:

- a probabilistic causal model $(\mathcal{M}, P(\mathbf{U}))$ over variables $\mathbf{X} \cup \{Y\}$ with laws such that $Y = f(\mathbf{X})$;
- a backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$, e.g., distance-based.

Then "$\mathbf{x}$ rather than $\mathbf{x}^*$ explains why $f(\mathbf{x}) = y$ rather than $y^* \neq y$" if such a change would be most likely to have come about through $\mathbf{x}^*$,

$$\mathbf{x}^* \in \arg\max_{\mathbf{x}^*} P_B(\mathbf{x}^* \mid y^*, \mathbf{x}, y).$$

Nearest CEs = maximum a-posteriori backtracking counterfactuals

Sparse CEs: $\arg\max_{\mathbf{z}^*} P_B(\mathbf{z}^* \mid y^*, \mathbf{x}, y)$ subject to $|\mathbf{Z}| \leq k, \mathbf{z}^* \neq \mathbf{z}$.

Original proposal: $\arg\max_{\mathbf{z}^*} P_B(\mathbf{z}^* \mid \mathbf{W}^* = \mathbf{w}, y^*, \mathbf{x}, y)$ where $\mathbf{W} = \mathbf{X} \setminus \mathbf{Z}$.

Backtracking Counterfactuals for XAI

Given:

- a probabilistic causal model $(\mathcal{M}, P(\mathbf{U}))$ over variables $\mathbf{X} \cup \{Y\}$ with laws such that $Y = f(\mathbf{X})$;
- a backtracking conditional $P_B(\mathbf{U}^* \mid \mathbf{U})$, e.g., distance-based.

Then "$\mathbf{x}$ rather than $\mathbf{x}^*$ explains why $f(\mathbf{x}) = y$ rather than $y^* \neq y$" if such a change would be most likely to have come about through $\mathbf{x}^*$,

$$\mathbf{x}^* \in \arg\max_{\mathbf{x}^*} P_B(\mathbf{x}^* \mid y^*, \mathbf{x}, y).$$

Nearest CEs = maximum a-posteriori backtracking counterfactuals

Sparse CEs: $\arg\max_{\mathbf{z}^*} P_B(\mathbf{z}^* \mid y^*, \mathbf{x}, y)$ subject to $|\mathbf{Z}| \leq k, \mathbf{z}^* \neq \mathbf{z}$.

Original proposal: $\arg\max_{\mathbf{z}^*} P_B(\mathbf{z}^* \mid \mathbf{W}^* = \mathbf{w}, y^*, \mathbf{x}, y)$ where $\mathbf{W} = \mathbf{X} \setminus \mathbf{Z}$.

Backtracking and Root Cause Analysis

Root cause analysis of outliers: explain an extreme value $Y = y$ (Budhathoki et al., 2022)

Main idea: exogenous (root) nodes $\mathbf{U}$ ultimately explain why $Y = y$

Approach: keep causal laws intact and vary each U_i according to some counterfactual distribution, keeping $\mathbf{U}_{-i}$ fixed, to quantify contributions,

$$P_B(\tau(y^*) \geq \tau(y) \mid \mathbf{U}^*_{-i} = \mathbf{u}_{-i}, \mathbf{U} = \mathbf{u}).$$

$\rightarrow$ a form of backtracking in disguise!

Outline

Related Work

Philosophy (Dorr, 2016; Esfeld, 2021; Fisher, 2017a,b; Hiddleston, 2005; Lee, 2017; Loewer, 2020; Woodward, 2021):

- logic-based semantics for Boolean variables
- minimise number of exogenous non-descendants that change

Cognitive science (Gerstenberg et al., 2013; Han et al., 2014; Lucas and Kemp, 2015; Rips, 2010):

- context & exact wording used to infer how antecedent has come about
- backtracking when diagnostically reasoning about causes of effects

History (Reiss, 2009; Tetlock and Belkin, 1996):

- minimal rewrite rule for historical counterfactuals
- typically interpreted in backtracking sense

Future Work and Concluding Thoughts

Future Work:

- Backtracking for causal fairness analysis
- Unified framework for backtracking and interventional counterfactuals

"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)

References I

[1] Sander Beckers. "Causal Explanations and XAI". In: *Proceedings of the First Conference on Causal Learning and Reasoning*. Vol. 177. PMLR, 2022, pp. 90–109.

[2] Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. "Causal structure-based root cause analysis of outliers". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2357–2369.

[3] Cian Dorr. "Against counterfactual miracles". In: *The Philosophical Review* 125.2 (2016), pp. 241–286.

[4] Michael Esfeld. "Super-Humeanism and free will". In: *Synthese* 198.7 (2021), pp. 6245–6258.

[5] Tyrus Fisher. "Causal counterfactuals are not interventionist counterfactuals". In: *Synthese* 194.12 (2017), pp. 4935–4957.

[6] Tyrus Fisher. "Counterlegal dependence and causation's arrows: Causal models for backtrackers and counterlegals". In: *Synthese* 194.12 (2017), pp. 4983–5003.

[7] 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*. Vol. 35. 2013.

[8] 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*. Vol. 36. 2014.

References II

[9] Eric Hiddleston. “A causal theory of counterfactuals”. In: *Noûs* 39.4 (2005), pp. 632–657.

[10] David Hume. *An Enquiry Concerning Human Understanding*. 1748.

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

[12] Amir-Hossein Karimi*, Julius von Kügelgen*, Bernhard Schölkopf, and Isabel Valera. “Towards Causal Algorithmic Recourse”. In: *xxAI-Beyond Explainable AI*. Vol. Lecture Notes in AI 13200. (*equal contribution). Springer. 2022, pp. 139–166.

[13] Kok Yong Lee. “Hiddleston’s causal modeling semantics and the distinction between forward-tracking and backtracking counterfactuals”. In: *Studies in Logic* 10.1 (2017).

[14] David Lewis. “Counterfactual dependence and time’s arrow”. In: *Noûs* (1979), pp. 455–476.

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

[16] Barry Loewer. *The consequence argument meets the Mentaculus*. Working papers, Rutgers University. 2020.

[17] Christopher G Lucas and Charles Kemp. “An improved probabilistic account of counterfactual reasoning.”. In: *Psychological review* 122.4 (2015), p. 700.

References III

[18] Divyat Mahajan, Chenhao Tan, and Amit Sharma. "Preserving causal constraints in counterfactual explanations for machine learning classifiers". In: *arXiv preprint arXiv:1912.03277* (2019).

[19] Judea Pearl. *Causality*. Cambridge university press, 2009.

[20] 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*. 2020, pp. 344–350.

[21] Julian Reiss. "Counterfactuals, thought experiments, and singular causal analysis in history". In: *Philosophy of Science* 76.5 (2009), pp. 712–723.

[22] Lance J Rips. "Two causal theories of counterfactual conditionals". In: *Cognitive science* 34.2 (2010), pp. 175–221.

[23] 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*. 2020, pp. 166–172.

[24] Philip E Tetlock and Aaron Belkin. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press, 1996.

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

References IV

[26] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[27] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. "The what-if tool: Interactive probing of machine learning models". In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

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