Scope-Restricted Backtracking Counterfactuals

Johan de Aguas JOHANMD@UIO.NO
Integreat: Norwegian Centre for Knowledge-driven Machine Learning, University of Oslo, Norway

Martin Jullum JULLUM@NR.NO
Norwegian Computing Center, Norway

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

Interventional counterfactuals evaluate statements such as "had A been a" by replacing the assignment for A while keeping the unit's exogenous background fixed. Backtracking counterfactuals reverse this logic: they preserve the structural assignments and instead vary background conditions across worlds. This semantics is natural for diagnostic explanation, but full backtracking may be too permissive because every exogenous coordinate can vary. We introduce *scope-restricted backtracking counterfactuals*, in which the modeler specifies which coordinates may vary while the remainder are shared across worlds. The scope determines the admissible carriers of the contrast, while a cross-world kernel governs the plausibility of their alternative values. We present model-based inference for general scope-restricted backtracking counterfactuals, derive sharp marginal bounds for a specified explanatory query, and establish frontier and compatibility conditions under which this explanatory query admits a modified-value policy representation. A linear-Gaussian example distinguishes local-noise, ancestor, irrelevant-noise, and infeasible explanations across scopes, while an application to the UCI Adult benchmark illustrates kernel specification, sensitivity, and how much of the explanation is carried by the cross-world coupling itself.

Keywords: Counterfactual explanations, algorithmic recourse, structural causal models

1. Introduction

Counterfactual reasoning is central to explainable artificial intelligence (XAI), where explanations often ask what would have needed to be different for an observed outcome to change (Wachter et al., 2018). In *algorithmic recourse*, this question is action-oriented: what feasible action could an agent take to obtain a more favorable outcome? (Ustun et al., 2019; Karimi et al., 2020). In *diagnostic explanations*, by contrast, the question is contrastive but not necessarily actionable: what circumstances would have had to differ for an alternative outcome to occur? (Miller, 2019; Beckers, 2022; Budhathoki et al., 2022).

From a structural-causal perspective, counterfactuals are model-defined cross-world quantities rather than merely alternative feature vectors produced by an explanation method. Within this framework, action-oriented and diagnostic interpretations correspond naturally to interventional and backtracking semantics, respectively. Under an *interventional counterfactual* (Pearl, 2009), the antecedent "had A been a" is evaluated by replacing the structural assignment for A while keeping the unit's background characteristics fixed. Under a *backtracking counterfactual* (von Kügelgen et al., 2023), the structural assignments are left unchanged and the antecedent is instead realized by revising the exogenous background through a cross-world kernel. Counterfactual worlds thus remain on the structural

© 2026 J. de Aguas & M. Jullum.

manifold, closely connecting backtracking to the plausibility and on-manifold constraints used in XAI (Poyiadzi et al., 2020; Crupi et al., 2024; Redelmeier et al., 2024).

Much of practical XAI, however, generates "counterfactuals" without an explicit causal graphical model. The method of Wachter et al. (2018), along with many approaches surveyed by Verma et al. (2024) and Guidotti (2024), searches directly in feature space for perturbations that alter a classifier's prediction. Other methods constrain this search to actionable or causally plausible regions: Mothilal et al. (2020), for example, allow features to be designated as immutable or bounded, while Mahajan et al. (2019) use partial structural models to preserve causal relations. Although such constraints can encode actionability or plausibility, without an explicit causal model they do not specify how changes to one feature propagate to its descendants.

Full backtracking uses structural information, but may be too permissive for explanation. In its unrestricted form, any exogenous coordinate may vary, including coordinates irrelevant to the antecedent or whose revision changes features intended to remain fixed. This conflicts with several common XAI goals: sparse explanations seek a small locus of change; recourse typically excludes immutable features such as age or place of birth; diagnostic queries may hold selected features fixed; and fairness queries may require fixing a protected attribute, varying it while preserving selected background facts, or restricting the pathways through which it affects a prediction (Kusner et al., 2017; Kilbertus et al., 2017; von Kügelgen et al., 2022). Because backtracking scope is defined over exogenous coordinates, translating such feature-level requirements into scope remains model-dependent.

This paper develops *scope-restricted* backtracking counterfactuals. The modeler selects which exogenous coordinates may backtrack, while the remainder stay at their factual values. Restricted backtracking is suggested in von Kügelgen et al. (2023) and appears in related computational forms: Kladny et al. (2024) impose sparsity on counterfactual representations, Hao et al. (2024) restrict variation to a learned manifold, and Fatemi et al. (2025) operationalize restricted backtracking with a deep generative model. Our focus is complementary: we give scope an explicit cross-world semantics, separate it from kernel specification, and characterize when the resulting diagnostic query admits a modified-value-policy representation. This gives the explanation a policy interpretation and, under additional actionability and effectiveness requirements, connects it to algorithmic recourse.

Contributions. We define scope-restricted backtracking through scoped cross-world laws, establish a simultaneous σ-regime representation, and develop model-based inference clarifying how scope, latent exogenous states, and kernel specification determine support-based antecedent feasibility. We show that if a downstream frontier intercepts all relevant effects of the scoped revision and query–response compatibility holds, the conditional explanation law admits a modified-value policy representation. When the full cross-world law is unavailable, Appendix C derives sharp marginal bounds for discrete or event-valued explanations from selected factual and counterfactual margins, together with denominator-based refinements. Finally, we illustrate scoped diagnostic explanations in XAI, their sensitivity to scope and kernel choices, and the additional requirements needed for a recourse interpretation.

Preliminaries. We work throughout with structural causal models (SCMs) by Pearl (2009) and the associated semantics of hard and soft interventions by Correa and Bareinboim (2020a,b). Appendix A collects the formal definitions, notation, and theoretical background

used in the paper, including the SCM solution map, graphical notation, hard-intervention counterfactuals, and the σ-regime formulation of soft interventions and its special cases.

Throughout, $\mathcal{M} = (\mathcal{V}, \mathcal{U}, \mathcal{G}, \mathcal{F}, P_{\mathcal{U}})$ denotes an SCM with endogenous variables $\mathcal{V}$, exogenous background $\mathcal{U}$, directed acyclic graph (DAG) $\mathcal{G}$, causal assignments $\mathcal{F} = \{f_V\}_{V \in \mathcal{V}}$, and background law $P_{\mathcal{U}}$; U_V denotes the exogenous coordinates entering f_V, $\Upsilon_{\mathcal{M}}$ the SCM solution map, and $\mathrm{pa}(\cdot; \mathcal{G})$, $\mathrm{an}(\cdot; \mathcal{G})$, $\mathrm{de}(\cdot; \mathcal{G})$, and $\mathrm{nde}(\cdot; \mathcal{G})$ the usual graphical parent, ancestor, descendant, and non-descendant operators. Formal definitions are given in Appendix A.

With an abuse of notation, we use $\mathrm{d}P$ both for the density/mass of P with respect to a suitable dominating measure μ_{dom} and for the corresponding probability element.

2. Backtracking Counterfactuals

Backtracking counterfactuals (von Kügelgen et al., 2023) keep the structural assignments of an SCM fixed and allow the exogenous background to differ across factual and counterfactual worlds. Instead of replacing components of $\mathcal{F}$, they augment $\mathcal{M}$ with a counterfactual copy of $\mathcal{U}$ and a cross-world *backtracking conditional kernel* $B(u^* \mid u)$. This kernel specifies which counterfactual backgrounds u^* are considered plausible relative to a factual background u, before conditioning on evidence. It induces cross-world laws of the form

$$\mathbb{P}(Y^* \in \mathcal{Y}, Z \in \mathcal{Z}) := \iint \mathbb{I}\{Y(u^*) \in \mathcal{Y}\}\, \mathbb{I}\{Z(u) \in \mathcal{Z}\}\, \mathrm{d}B(u^* \mid u)\, \mathrm{d}P_{\mathcal{U}}(u), \tag{1}$$

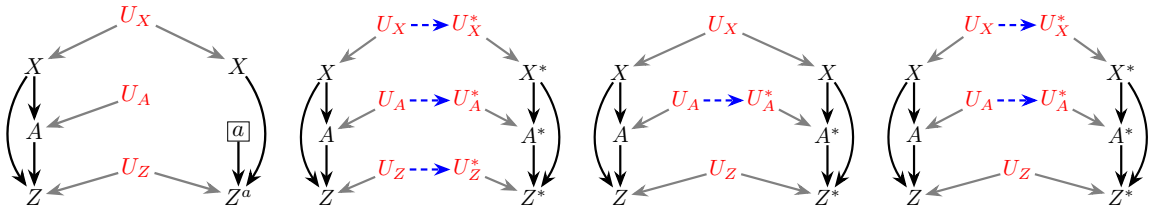
for measurable $\mathcal{Y} \subseteq \mathrm{dom}\, Y$ and $\mathcal{Z} \subseteq \mathrm{dom}\, Z$. Here $Y(u^*)$ is obtained by solving the unchanged structural equations $\Upsilon_{\mathcal{M}}$ at the counterfactual exogenous input u^*. Given factual evidence $H = h$, the factual background distribution $P_{\mathcal{U}}$ in (1) is updated by abduction. Conditional backtracking queries are then obtained by ordinary probabilistic conditioning on the relevant factual and counterfactual events, whenever these have positive probability.

A common modeling choice is a distance-based Gibbs kernel $\mathrm{d}B(u^* \mid u) \propto \exp\{-\beta\, d(u^*, u)\}$, which favors counterfactual backgrounds close to the factual one. The choice of B constitutes a cross-world structure: observational data may inform the distribution of factual states, but do not generally determine how a factual background u should be paired with a counterfactual one u^*. A generative model may provide a latent representation and meaningful distances, but the cross-world coupling still requires a modeling choice, calibration, and sensitivity analysis. Under scope restriction, this distinction becomes explicit: the scope determines *which* coordinates may change, while the kernel determines *how* they change.

Two points are important for what follows. First, an antecedent may be infeasible: if $X(u^*) = x^*$ has no support under the unchanged structural assignments and the chosen kernel, the corresponding conditional law is undefined. Second, because full backtracking absorbs the contrast through the exogenous state, it may revise ancestors, local noises, and descendants of the antecedent, including coordinates unrelated to realizing it. Section 5 illustrates both phenomena in a linear-Gaussian SCM.

3. Restricted Backtracking Counterfactuals

Instead of drawing a wholly new background u^*, choose a scope $\mathcal{E} \subseteq \mathcal{U}$ of exogenous coordinates that may backtrack and keep $\bar{\mathcal{E}} = \mathcal{U} \setminus \mathcal{E}$ fixed. Write $u = (e, \bar{e})$ and $u^* = (e^*, \bar{e}^*)$ for the corresponding decomposition into scoped and pinned coordinates.



(a) Interventional (b) Full backtracking (c) $\mathcal{E} = \{U_A\}$ (d) $\mathcal{E} = \{U_X, U_A\}$

Figure 1: Twin-network (Balke and Pearl, 1994) comparison of interventional, full-backtracking, and scope-restricted backtracking in a Markovian SCM. (a) The antecedent A is fixed by an intervention to a, represented by the boxed node, and its incoming arrows are removed. (b) Under full backtracking, all exogenous coordinates may differ across factual and counterfactual worlds, coupled by B. (c) Under restricted backtracking with scope $\mathcal{E} = \{U_A\}$, only the local exogenous input of A may change, while all remaining ones are shared across worlds. (d) Under scope $\mathcal{E} = \{U_X, U_A\}$, the background of the ancestor X may also change, while U_Z remains pinned to its factual value.

Definition (Scope-restricted backtracking kernel). A *scope-restricted backtracking kernel with scope* $\mathcal{E}$ is a cross-world kernel of the form

$$\mathrm{d}B(u^* \mid u) = \mathrm{d}K(e^* \mid u)\, \delta_{\bar{e}}(\bar{e}^*), \tag{2}$$

where K specifies how the coordinates in $\mathcal{E}$ may change and $\delta_{\bar{e}}$ denotes the density/mass of the Dirac measure at $\bar{e}$, which pins all coordinates in $\bar{\mathcal{E}}$ to their factual values. With a slight abuse of notation, (2) should be understood, at the level of distribution functions as $B(e^*, \bar{e}^* \mid u) = K(e^* \mid u)\mathbb{I}\{\bar{e} \le \bar{e}^*\}$, with the inequality taken componentwise.

Given a factual background $u = (e, \bar{e})$, draw $e^* \sim K(\cdot \mid e, \bar{e})$ and set $u^* = (e^*, \bar{e})$. For any endogenous variable V, write $V^* := V(e^*, \bar{e})$ for its value obtained by solving the unchanged SCM at this partially revised background. Thus, for endogenous $Y, Z \subseteq \mathcal{V}$ and measurable $\mathcal{Y}, \mathcal{Z}$, the scoped cross-world law is

$$\mathbb{P}(Y^* \in \mathcal{Y}, Z \in \mathcal{Z}) = \iint \mathbb{I}\{Y(e^*, \bar{e}) \in \mathcal{Y}\}\, \mathbb{I}\{Z(e, \bar{e}) \in \mathcal{Z}\}\, \mathrm{d}K(e^* \mid e, \bar{e})\, \mathrm{d}P_{\mathcal{U}}(e, \bar{e}). \tag{3}$$

If $\mathcal{E} = \varnothing$, this gives the observational law; if $\mathcal{E} = \mathcal{U}$ and $K = B$, it gives full backtrack (1).

Scope restriction is not the same as querying a factual variable alongside a counterfactual one. If V is downstream of the antecedent, the factual value $V(u)$ remains unchanged and therefore does not propagate the counterfactual contrast. By contrast, the full backtracking value $V(u^*)$ propagates the contrast while also allowing its local noise u_V^* to change, whereas the restricted value $V(e^*, \bar{e})$ propagates the contrast while retaining the factual noise whenever $U_V \in \bar{\mathcal{E}}$. Thus, restricted backtracking is a sub-family of full backtracking obtained through kernel (2), but it makes non-revision of $\bar{\mathcal{E}}$ a hard constraint rather than merely encouraging it through a distance penalty. Example 1 shows the consequence: pinning U_Z removes the irrelevant σ_Z^2 variance contributed by full backtracking.

3.1. Choosing the scope and kernel

The scope and kernel encode distinct modeling choices. The scope $\mathcal{E}$ determines which exogenous coordinates are permitted to differ across worlds, while K determines which

revisions of those coordinates are supported and how they are weighted. Thus, changing the scope alters the admissible directions of revision, whereas, for a fixed support, changing K changes their relative plausibility. We say that K is *decomposable* if, for every u, one has $K(e^* \mid u) = \prod_{U_j \in \mathcal{E}} K_j(u_j^* \mid u_j)$.

Figure 1 contrasts interventional and full-backtracking counterfactuals with two scope-restricted constructions under a decomposable kernel.

There are several practical ways to specify K. One option is to impose a structurally motivated *close-to-data-manifold* kernel, for example $\mathrm{d}K(e^* \mid u) \propto \exp\{-\beta\, d_{\mathcal{E}}(e^*, e)\}$ or, in a decomposable continuous model, $U_j^* \mid (U_j = u_j) \sim \mathcal{N}(u_j, \kappa_j^2)$ for $U_j \in \mathcal{E}$. Here, $d_{\mathcal{E}}$ encodes which revisions are considered similar, while β or κ_j controls the strength of cross-world similarity. Alternatively, a fitted generative model may provide latent coordinates and a useful geometry for defining $d_{\mathcal{E}}$. When no single kernel is clearly defensible, β, κ_j, or other aspects of K can instead be treated as sensitivity parameters, or one may target the partial-identification analysis in Appendix C.

Scope restriction provides control over which background coordinates may change, but an overly narrow scope may prevent the desired counterfactual contrast from being realized at all. Feasibility is therefore a prerequisite for a conditional backtracking query and provides a basic criterion for scope selection. Let $V \subseteq \mathcal{V}$ denote the endogenous variables appearing in the antecedent. Given factual evidence $Z = z$, we say that an antecedent $V^* = v$ is *feasible* under scope $\mathcal{E}$ if v lies in the support of the scoped conditional law of V^* given $Z = z$. In the discrete case, this reduces to $\mathbb{P}(V^* = v, Z = z) > 0$ whenever the factual event $Z = z$ itself has positive probability. This yields a simple monotonicity principle for nested scopes.

Proposition 1 (Support monotonicity) *Let $\mathcal{E}_1 \subseteq \mathcal{E}_2$, write $\mathcal{D} = \mathcal{E}_2 \setminus \mathcal{E}_1$, and let $K_{\mathcal{E}_j}(\cdot \mid u)$ denote the scope kernel associated with $\mathcal{E}_j$, $j \in \{1, 2\}$. Suppose that, for every factual background u, $\mathrm{supp}\{K_{\mathcal{E}_1}(\cdot \mid u)\} \times \{d(u)\} \subseteq \mathrm{supp}\{K_{\mathcal{E}_2}(\cdot \mid u)\}$, where $d(u)$ denotes the factual value of the coordinates in $\mathcal{D}$. Then every antecedent feasible under $\mathcal{E}_1$ is feasible under $\mathcal{E}_2$.*

The property does not compare scopes that are not nested, since different scopes may activate distinct paths for realizing the same antecedent (see Appendix B.2 and Example 1).

3.2. Model-based counterfactual inference

Suppose the relevant structural mechanisms of the SCM are known, or are represented by a fitted generative model that supports abduction and prediction, and let the scope-restricted kernel K be specified. Conditional queries such as $\mathbb{P}(X^* \in \mathcal{X} \mid Y^* = y^*, Z = z)$, where Z here denotes generically the available factual evidence, can then be computed by the usual abduction–prediction logic, subject to the restriction $\bar{e}^* = \bar{e}$:

1. **Partial cross-world abduction.** Update the joint law $K(e^* \mid e, \bar{e})\, P_{\mathcal{U}}(e, \bar{e})$ using the counterfactual and factual evidence:

$$\mathrm{d}P_{\mathrm{post}}(e^*, e, \bar{e} \mid Y^* = y^*, Z = z) \propto \mathbb{I}\{Y(e^*, \bar{e}) = y^*\}\, \mathbb{I}\{Z(e, \bar{e}) = z\}\, \mathrm{d}K(e^* \mid e, \bar{e})\, \mathrm{d}P_{\mathcal{U}}(e, \bar{e}),$$

with normalizing constant obtained by integration over $(e^*, e, \bar{e})$.

2. **Marginalization.** Remove factual coordinates not needed for prediction:

$$\mathrm{d}P_{\mathrm{marg}}(u^* \mid Y^* = y^*, Z = z) = \int \mathrm{d}P_{\mathrm{post}}(e^*, e, \bar{e} \mid Y^* = y^*, Z = z)\, \mathrm{d}\mu_{\mathrm{dom}}(e),$$

   where $u^* = (e^*, \bar{e})$. Thus, the resulting posterior is supported only on counterfactual worlds sharing the factual coordinates $\bar{e}$.

3. **Prediction.** Evaluate the unchanged structural assignments under this posterior:

$$\mathbb{P}(X^* \in \mathcal{X} \mid Y^* = y^*, Z = z) = \int \mathbb{I}\{X(u^*) \in \mathcal{X}\}\, \mathrm{d}P_{\mathrm{marg}}(u^* \mid Y^* = y^*, Z = z).$$

For continuous conditioning variables, the displays above are understood through the corresponding densities or regular conditional distributions.

3.3. Scoped counterfactual explanations

A natural counterfactual-explanation query is to rank alternative values of endogenous explanatory variables $X \subseteq \mathrm{nde}(Y; \mathcal{G})$ that accompany a desired outcome $Y^* = y^*$ under the chosen scope, conditional on the factual information available for the unit.

Definition (MAP scoped backtracking explanation). Let $H = (X, Y, Z)$ collect the factual evidence, with realization $h = (x, y, z)$, where $Z = z$ denotes any additional pre-X factual information beyond the explanatory variables $X = x$ and factual outcome $Y = y$. Under scope $\mathcal{E} \subseteq \mathcal{U}$, the posterior predictive density or mass

$$p_{X^*}(x^* \mid Y^* = y^*, H = h) \tag{4}$$

ranks candidate values x^* among scoped counterfactual worlds consistent with the factual evidence and realizing the desired outcome (Fatemi et al., 2025). Here, X^* is induced by revisions of the exogenous coordinates in $\mathcal{E}$ propagated through the unchanged structural assignments, so $\mathcal{E}$ need not correspond directly to the coordinates of X. A *maximum a posteriori* (MAP) scoped backtracking explanation is any

$$x^*_{\mathrm{MAP}} \in \arg\max_{x^*} p_{X^*}(x^* \mid Y^* = y^*, H = h). \tag{5}$$

This posterior-mode view parallels mode-based backtracking explanations (Kladny et al., 2024; Kinjo, 2026), while related probabilistic approaches characterize explanations through the full posterior (Raman et al., 2023). Intuitively, x^*_{MAP} is the most plausible value of X among scoped worlds that agree with the factual evidence $H = h$ and realize the desired contrast $Y^* = y^*$. It thus provides a point-valued diagnostic explanation, with admissible background revisions controlled by $\mathcal{E}$ and their relative plausibility governed by K.

Partial identification. When a fitted generative model for the SCM is unavailable, or the scoped cross-world law is only partially specified, the explanation query in (4) may not be point identified. Proposition 4 in Appendix C derives sharp marginal bounds for discrete or event-valued versions of this target from selected factual and counterfactual margins. The appendix also gives explicit refinements when the query denominator is known or bounded and a general feasible-set template for incorporating additional kernel restrictions.

4. Backtracking Counterfactuals and Modified-Value Policy Frontiers

Restricted backtracking acts by revising part of the exogenous background and then propagating the original structural assignments. This section asks when the same explanatory behavior can instead be represented more locally, by modifying a collection of *endogenous* variables. Specifically, we ask whether the effect of the scoped revision on the pair (X^*, Y^*) can be "pushed forward" to an endogenous *frontier* $A \subseteq \mathcal{V}$: once the revised value A^* at that frontier is known, can the upstream background revision be discarded and replaced by an intervention on A?

Such a reduction is useful because it connects a backtracking query to the language of stochastic and modified-value interventions (Haneuse and Rotnitzky, 2013; Sani et al., 2020) without requiring the entire scoped cross-world mechanism to admit such a representation. The frontier need not coincide with the explanatory variable X: it may lie upstream or downstream of X, contain X itself, or consist of several variables jointly. Moreover, the induced policy may depend on the natural frontier value $A^\natural = A(u)$. This dependence is important: a backtracking revision can modify a variable differently depending on where that variable started, and conditioning on $A^\natural$ retains this factual–counterfactual coupling.

A simultaneous σ-regime representation of the full scoped law is given in Appendix B.1. Here the aim is deliberately more local: to reproduce only the joint behavior of (X^*, Y^*) needed by the explanatory query (4).

4.1. Reducibility to a modified-value policy

Let

$$S_{\mathcal{E}} := \{V \in \mathcal{V} : U_V \cap \mathcal{E} \neq \varnothing\} \tag{6}$$

denote the endogenous variables whose *local* exogenous arguments can be directly altered by the scope. For a candidate frontier $A \subseteq \mathcal{V}$, let

$$C := (X, Y), \qquad C^a(u) = (X^a(u), Y^a(u)).$$

The latter is the value of this target pair under $\mathrm{do}(A = a)$ and factual background $\mathcal{U} = u$.

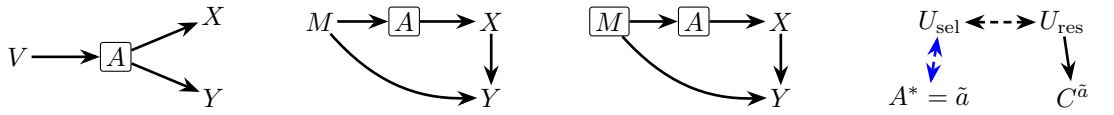
Two distinct obstacles can prevent such a reduction. The first is *structural*: a scoped revision may reach (X, Y) through a directed route that bypasses the proposed frontier. The second is *probabilistic*: even after all relevant directed effects pass through A, the particular value A^* selected by the backtracking construction may reveal latent factual information that changes how (X, Y) respond to that value. Figure 2 illustrates these two obstacles.

Proposition 2 (Reducibility through an endogenous frontier) *Let $\mathcal{E} \subseteq \mathcal{U}$ be a restricted backtracking scope, let $A \subseteq \mathcal{V}$ be a candidate endogenous frontier, and fix factual evidence $H = h$. Let $A^\natural := A(u)$ denote the natural value of the frontier. Suppose:*

(i) ***Downstream frontier.*** *For every factual background $u = (e, \bar{e})$ compatible with $H = h$ and every admissible scoped revision e^*, let $u^* = (e^*, \bar{e})$. Then*

$$C(u^*) = C^{A(u^*)}(u). \tag{7}$$

Thus, once the scoped revision has selected $A^* = A(u^*)$, assigning that same frontier value in the factual background reproduces both X^* and Y^*.



(a) Frontier, $V \in S_{\mathcal{E}}$ (b) Bypass of $\{A\}$ (c) Frontier $\{M, A\}$ (d) Selection vs. response information

Figure 2: Two requirements for reducibility. Boxes indicate the candidate frontier. (a) If all relevant effects of the scoped revision pass through A, then fixing A to its revised value can absorb the upstream change. (b) With scope affecting M and A, the route $M \to Y$ bypasses the singleton frontier $\{A\}$. (c) Enlarging the frontier to $\{M, A\}$ removes this structural bypass. (d) Structural interception is not sufficient: A^* may depend on latent selection information U_{sel}, while $C^{\tilde{a}}$ depends on latent response information U_{res}. Conditional dependence between these components can make selection response-informative; Corollary 3(ii) rules this out through a latent-separation criterion.

(ii) ***Query–response compatibility.*** *Let $W \subseteq \mathrm{nde}(A; \mathcal{G})$ denote observed pre-frontier information allowed to enter the policy. The variables W and H may overlap; whenever both appear, conditioning is understood on their joint information. For every $\tilde{a}$ in the conditional support of A^*,*

$$P_{C^{\tilde{a}}}\big(c \,|\, A^* = \tilde{a}, W = w, A^\natural = a^\natural, H = h\big) = P_{C^{\tilde{a}}}\big(c \,|\, W = w, A^\natural = a^\natural, H = h\big) \tag{8}$$

for every c. Thus, among factual units with the same $(W, A^\natural, H)$, learning that the backtracking construction selected $\tilde{a}$ provides no additional information about their response $C^{\tilde{a}}$ to that same frontier value; that is, selection itself is not response-informative.

Define the modified-value policy

$$\nu_h(\cdot \,|\, w, a^\natural) := P_{A^*}\big(\cdot \,|\, W = w, A^\natural = a^\natural, H = h\big). \tag{9}$$

After abduction on $H = h$, let the policy retain $(W, A^\natural)$, draw $A^{\nu_h} \sim \nu_h(\cdot \,|\, W, A^\natural)$ independently of the remaining latent background conditional on $(W, A^\natural, H)$, and propagate the original structural assignments. Then

$$p_{X^*}(x^* \,|\, Y^* = y^*, H = h) = p_{X^{\nu_h}}(x^* \,|\, Y^{\nu_h} = y^*, H = h). \tag{10}$$

The proposition separates reducibility into a causal-path condition and a dependence condition. Condition *(i)* gives the unit-level identity $C^* = C^{A^*}$: once the scoped revision has selected its frontier value, its downstream effect on (X, Y) can be reproduced by assigning that same value in the factual background. Condition *(ii)* then asks whether the remaining dependence between the selected value and the factual latent state can safely be discarded. A proof is given in Appendix B.3. Although conditions *(i)–(ii)* are stated at the level needed for reducibility, they can be guaranteed by stronger but more readily interpretable graphical and structural criteria. The following corollary gives one such pair: a directed-cut criterion for the frontier and a latent-separation criterion ensuring query–response compatibility.

Corollary 3 (Graphical and structural sufficient criteria) *Suppose:* (i) *the directly scope-affected target variables are themselves included in the frontier, $S_{\mathcal{E}} \cap (X \cup Y) \subseteq A$ and, for every $V \in S_{\mathcal{E}} \setminus A$ and $V^\dagger \in (X \cup Y) \setminus A$,*

$$\text{every directed path } V \to \cdot \to V^\dagger \text{ in } \mathcal{G} \text{ intersects } A; \tag{11}$$

and (ii) *there exist latent subvectors* $U_{\text{sel}}, U_{\text{res}} \subseteq \mathcal{U}$*, with complements in* $\mathcal{U}$ *denoted by* $\overline{U}_{\text{sel}}$ *and* $\overline{U}_{\text{res}}$*. Let* $Q = (W, A^\natural)$ *and* $L = (Q, H) = (W, A^\natural, H)$*. For every* $\tilde{a}$ *in the conditional support of* A^**, suppose*

$$A^* \perp\!\!\!\perp \overline{U}_{\text{sel}} \,|\, U_{\text{sel}}, L, \quad C^{\tilde{a}} \perp\!\!\!\perp \overline{U}_{\text{res}} \,|\, U_{\text{res}}, L, \quad U_{\text{sel}} \perp\!\!\!\perp U_{\text{res}} \,|\, L. \tag{12}$$

Then condition (i) *implies the downstream-frontier condition* (7)*, while condition* (ii) *implies the query–response compatibility condition* (8)*. Consequently, the scoped explanatory query* (4) *is reducible to the modified-value query* (10)*.*

Condition (11) requires the frontier to contain any target directly affected by the scope and to intercept every remaining directed route from a scope-affected mechanism to the target pair (X, Y). Condition (12), in turn, separates latent information relevant to selecting A^* from latent information relevant to the responses $C^{\tilde{a}}$, conditional on $L = (W, A^\natural, H)$. Thus, the first criterion can be assessed from the causal graph together with the scope-to-mechanism map $S_{\mathcal{E}}$, whereas the second concerns the joint SCM–backtracking construction. Both are sufficient rather than necessary, but together they provide a concrete way to diagnose when the more general conditions of Proposition 2 are satisfied.

Overall, the reduction clarifies when a scoped backtracking explanation can be represented by a modified-value policy: the scoped change must be absorbed by an endogenous frontier, and the way that frontier value is selected must not carry additional response-relevant information. When either requirement fails, the backtracking query remains genuinely a cross-world counterfactual. When reduction does hold, the resulting policy should still be interpreted according to the information it uses: if it depends on post-outcome evidence such as Y, it is a diagnostic representation. A prospective deployment interpretation additionally requires the policy inputs to be available before frontier assignment; a recourse interpretation further requires the frontier variables to be manipulable or actionable and the induced policy to satisfy an appropriate effectiveness criterion.

5. A Linear-Gaussian Example

This example makes the roles of scope, feasibility, and kernel choice concrete. Different scopes isolate local-noise and ancestor explanations, exclude irrelevant downstream revisions, or make the antecedent infeasible.

Example 1 *Consider the Markovian SCM*

$$X := U_X, \qquad Y := X + U_Y, \qquad Z := X + Y + U_Z, \tag{13}$$

with factual observation $(X, Y, Z) = (1, 2, 2)$*, so* $u = (U_X, U_Y, U_Z) = (1, 1, -1)$*. Notice that the exogenous realization is not separately observed here; it is recovered by deterministic abduction from the structural assignments. We ask the diagnostic question: what would* (X^*, Z^*) *have been, had* $Y^* = 3$*? A hard intervention* $\text{do}(Y = 3)$ *changes the mechanism for* Y*, whereas full backtracking keeps all mechanisms unchanged and allows all exogenous coordinates to differ. Scope restriction lies between these extremes by specifying which coordinates may carry the contrast. Table 1 summarizes the resulting explanations.*

For a Gaussian full-backtracking kernel $B(u^* \mid u) \equiv \mathcal{N}\{u, \mathrm{diag}(\sigma_X^2, \sigma_Y^2, \sigma_Z^2)\}(u^*)$*, which is decomposable with coordinate-wise conditional factors* $\mathcal{N}(u_X, \sigma_X^2)$*,* $\mathcal{N}(u_Y, \sigma_Y^2)$*, and* $\mathcal{N}(u_Z, \sigma_Z^2)$*, conditioning on* $Y^* = 3$ *imposes* $U_Y^* = 3 - U_X^*$*. Define*

$$\xi := \frac{\sigma_Y^2}{\sigma_Y^2 + \sigma_X^2} \in (0, 1).$$

Then

$$(X^*, Z^*) \mid (Y^* = 3, u) \sim \mathcal{N}\left(\begin{pmatrix} 2 - \xi \\ 4 - \xi \end{pmatrix}, \xi\sigma_X^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \sigma_Z^2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Full backtracking therefore allows U_Z^* *to vary even though* U_Z^* *is not needed to realize the antecedent. Restricted scopes separate the possible explanations:*

1. $\mathcal{E} = \{U_Y\}$ *pins* $U_X^* = 1$ *and* $U_Z^* = -1$*; the antecedent forces* $U_Y^* = 2$*, so* $(X^*, Z^*) = (1, 3)$*. This agrees with the hard counterfactual after* $\mathrm{do}(Y = 3)$*.*
2. $\mathcal{E} = \{U_X\}$ *pins* $U_Y^* = 1$ *and* $U_Z^* = -1$*; the antecedent forces* $U_X^* = 2$*, so* $(X^*, Z^*) = (2, 4)$*. The contrast in* Y *is explained by a different upstream* X*.*
3. $\mathcal{E} = \{U_X, U_Y\}$ *with independent Gaussian coordinate kernels gives*

$$U_X^* \mid (Y^* = 3, u) \sim \mathcal{N}\left(2 - \xi, \xi\sigma_X^2\right), \qquad U_Y^* = 3 - U_X^*,$$

   and $U_Z^* = -1$ *is pinned. Hence*

$$(X^*, Z^*) \mid (Y^* = 3, u) \sim \mathcal{N}\left(\begin{pmatrix} 2 - \xi \\ 4 - \xi \end{pmatrix}, \xi\sigma_X^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right).$$

   The mean agrees with full backtracking, but the variance contribution σ_Z^2 *disappears.*
4. $\mathcal{E} = \{U_Z\}$ *makes the antecedent infeasible. Since* U_X^* *and* U_Y^* *are pinned,* $Y^* = 1 + 1 = 2$*.*

The Gaussian kernel also makes the role of kernel sensitivity transparent. For the scope $\{U_X, U_Y\}$, the relative variances determine which admissible coordinate tends to carry the contrast: $\xi = \sigma_Y^2/(\sigma_Y^2 + \sigma_X^2)$ shifts the posterior mean of X^* between the local-noise and ancestor explanations. Thus the scope answers which explanations are admissible, while (σ_X^2, σ_Y^2) express how strongly the analyst prefers one type of admissible revision over another. Reporting results over a range of these parameters is a direct sensitivity analysis when no single cross-world kernel is scientifically privileged.

6. Application Case: Diagnostic Explanations and Recourse

We illustrate the framework on the UCI Adult benchmark (Becker and Kohavi, 1996), a standard tabular dataset for predicting whether annual income exceeds \$50K from demographic and socioeconomic variables. The dataset has also been used in causal fairness and recourse (von Kügelgen et al., 2022). We adopt the feature graph studied by Plečko and Meinshausen (2020), shown in Figure 3, and append a learned binary classifier

$$Y = f_\theta(X), \qquad X = (E, R, J, O),$$

Table 1: Comparison of intervention and backtracking scopes for the contrast $Y^* = 3$ in (13). The scope determines which background coordinates may account for the contrast; the kernel determines the plausibility of alternatives within that scope.

Semantics/Scope	How the contrast is realized	Consequence
Hard do$(Y = 3)$	mechanism for Y is replaced	$(X^*, Z^*) = (1, 3)$
$\mathcal{E} = \{U_Y\}$	local noise of Y changes	$(1, 3)$
$\mathcal{E} = \{U_X\}$	upstream X changes	$(2, 4)$
$\mathcal{E} = \{U_X, U_Y\}$	upstream and local noises share the contrast	distribution over (X^*, Z^*) with U_Z pinned
Full $\mathcal{E} = \mathcal{U}$	every background coordinate may change	adds uncertainty from U_Z^*, not needed for $Y^* = 3$
$\mathcal{E} = \{U_Z\}$	only downstream noise may change	antecedent infeasible

where E is education, R workclass, J hours per week, and O occupation. Non-root structural assignments are fitted with random forest regressors or classifiers according to variable type, and f_θ is itself a random-forest classifier; full preprocessing, learner, structural-fitting, scope, and kernel details are given in Appendix D. The fitted decision achieves accuracy 0.789 and ROC-AUC 0.797 on the held-out test set.

For individuals receiving the unfavorable decision $Y = 0$, we consider the diagnostic explanation query in (4) with desired outcome $y^* = 1$ and under different scopes $\mathcal{E}$ and a Gaussian-copula sensitivity kernel indexed by $\rho \in [0, 1]$. Within the scoped coordinates, $\rho = 0$ corresponds to independent factual and counterfactual draws, whereas $\rho = 1$ makes them identical. We consider $\rho \in \{0, 0.5, 0.9\}$ and focus on

$$\mathcal{E}_{\text{edu}} = \{U_E\}, \qquad \mathcal{E}_X = \{U_E, U_R, U_J, U_O\}, \qquad \mathcal{E}_{\text{full}} = \mathcal{U}.$$

The first permits revision only of the education disturbance, the second permits revision of disturbances local to the explanatory vector X, and the third allows full backtracking.

The resulting explanations exhibit the structural distinction motivating scope restriction. Under $\mathcal{E}_{\text{edu}}$ and $\rho = 0$, the equal-case average of the case-specific conditional probabilities $\mathbb{P}(E^* > E \mid Y^* = 1, H = h)$ is approximately 0.996; the corresponding averages for increasing hours and changing occupation are 0.766 and 0.633. Thus, although U_J and U_O are pinned, J^* and O^* need not remain factual: they can change through propagation of the revised education value along the original structural equations. Scope therefore restricts which background disturbances may carry the contrast, rather than directly fixing downstream endogenous variables.

Scope determines which coordinates may be revised, while the kernel determines how probability mass is distributed across the permitted revisions. Averaged over 200 factual negative cases, at $\rho = 0$, the mean case-specific values of $\mathbb{P}(Y^* = 1 \mid H = h)$ are 0.076, 0.082, and 0.089 under $\mathcal{E}_{\text{edu}}$, $\mathcal{E}_X$, and $\mathcal{E}_{\text{full}}$, respectively. Thus, in this fitted benchmark, broader scopes assign more probability to counterfactual backgrounds that reach the favorable decision. Increasing ρ has the opposite effect: at $\rho = 0.9$, the corresponding means fall to 0.032, 0.041, and 0.051. Since larger ρ forces the counterfactual state to be closer to the factual one, successful revisions become less likely.

The application also gives a direct instance of Proposition 2. Since $Y = f_\theta(X)$, taking the endogenous frontier to be $A = X$ intercepts every directed route to Y. Moreover, since f_θ is a fixed deterministic function once trained, setting $X = \tilde{x}$ fixes the target pair (X, Y) to $(\tilde{x}, f_\theta(\tilde{x}))$, so query–response compatibility holds automatically. Hence the scoped diagnostic explanation admits a modified-value policy representation on observable X. Because $H = (X, Y, Z)$, $A^\natural = X$, and $Y = f_\theta(X)$, the Y-component of H is redundant once $(Z, X^\natural)$ is known. Thus, in this application, the induced policy can be indexed only by pre-frontier Z and the natural frontier value $X^\natural$, giving the representation a modified-value policy interpretation. Interpreting it specifically as recourse additionally requires the components of X to be regarded as jointly actionable.

We use the fitted model to illustrate the marginal bounds derived in Appendix C. Let

$$F = \{Y = 0\}, \qquad T^* = \{Y^* = 1\}, \qquad I^* = \{E^* \geq 13, Y^* = 1\}.$$

The target

$$\mathbb{P}(E^* \geq 13 \,|\, Y = 0, Y^* = 1, S)$$

asks, within each sex stratum (S), among individuals with an unfavorable factual decision who attain a favorable counterfactual decision, how often the successful revision has $E^* \geq 13$, corresponding to a bachelor's degree or above on the Adult education-number scale.

If the fully specified fitted cross-world law is replaced by only its factual and counterfactual margins, the correspondence between factual negatives and counterfactual successes is lost. As a result, the margins alone are largely uninformative for the conditional query: in 16 of the 18 scope–kernel–sex configurations, they permit zero probability for the conditioning event, giving the trivial envelope $[0, 1]$. Providing just one extra cross-world quantity already sharpens the result considerably. If the overlap $\eta = \mathbb{P}(F \cap T^* \,|\, S)$ is known, thereby fixing the amount of probability mass shared by factual negatives and counterfactual successes, the median interval width across configurations falls to 0.122. In every case, the refined interval still contains the corresponding value under the fully specified fitted law.

7. Discussion and Conclusion

We introduced *scope-restricted backtracking counterfactuals*, which refine full backtracking by allowing only a designated subset of exogenous coordinates to differ across factual and counterfactual worlds. The framework separates the *scope*, determining where a contrast may arise, from the cross-world *kernel*, determining which alternatives are plausible, while revised exogenous values still propagate to endogenous descendants. We also characterize when restricted-backtracking queries admit modified-value policy representations and derive partial-identification bounds under limited cross-world information.

The framework nevertheless relies on modeling assumptions that should be made explicit. Latent exogenous variables can be handled through posterior abduction, but the cross-world kernel is not identified from observational data and must be scientifically specified or varied in sensitivity analysis. Corollary 3 provides sufficient rather than necessary conditions, and the sharp bounds in Appendix C apply to event-valued rather than pointwise density queries. Finally, the Adult analysis uses a benchmark SCM rather than an identified graph, and its Monte Carlo summaries do not propagate learner uncertainty.

Code availability

Code implementing the examples and reproducing the associated results is available at https://github.com/johandh2o/scope-restricted-backtracking-counterfactuals.

Acknowledgments

This work was supported by the Research Council of Norway, Integreat — Norwegian Centre for knowledge-driven machine learning, project number 332645.

References

Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, AAAI'94, pages 230–237. AAAI Press, 1994.

Barry Becker and Ronny Kohavi. Adult, 1996.

Sander Beckers. Causal explanations and XAI. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 90–109. PMLR, 11–13 Apr 2022.

Kailash Budhathoki, Lenon Minorics, Patrick Bloebaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 17–23 Jul 2022.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10093–10100, 2020a. doi: 10.1609/aaai.v34i06.6567.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 10902–10912. Curran Associates, Inc., 2020b. doi: 10.5555/3495724.3496639.

Riccardo Crupi, Alessandro Castelnovo, Daniele Regoli, and Beatriz San Miguel Gonzalez. Counterfactual explanations as interventions in latent space. *Data Mining and Knowledge Discovery*, 38(5):2733–2769, 2024. doi: 10.1007/s10618-022-00889-2.

Philip Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002. doi: 10.1111/j.1751-5823.2002.tb00354.x.

Johan de Aguas. Interpolated stochastic interventions based on propensity scores, target policies and treatment-specific costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 36654–36662, 2026. doi: 10.1609/aaai.v40i43.40989.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490, 2021.

Iván Díaz, Nicholas Williams, Katherine Hoffman, and Edward Schenck. Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857, 2023. doi: 10.1080/01621459.2021.1955691.

Robin Evans and Thomas Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, 42(4):1452–1482, 2014. doi: 10.1214/14-AOS1206.

Pouria Fatemi, Ehsan Sharifian, and Mohammad Hossein Yassaee. A new approach to backtracking counterfactual explanations: A unified causal framework for efficient model interpretability. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 16273–16285. PMLR, 13–19 Jul 2025.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38:2770–2824, 2024. doi: 10.1007/s10618-022-00831-6.

Sebastien Haneuse and Andrea Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in Medicine*, 32(30):5260–5277, 2013. doi: 10.1002/sim.5907.

Guang-Yuan Hao, Jiji Zhang, Biwei Huang, Hao Wang, and Kun Zhang. Natural counterfactuals with necessary backtracking. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33 of *NIPS '20*, pages 265–277, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Edward Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019. doi: 10.1080/01621459.2017.1422737.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 656–666, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Keita Kinjo. A generalized-bayes perspective on counterfactual explanations: Posterior-based decision-making and evaluation. *arXiv preprint arXiv:2607.29077*, 2026.

Klaus-Rudolf Kladny, Julius von Kügelgen, Bernhard Schölkopf, and Michael Muehlebach. Deep backtracking counterfactuals for causally compliant explanations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Alexander Levis, Edward Kennedy, Alec McClean, Sivaraman Balakrishnan, and Larry Wasserman. Stochastic interventions, sensitivity analysis, and optimal transport. *arXiv preprint arXiv:2411.14285*, 2024. doi: 10.48550/arXiv.2411.14285.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. In *CausalML: Machine Learning and Causal Inference for Improved Decision Making Workshop at NeurIPS 2019*, 2019.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi: 10.1016/j.artint.2018.07.007.

Ramaravind Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 978-0-521-89560-6.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

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*, AIES '20, pages 344–350, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375850.

Natraj Raman, Daniele Magazzeni, and Sameena Shah. Bayesian hierarchical models for counterfactual estimation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1115–1128. PMLR, 2023.

Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. MCCE: Monte carlo sampling of valid and realistic counterfactual explanations for tabular data. *Data Mining and Knowledge Discovery*, 38:1830–1861, 2024. doi: 10.1007/s10618-024-01017-y.

Numair Sani, Jaron Lee, and Ilya Shpitser. Identification and estimation of causal effects defined by shift interventions. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 949–958. PMLR, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3287560.3287566.

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan Hines, John Dickerson, and Chirag Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys*, 56(12):1–42, 2024. doi: 10.1145/3677119.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594, 2022. doi: 10.1609/aaai.v36i9.21192.

Julius von Kügelgen, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 11–14 Apr 2023.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018. doi: 10.2139/ssrn.3063289.

Appendix A. Preliminaries, Notation, and Theoretical Background

A.1. Structural causal models and notation

A *structural causal model* (SCM) is a tuple $\mathcal{M} = (\mathcal{V}, \mathcal{U}, \mathcal{G}, \mathcal{F}, P_{\mathcal{U}})$, where $\mathcal{V}$ is a finite set of endogenous variables, $\mathcal{U}$ is a finite set of background noise variables, $\mathcal{G}$ is a directed acyclic graph (DAG) over $\mathcal{V}$, $\mathcal{F} = \{f_V\}_{V \in \mathcal{V}}$ is a collection of measurable structural assignments, and $P_{\mathcal{U}}$ is a probability measure on $\operatorname{dom}\mathcal{U}$ (Pearl, 2009). For each $V \in \mathcal{V}$, let U_V denote the subvector of exogenous coordinates entering the structural assignment for V. Thus

$$V = f_V\big(\operatorname{pa}(V;\mathcal{G}), U_V\big) \quad P_{\mathcal{U}}\text{–almost surely.} \tag{14}$$

In the Markovian case, the subvectors U_V are node-specific and mutually independent. In the semi-Markovian case, latent common exogenous variables are represented graphically by bidirected edges and the projected graph over observables is an acyclic directed mixed graph (ADMG) (Evans and Richardson, 2014).

We assume the structural assignments admit a unique solution for every $u \in \operatorname{dom}\mathcal{U}$, as is guaranteed in particular by acyclicity. Let $\Upsilon_{\mathcal{M}} : \operatorname{dom}\mathcal{U} \to \operatorname{dom}\mathcal{V}$ denote the measurable solution map obtained by recursively solving (14) along a topological order of $\mathcal{G}$.

For $A \subseteq \mathcal{V}$, $\operatorname{pa}(A;\mathcal{G})$, $\operatorname{an}(A;\mathcal{G})$, $\operatorname{de}(A;\mathcal{G})$, and $\operatorname{nde}(A;\mathcal{G})$ denote its parents, ancestors, descendants, and non-descendants.

A.2. Hard and soft intervention and σ-regimes

For $A \in \mathcal{V}$, $a \in \operatorname{dom} A$, and $u \in \operatorname{dom} \mathcal{U}$, the hard-intervention counterfactual $Y^a(u)$ is the value of the Y-component evaluated at u in the recursive solution map $\Upsilon^a_{\mathcal{M}}$ of the structural equations obtained by replacing the assignment for A by the constant assignment $A = a$. The intervened graph removes arrows entering A in $\mathcal{G}$ (Pearl, 2009) and the corresponding population-level interventional law is

$$\mathbb{P}(Y \in \mathcal{Y} \mid \operatorname{do}(A = a)) \equiv \mathbb{P}(Y^a \in \mathcal{Y}) := \int \mathbb{I}\{Y^a(u) \in \mathcal{Y}\} \, \mathrm{d}P_{\mathcal{U}}(u), \qquad \mathcal{Y} \subseteq \operatorname{dom} Y,$$

with density or mass $p_Y(y \mid \operatorname{do}(A = a))$. Hard interventions provide the canonical semantics for action-oriented questions but are less natural for diagnostic questions that ask how the background conditions would have had to differ for an event to occur on-manifold.

Several formulations of stochastic and soft interventions have been proposed, including modified treatment policies (Haneuse and Rotnitzky, 2013; Díaz et al., 2023), shift intervention policies (Sani et al., 2020), incremental propensity-score interventions (IPI) (Kennedy, 2019), cost-interpolated IPIs (de Aguas, 2026), and generalized policies (Levis et al., 2024). One general SCM formulation is given by Correa and Bareinboim (2020a,b) through σ-regimes, which we adopt and present in a convenient unit-level form.

Fix $A \in \mathcal{V}$ and a unit with exogenous realization $\mathcal{U} = u$. Let ω be an auxiliary noise variable on a measurable space Ω with conditional law $P_{\omega \mid \mathcal{U}}(\cdot \mid u)$. A (unit-level) *soft intervention* or *σ-regime* on A is a mechanism replacement

$$\mathfrak{i}_A(u, \omega) = (X, g, \omega),$$

where $X \subseteq \operatorname{nde}(A; \mathcal{G})$ is a replacement parent set (within the original non-descendants of A), and $g : \operatorname{dom} X \times \operatorname{dom} U_A \times \Omega \to \operatorname{dom} A$ is a measurable replacement mechanism. The intervention replaces the original assignment of A at u

$$A(u) = f_A(\operatorname{pa}(A; \mathcal{G})(u), u_A) \quad \text{by} \quad A(u, \omega) = g(X(u), u_A, \omega),$$

leaving all other assignments untouched. Let $\Upsilon^{\mathfrak{i}_A}_{\mathcal{M}}$ denote the solution map of the new structural equations and $Y^{\mathfrak{i}_A}(u, \omega)$ be its Y-component at (u, ω). Then one can denote

$$\mathbb{P}(Y^{\mathfrak{i}_A} \in \mathcal{Y}) := \iint \mathbb{I}\Big\{Y^{\mathfrak{i}_A}(u, \omega) \in \mathcal{Y}\Big\} \, \mathrm{d}P_{\omega \mid \mathcal{U}}(\omega \mid u) \, \mathrm{d}P_{\mathcal{U}}(u).$$

For marginal queries, this can also be denoted with the semi-colon notation of Correa and Bareinboim (2020a,b): $\mathbb{P}(Y \in \mathcal{Y} \,;\, \mathfrak{i}_A)$. When ω is absent, we work with the singleton space $\{\varnothing\}$ and the degenerate law $\mathrm{d}P_{\omega \mid \mathcal{U}} = \delta_{\varnothing}$. The associated *operated graph* $\mathcal{G}^{\mathfrak{i}_A}$ deletes incoming arrows into A, adds arrows $X \to A$, and may mark the regime node $\mathfrak{i}_A \to A$ (Dawid, 2002).

The standard intervention classes are recovered as special cases. Hard interventions correspond to $X = \varnothing$ with constant-valued g and degenerate $\mathrm{d}P_{\omega \mid \mathcal{U}} = \delta_{\varnothing}$ (noise-free). Conditional interventions $\operatorname{do}\{A = h(X)\}$ correspond to noise-free $g(X, u_A, \varnothing) = h(X)$. Pure stochastic interventions reproducing a conditional distribution $P^{\dagger}(\cdot \mid X)$ correspond to $\omega \sim \operatorname{Unif}[0, 1]$, with $\omega \perp\!\!\!\perp \mathcal{U}$, and the conditional-quantile transform $g(X, u_A, \omega) = \inf\{a : P^{\dagger}(a \mid X) \geq \omega\}$. Except in the atomic case, the regime-level law $p_Y(y \,;\, \mathfrak{i}_A)$ is in general distinct from the conditional law $p_Y(y \mid A = a \,;\, \mathfrak{i}_A)$ (Correa and Bareinboim, 2020a,b).

Appendix B. Derivations and Proofs

B.1. Simultaneous σ-representation of restricted backtracking

For completeness, restricted backtracking can be written equivalently as a simultaneous σ-regime. By a simultaneous σ-regime we mean a collection $\{\mathfrak{i}_V : V \in \mathcal{V}\}$ of mechanism replacements applied jointly and driven by one shared auxiliary variable ω, whose conditional law may depend on the factual background $\mathcal{U} = u$. The resulting acyclic system is solved using the same draw ω in every replacement; here we take $\omega = u^* = (e^*, \bar{e})$, where $e^* \sim K(\cdot \mid e, \bar{e})$. Given a factual background $u = (e, \bar{e})$ and such a scoped revision e^*, define

$$\mathfrak{i}(u, e^*) = \{\mathfrak{i}_V(u, e^*) : V \in \mathcal{V}\}, \qquad \mathfrak{i}_V(u, e^*) = \big(\mathrm{pa}(V; \mathcal{G}), g_V, u^*\big),$$

with

$$g_V(x, u_V, u^*) := f_V(x, u^*_V),$$

where f_V is the original causal mechanism for V in the SCM. Thus every structural assignment retains its original parent set and structural map, but is evaluated using the same partially revised background u^*.

Since $\mathcal{G}$ is acyclic, solving the resulting assignments in topological order is exactly the ordinary SCM solution at u^*. Hence, pointwise,

$$Y^{\mathfrak{i}}(u, e^*) = Y(u^*) = Y(e^*, \bar{e}),$$

while factual variables remain evaluated at $u = (e, \bar{e})$. Integrating over $e^* \sim K(\cdot \mid e, \bar{e})$ and $u \sim P_{\mathcal{U}}$ therefore gives

$$\mathbb{P}(Y^{\mathfrak{i}} \in \mathcal{Y}, Z \in \mathcal{Z}) = \mathbb{P}(Y^* \in \mathcal{Y}, Z \in \mathcal{Z}),$$

which is precisely the scoped cross-world law in (3). Thus the simultaneous σ-regime is an equivalent notation for restricted backtracking.

B.2. Proof of Proposition 1

Write $\mathcal{D} = \mathcal{E}_2 \setminus \mathcal{E}_1$, and let $d(u)$ denote the factual value of the coordinates in $\mathcal{D}$. Fix a factual background u. Under scope $\mathcal{E}_1$, any feasible counterfactual background has the form

$$u^* = (e_1^*, d(u), \bar{e}_2), \quad \text{where} \quad e_1^* \in \mathrm{supp}\{K_{\mathcal{E}_1}(\cdot \mid u)\},$$

and $\bar{e}_2$ denotes the factual value of the coordinates outside $\mathcal{E}_2$. Hence

$$(e_1^*, d(u)) \in \mathrm{supp}\{K_{\mathcal{E}_1}(\cdot \mid u)\} \times \{d(u)\}.$$

By assumption,

$$\mathrm{supp}\{K_{\mathcal{E}_1}(\cdot \mid u)\} \times \{d(u)\} \subseteq \mathrm{supp}\{K_{\mathcal{E}_2}(\cdot \mid u)\},$$

so $(e_1^*, d(u)) \in \mathrm{supp}\{K_{\mathcal{E}_2}(\cdot \mid u)\}$. Since the coordinates outside $\mathcal{E}_2$ are fixed at the same factual values under both scopes, the same full counterfactual background

$$u^* = (e_1^*, d(u), \bar{e}_2)$$

is therefore feasible under $\mathcal{E}_2$ whenever it is feasible under $\mathcal{E}_1$.

Consequently, any antecedent $v = V(u^*)$ attainable under $\mathcal{E}_1$ is also attainable under $\mathcal{E}_2$. Hence every antecedent feasible under $\mathcal{E}_1$ is feasible under $\mathcal{E}_2$. The result does not impose an ordering on incomparable scopes. For example, Example 1 gives $\{U_Z\}$ and $\{U_Y\}$ as an incomparable pair. □

B.3. Proof of Proposition 2 and Corollary 3

Proof of Proposition 2. Fix $H = h$ and write

$$Q := (W, A^\natural).$$

Under the scoped backtracking construction, condition (7) implies

$$C^* = C^{A^*}(u) \qquad \text{almost surely given } H = h. \tag{15}$$

Let φ be any bounded measurable function on the state space of $C = (X, Y)$. Then, using (15),

$$\mathbb{E}[\varphi(C^*) \mid H = h] = \int \mathbb{E}[\varphi\{C^a(u)\} \mid A^* = a, Q = q, H = h] \, \mathrm{d}P_{A^*,Q}(a, q \mid H = h). \tag{16}$$

Query–response compatibility (8) removes the additional conditioning on the selected frontier value from the response term. Since

$$\nu_h(\cdot \mid q) = P_{A^*}(\cdot \mid Q = q, H = h),$$

equation (16) becomes

$$\mathbb{E}[\varphi(C^*) \mid H = h] = \iint \mathbb{E}[\varphi\{C^a(u)\} \mid Q = q, H = h] \, \mathrm{d}\nu_h(a \mid q) \, \mathrm{d}P_Q(q \mid H = h). \tag{17}$$

Now consider the modified-value policy. Conditional on $(Q, H = h)$, the policy draws

$$A^{\nu_h} \sim \nu_h(\cdot \mid Q)$$

independently of the remaining factual latent background and propagates the original structural assignments. Hence

$$C^{\nu_h} = C^{A^{\nu_h}}(u),$$

and the conditional independence built into the policy yields

$$\mathbb{E}[\varphi(C^{\nu_h}) \mid H = h] = \iint \mathbb{E}[\varphi\{C^a(u)\} \mid Q = q, H = h] \, \mathrm{d}\nu_h(a \mid q) \, \mathrm{d}P_Q(q \mid H = h). \tag{18}$$

The right-hand sides of (17) and (18) coincide for every bounded measurable φ. Therefore

$$C^* \mid (H = h) \quad \text{and} \quad C^{\nu_h} \mid (H = h)$$

have the same probability law. Since $C = (X, Y)$,

$$(X^*, Y^*) \mid (H = h) \overset{d}{=} (X^{\nu_h}, Y^{\nu_h}) \mid (H = h). \tag{19}$$

Whenever the corresponding conditional density or mass function exists and the conditioning value has positive normalizing mass, equality of the joint laws in (19) gives

$$p_{X^*}(x^* \mid Y^* = y^*, H = h) = p_{X^{\nu_h}}(x^* \mid Y^{\nu_h} = y^*, H = h),$$

which is (10). □

Proof of Corollary 3(i). Fix a factual background $u = (e, \bar{e})$ compatible with $H = h$, an admissible scoped revision e^*, and let $u^* = (e^*, \bar{e})$ and $a^* = A(u^*)$. Compare (a) the ordinary SCM solution under u^* with (b) the solution under factual background u after $\mathrm{do}(A = a^*)$. By construction, the two solutions agree at every node in A.

If a target $V^\dagger \in X \cup Y$ belongs to A, the two solutions therefore agree at $V^\dagger$. Otherwise, condition (i) gives $V^\dagger \notin S_{\mathcal{E}}$, so a discrepancy at $V^\dagger$ cannot originate from a changed local exogenous argument at $V^\dagger$.

Let $\Delta \subseteq \mathcal{V}$ denote the set of nodes at which the two solutions differ. As just noted, $\Delta \cap A = \varnothing$. Moreover, every $V' \in \Delta$ either satisfies $U_{V'} \cap \mathcal{E} \neq \varnothing$, so that $V' \in S_{\mathcal{E}} \setminus A$, or else has at least one parent in Δ, since otherwise both its parent values and its local exogenous argument would coincide across the two constructions and its assignment $f_{V'}$ would return the same value. Because $\mathcal{G}$ is acyclic, induction along a topological order therefore shows that every $V' \in \Delta$ is reached from some $V \in S_{\mathcal{E}} \setminus A$ by a directed path lying entirely within Δ, and hence disjoint from A. If some target $V^\dagger \in (X \cup Y) \setminus A$ belonged to Δ, this would produce a directed path $V \to \cdot \to V^\dagger$ avoiding A, contradicting (11).

Hence the two constructions agree at both target nodes,

$$C(u^*) = C^{A(u^*)}(u),$$

which is the downstream-frontier condition (7). □

Proof of Corollary 3(ii). Fix $\tilde{a}$ and write $Q = (W, A^\natural)$ and $L = (Q, H)$. Conditional on the factual latent state $\mathcal{U}$, $C^{\tilde{a}}$ is fixed by the SCM, whereas the randomness in A^* is supplied by the backtracking kernel. The first two conditional independences in (12) therefore imply

$$P_{A^*}(a^* \,|\, \mathcal{U}, L) = P_{A^*}(a^* \,|\, U_{\mathrm{sel}}, L), \qquad P_{C^{\tilde{a}}}(c \,|\, \mathcal{U}, L) = P_{C^{\tilde{a}}}(c \,|\, U_{\mathrm{res}}, L).$$

Because $C^{\tilde{a}}$ is a deterministic function of $\mathcal{U}$, the pair $(A^*, C^{\tilde{a}})$ is conditionally independent given $(\mathcal{U}, L)$, so iterated expectation gives

$$P_{A^*, C^{\tilde{a}}}(a^*, c \,|\, L) = \mathbb{E}[P_{A^*}(a^* \,|\, U_{\mathrm{sel}}, L) P_{C^{\tilde{a}}}(c \,|\, U_{\mathrm{res}}, L) \,|\, L]\,.$$

Since $U_{\mathrm{sel}} \perp\!\!\!\perp U_{\mathrm{res}} \,|\, L$, this factorizes as

$$P_{A^*, C^{\tilde{a}}}(a^*, c \,|\, L) = P_{A^*}(a^* \,|\, L)\, P_{C^{\tilde{a}}}(c \,|\, L),$$

and hence

$$A^* \perp\!\!\!\perp C^{\tilde{a}} \,|\, L.$$

Evaluating at $A^* = \tilde{a}$ gives

$$P_{C^{\tilde{a}}}\big(c \,|\, A^* = \tilde{a}, W, A^\natural, H\big) = P_{C^{\tilde{a}}}\big(c \,|\, W, A^\natural, H\big),$$

which is query–response compatibility (8). □

B.4. Understanding query–response compatibility

The downstream-frontier condition and query–response compatibility address different requirements. The former is structural: once the revised frontier value A^* is known, it asks whether the upstream scoped revision can be replaced by $\mathrm{do}(A = A^*)$. Query–response compatibility is instead probabilistic: after this replacement, it asks whether independently redrawing the selected frontier value would discard latent information that matters for the corresponding response.

To see this, write $Q = (W, A^\natural)$ and $L = (Q, H)$. Compatibility requires

$$P_{C^a}\big(c \,|\, A^* = a, L\big) = P_{C^a}\big(c \,|\, L\big)$$

for every c. Thus, $A^* = a$ may reveal latent information beyond L; the requirement is only that this information be irrelevant to the response C^a under the value actually selected. In particular, it does *not* require the stronger condition

$$A^* \perp\!\!\!\perp \mathcal{U} \,|\, W, A^\natural, H. \tag{20}$$

The following two examples illustrate the distinction.

Example 1: harmless latent dependence. Consider the Markovian SCM

$$Z = f_Z(U_Z), \qquad X = \mathbb{I}\{U_X \le \pi(Z)\}, \qquad Y = g(X, Z, U_Y),$$

with mutually independent (U_Z, U_X, U_Y). Restrict backtracking to $\mathcal{E} = \{U_X\}$ and let

$$U_X^* \,|\, (U_X = u_X) \sim \mathcal{N}(u_X, \sigma^2).$$

Take the frontier to be $A = X$ and let $W = Z$.

Conditioning on the factual value of X reveals only which side of the threshold contains U_X. Since the backtracking kernel is centered at the factual U_X, the selected value X^* generally retains additional information about U_X, so

$$X^* \not\perp\!\!\!\perp U_X \,|\, X, Z$$

in general. Hence the stronger latent-independence condition (20) need not hold.

This residual dependence is nevertheless harmless for the response. Under the intervention $X = a$,

$$C^a = (X^a, Y^a) = \big(a, g(a, Z, U_Y)\big),$$

which no longer depends on U_X. In the terminology of Corollary 3(ii), one may take $U_{\mathrm{sel}} = U_X$ and $U_{\mathrm{res}} = U_Y$: the selected frontier value depends on the former, the response depends on the latter, and their conditional separation follows from the Markovian exogenous structure. Hence the additional information about U_X conveyed by $X^* = a$ does not change the conditional distribution of C^a. Query–response compatibility can therefore hold even though X^* remains associated with the factual latent background.

Example 2: response-relevant latent selection. Now consider a factual exogenous background (U_X, U_R) satisfying

$$\mathbb{P}(U_X = 0, U_R = 0) = \mathbb{P}(U_X = 1, U_R = 1) = \frac{1}{2},$$

and define

$$X = \mathbb{I}\{U_X = 2\}, \qquad Y = XU_R.$$

Under the factual distribution, $(X, Y) = (0, 0)$ almost surely, so factual evidence $H = (X, Y) = (0, 0)$ does not distinguish the two latent backgrounds.

Restrict the scope to $\mathcal{E} = \{U_X\}$ and set $U_X^* = U_X + 1$. With frontier $A = X$,

$$X^* = \mathbb{I}\{U_X^* = 2\} = \mathbb{I}\{U_X = 1\} = U_R.$$

Moreover, the downstream-frontier condition holds: once X^* is fixed, the scoped change can be reproduced by intervening on X at that value.

Compatibility nevertheless fails because the selected frontier value now reveals a latent response modifier. Under $X = 1$,

$$Y^1 = U_R,$$

and since $X^* = U_R$,

$$\mathbb{P}\big(Y^1 = 1 \,|\, X^* = 1, H\big) = 1, \qquad \mathbb{P}\big(Y^1 = 1 \,|\, H\big) = \frac{1}{2}.$$

Here $U_{\text{sel}} = U_X$ and $U_{\text{res}} = U_R$ are perfectly dependent, so the latent-separation requirement in Corollary 3(ii) fails. Accordingly, $X^* = 1$ selects precisely those factual backgrounds with the high response $Y^1 = 1$. An independently redrawn modified-value policy can match the conditional distribution of X^*, but it cannot preserve this coupling between frontier selection and the latent response modifier.

The two examples isolate the role of query–response compatibility. Dependence between the backtracked frontier value and the factual latent background is not itself an obstacle to reducibility. Reducibility fails only when the dependence discarded by independently redrawing the frontier value carries response information under that same selected value.

Appendix C. Partial Identification

This section develops partial-identification results for the scoped explanation query in (4) when the joint dependence between factual and counterfactual variables is only partially specified. The sharp marginal result is stated for discrete or event-valued explanations and depends on selected factual and counterfactual margins rather than on observational information alone; for continuous variables, the corresponding statement concerns measurable events of positive conditional probability, not pointwise densities. Additional denominator information yields explicit refinements, while kernel restrictions can be incorporated through a feasible-set refinement when they can be translated into constraints on the relevant cross-world overlaps.

C.1. Sharp marginal bounds

We first consider the information contained in selected margins alone. Because (4) conditions on factual information $Z = z$, let

$$\mathbb{P}_z(\cdot) := \mathbb{P}(\cdot \mid Z = z),$$

with $\mathbb{P}_z = \mathbb{P}$ when no additional Z is present. We first write the construction for discrete X and Y and then state the event-valued analogue, which also covers measurable-set queries for continuous variables.

For a candidate explanation x^*, define

$$F := \{X = x, Y = y\}, \qquad T^* := \{Y^* = y^*\}, \qquad I^* := \{X^* = x^*, Y^* = y^*\}.$$

For example, suppose Y indicates whether a credit application is approved and X indicates whether the applicant has high debt. Then $F = \{X = 1, Y = 0\}$ represents applicants who are factually high-debt and rejected, $T^* = \{Y^* = 1\}$ represents counterfactual worlds in which the application is approved, and $I^* = \{X^* = 0, Y^* = 1\}$ represents counterfactual worlds in which the applicant has low debt and is approved. Thus I^* represents one specific candidate explanation for the desired outcome encoded by T^*. We have $I^* \subseteq T^*$ and

$$\mathbb{P}_z(I^* \mid F \cap T^*) = \mathbb{P}(X^* = x^* \mid Y^* = y^*, H = h).$$

Suppose the three stratum-specific margins

$$\phi_z := \mathbb{P}_z(F), \qquad \tau_z := \mathbb{P}_z(T^*), \qquad \gamma_z := \mathbb{P}_z(I^*)$$

are identified or otherwise specified, and define

$$\begin{aligned} r_z &:= \max\{0, \phi_z + \tau_z - 1\}, & s_z &:= \min\{\phi_z, \tau_z\}, \\ \ell_z &:= \max\{0, \phi_z + \gamma_z - 1\}, & m_z &:= \min\{\phi_z, \gamma_z\}. \end{aligned}$$

Proposition 4 (Sharp marginal explanation bounds) *Under the setup above, suppose $r_z > 0$, so that $F \cap T^*$ has positive probability under every cross-world coupling compatible with the specified margins. For a discrete candidate x^*, the sharp marginal bounds are*

$$p_{\mathrm{low},z}(x^*) = \frac{\ell_z}{r_z}, \qquad p_{\mathrm{upp},z}(x^*) = \min\left\{1, \frac{m_z}{r_z}\right\}. \tag{21}$$

Both endpoints are attainable over all cross-world couplings compatible with $(\phi_z, \tau_z, \gamma_z)$. More generally, if the point events defining F, T^*, and I^* are replaced by measurable events of positive conditional probability, the same formulas sharply bound $\mathbb{P}_z(I^* \mid F \cap T^*)$. Thus, for continuous variables the result applies to event-valued explanations; it does not by itself provide pointwise density or MAP-density bounds.

For discrete candidate explanations, a candidate x^* cannot be a MAP explanation when

$$p_{\mathrm{upp},z}(x^*) < \sup_{\tilde{x}} p_{\mathrm{low},z}(\tilde{x}),$$

and it is the unique MAP explanation whenever

$$p_{\mathrm{low},z}(x^*) > \sup_{\tilde{x} \neq x^*} p_{\mathrm{upp},z}(\tilde{x}).$$

Proof and sharpness. Let

$$a_z := \mathbb{P}_z(F \cap I^*), \qquad b_z := \mathbb{P}_z(F \cap T^*),$$

so that the target equals a_z/b_z. Fréchet inequalities and $I^* \subseteq T^*$ imply

$$\ell_z \le a_z \le m_z, \qquad r_z \le b_z \le s_z, \qquad 0 \le a_z \le b_z, \qquad 0 \le b_z - a_z \le \tau_z - \gamma_z.$$

For a fixed feasible denominator b, the smallest and largest feasible numerators are therefore

$$a_{\min}(b) = \max\{\ell_z, b - (\tau_z - \gamma_z)\}, \qquad a_{\max}(b) = \min\{m_z, b\}.$$

Hence the smallest and largest attainable ratios at b are

$$\max\left\{\frac{\ell_z}{b}, 1 - \frac{\tau_z - \gamma_z}{b}\right\}, \qquad \min\left\{1, \frac{m_z}{b}\right\},$$

respectively. Write $k_z := \tau_z - \gamma_z$, so that $a_{\min}(b)/b$ equals ℓ_z/b for $b \le \ell_z + k_z$ and $1 - k_z/b$ thereafter; the first branch is decreasing and the second increasing, so this map is not monotone on all of $(0, \infty)$. Since $\ell_z = \max\{0, r_z - k_z\}$, however, one has $\ell_z + k_z = \max\{r_z, k_z\}$. If $r_z \ge k_z$ this gives $\ell_z + k_z = r_z$, so only the increasing branch is active on $[r_z, s_z]$; if $r_z < k_z$ then $\ell_z = 0$ and the decreasing branch contributes the constant value 0. In either case $b \mapsto a_{\min}(b)/b$ is nondecreasing on $[r_z, s_z]$, so the lower ratio is minimized at $b = r_z$. Since $b \mapsto \min\{m_z, b\}/b$ is nonincreasing, the upper ratio is also maximized at $b = r_z$. This yields (21).

Both endpoints are attainable by allocating the factual mass F across the three disjoint events I^*, $T^* \setminus I^*$, and $\neg T^*$, whose probabilities are γ_z, $\tau_z - \gamma_z$, and $1 - \tau_z$. Hence the bounds are sharp. □

Intuitively, the bounds quantify how much factual mass can be assigned to the candidate explanation I^* among all cross-world couplings consistent with the supplied margins. Here r_z is the smallest forced overlap between the factual event and the desired counterfactual outcome, while ℓ_z and m_z are the smallest and largest possible overlaps with the candidate explanation. Thus the bounds tighten when the margins force more candidate-specific overlap or limit how much mass I^* can contain. The condition $r_z > 0$ guarantees a nonvanishing query denominator.

Notice that ϕ_z is observational, whereas τ_z and γ_z are counterfactual margins and must come from structural assumptions, a specified SCM and backtracking kernel, external information, a valid modified-value reduction, or sensitivity ranges.

C.2. Refinements

The marginal bounds deliberately allow every cross-world coupling compatible with the three specified margins. Thus, any additional information that rules out some of these couplings can sharpen the same query. Let

$$a_z := \mathbb{P}_z(F \cap I^*), \qquad \eta_z := \mathbb{P}_z(F \cap T^*),$$

so that the target is a_z/η_z, and define the margin-compatible region

$$\mathcal{B}_z := \{(a, b) : \ell_z \le a \le m_z,\ r_z \le b \le s_z,\ 0 \le a \le b,\ 0 \le b - a \le \tau_z - \gamma_z\}.$$

If additional information $\mathcal{I}$ can be translated into an exact restriction of the admissible pairs to $\mathcal{B}_z(\mathcal{I}) \subseteq \mathcal{B}_z$, then the corresponding sharp refinement is

$$\inf_{(a,b)\in\mathcal{B}_z(\mathcal{I}):\, b>0} \frac{a}{b} \leq \mathbb{P}(X^* = x^* \mid Y^* = y^*, H = h) \leq \sup_{(a,b)\in\mathcal{B}_z(\mathcal{I}):\, b>0} \frac{a}{b}. \tag{22}$$

Thus extra information can shrink the feasible numerator–denominator region, and the explanation probability is re-optimized over the smaller set. When $\mathcal{B}_z(\mathcal{I})$ is only a conservative outer restriction rather than an exact characterization, the same optimization yields valid but not necessarily sharp bounds. Importantly, joint restrictions should be imposed directly on (a, b) rather than converted into separate numerator and denominator bounds, which can unnecessarily lose sharpness.

Kernel restrictions. Information about the scope $\mathcal{E}$ and backtracking kernel K can rule out cross-world couplings that are allowed by the margins alone. If these restrictions can be translated into a set $\mathcal{B}_{\mathcal{E},K}(z) \subseteq \mathcal{B}_z$ of compatible numerator–denominator pairs, then (22) applies with $\mathcal{I} = (\mathcal{E}, K)$; it is sharp when that feasible set is characterized exactly. Intuitively, the kernel describes which counterfactual backgrounds are plausible given the factual one, and therefore constrains which numerator–denominator combinations can actually arise. The display therefore provides a general refinement template rather than a closed-form kernel-specific bound: a concrete kernel must first be translated into restrictions on (a, b). Such information may be available when the backtracking kernel is substantively specified as part of the model, calibrated from domain knowledge, or restricted through scientifically meaningful features such as locality or similarity.

Known denominator. A particularly useful intermediate case arises when the numerator remains unknown but

$$\eta_z = \mathbb{P}_z(F \cap T^*)$$

is known. In words, η_z is the probability, within the factual stratum $Z = z$, that the factual event F holds and the desired counterfactual outcome T^* is realized for the same unit. For instance, if Y indicates whether a credit application is approved and X indicates whether the applicant has high debt, then for $F = \{X = 1, Y = 0\}$ and $T^* = \{Y^* = 1\}$, η_z is the probability, among applicants with $Z = z$, of being factually high-debt and rejected while being approved in the counterfactual world.

If $0 < \eta_z \in [r_z, s_z]$ then the feasible numerator is exactly restricted to

$$\max\{0, \ell_z, \eta_z - (\tau_z - \gamma_z)\} \leq a_z \leq \min\{m_z, \eta_z\}.$$

Provided this interval is nonempty, the resulting sharp bounds are

$$p^{\eta}_{\mathrm{low},z}(x^*) = \frac{\max\{0, \ell_z, \eta_z - (\tau_z - \gamma_z)\}}{\eta_z}, \qquad p^{\eta}_{\mathrm{upp},z}(x^*) = \frac{\min\{m_z, \eta_z\}}{\eta_z}. \tag{23}$$

Knowing η_z removes uncertainty about the total factual–counterfactual overlap entering the query denominator; only its allocation between I^* and $T^* \setminus I^*$ remains unknown.

An exactly known denominator is stronger information than the marginal inputs and will not usually be directly available from observational data. Alternatively η_z can itself be treated as a sensitivity parameter whose variation measures how strongly the explanation depends on the factual–counterfactual overlap.

Bounded denominator. Suppose instead that one knows or assumes

$$\eta_z^- \le \eta_z \le \eta_z^+.$$

If $\mathcal{I}_z := [\eta_z^-, \eta_z^+] \cap [r_z, s_z]$ is nonempty and bounded away from zero, then

$$\inf_{b \in \mathcal{I}_z} \frac{\max\{0, \ell_z, b - (\tau_z - \gamma_z)\}}{b} \le \mathbb{P}(X^* = x^* \mid Y^* = y^*, H = h) \le \sup_{b \in \mathcal{I}_z} \frac{\min\{m_z, b\}}{b}.$$

By the same monotonicity argument as in Appendix C.1, the lower ratio is nondecreasing and the upper ratio is nonincreasing over the feasible denominator range, so both extrema are attained at the smallest feasible $b \in \mathcal{I}_z$.

This case is often more realistic than specifying η_z exactly: scientific knowledge may only constrain the fraction of factual units that could reach $Y^* = y^*$, or external information may provide an interval rather than a point value. The endpoints (η_z^-, η_z^+) therefore have a direct sensitivity-analysis interpretation and can be widened or narrowed to assess robustness.

C.3. Toy Bound Analysis

This subsection illustrates the partial-identification hierarchy with simple simulated cross-world laws using the credit example above. Let Y indicate whether a credit application is approved, X whether the applicant has high debt, and Z additional observed applicant information, such as an income category. For simplicity, we omit Z here, so that $\mathbb{P}_z = \mathbb{P}$. We take

$$F = \{X = 1, Y = 0\}$$

to represent applicants who are factually high-debt and rejected,

$$T^* = \{Y^* = 1\}$$

to represent counterfactual approval, and

$$I_j^* = \{X^* = j, Y^* = 1\}, \qquad j \in \{0, 1\},$$

to represent approval accompanied by counterfactual debt status j. Thus $\phi = \mathbb{P}(F)$, $\tau = \mathbb{P}(T^*)$, and $\gamma_j = \mathbb{P}(I_j^*)$. The examples do not estimate a full SCM; they only illustrate what the bounds imply once these margins are supplied. The same calculations apply within any fixed stratum $Z = z$.

A uniquely ranked MAP explanation. First, simulate $n = 10^6$ units with

$$\mathbb{I}(F) \sim \text{Bernoulli}(0.7), \qquad \mathbb{I}(T^*) \sim \text{Bernoulli}(0.6),$$

independently, and, among units satisfying T^*, set $X^* = 1$ with probability $11/12$ and $X^* = 0$ otherwise. The resulting estimates are

$$\widehat{\phi} = 0.700, \qquad \widehat{\tau} = 0.601, \qquad \widehat{\gamma}_1 = 0.551, \qquad \widehat{\gamma}_0 = 0.050.$$

Applying (21) gives

$$p_{\text{low}}(1) = 0.834, \qquad p_{\text{upp}}(1) = 1,$$

whereas

$$p_{\text{low}}(0) = 0, \qquad p_{\text{upp}}(0) = 0.166.$$

Hence

$$p_{\text{low}}(1) > p_{\text{upp}}(0),$$

so $X^* = 1$ is the unique MAP explanation for every cross-world coupling compatible with the supplied margins. In the credit example, among applicants who are factually high-debt and rejected but would be approved in the counterfactual world, the most likely accompanying debt state is necessarily still high debt. The corresponding simulated probabilities are approximately 0.917 for $X^* = 1$ and 0.083 for $X^* = 0$, although these values are not used to construct the bounds.

An inconclusive marginal analysis. Now simulate $n = 10^6$ units with

$$\mathbb{I}(F) \sim \text{Bernoulli}(0.65), \qquad \mathbb{I}(T^*) \sim \text{Bernoulli}(0.5),$$

independently, and, among units satisfying T^*, set $X^* = 1$ with probability $1/2$ and $X^* = 0$ otherwise. The estimated margins are

$$\widehat{\phi} = 0.650, \qquad \widehat{\tau} = 0.499, \qquad \widehat{\gamma}_1 = 0.250, \qquad \widehat{\gamma}_0 = 0.250.$$

For either debt state, (21) gives

$$p_{\text{low}}(j) = 0, \qquad p_{\text{upp}}(j) = 1, \qquad j \in \{0, 1\}.$$

Thus the margins alone cannot determine whether counterfactual approval is more likely to accompany high or low debt among the factually rejected applicants. In the simulated law both conditional probabilities are close to 0.5, but many other cross-world couplings share the same margins.

Refinement using denominator information. Finally, simulate $n = 10^6$ units with

$$\mathbb{I}(T^*) \sim \text{Bernoulli}(0.5), \qquad \mathbb{P}\{\mathbb{I}(F) = 1 \,|\, \mathbb{I}(T^*) = 1\} = 0.7, \qquad \mathbb{P}\{\mathbb{I}(F) = 1 \,|\, \mathbb{I}(T^*) = 0\} = 0.6,$$

and, among units satisfying T^*, set $X^* = 1$ with probability 0.6. The estimated margins are

$$\widehat{\phi} = 0.649, \qquad \widehat{\tau} = 0.499, \qquad \widehat{\gamma}_1 = 0.300.$$

Using only these margins, (21) gives the vacuous interval

$$0 \leq \mathbb{P}(I_1^* \,|\, F \cap T^*) \leq 1.$$

Thus the margins alone provide no information about the probability that a factually high-debt, rejected applicant who would instead be approved would also remain high-debt in that counterfactual world.

Suppose, however, that additional information determines

$$\eta = \mathbb{P}(F \cap T^*),$$

whose population value under the simulated cross-world law is 0.35 and whose Monte Carlo estimate is $\widehat{\eta} = 0.349$. Here η is the probability of being factually high-debt and rejected while being approved in the counterfactual world. The known-denominator refinement in (23) yields

$$0.430 \leq \mathbb{P}(I_1^* \mid F \cap T^*) \leq 0.860.$$

Indeed,

$$p^{\eta}_{\text{low}}(1) = \frac{\max\{0, \widehat{\phi} + \widehat{\gamma}_1 - 1, \widehat{\eta} - (\widehat{\tau} - \widehat{\gamma}_1)\}}{\widehat{\eta}} = 0.430,$$

whereas

$$p^{\eta}_{\text{upp}}(1) = \frac{\min\{\widehat{\phi}, \widehat{\gamma}_1, \widehat{\eta}\}}{\widehat{\eta}} = 0.860.$$

The true simulated value is approximately 0.6. Thus the explanation is only partially identified: among applicants who are factually high-debt and rejected but would be approved counterfactually, the probability that counterfactual approval is accompanied by remaining high-debt lies between roughly 43% and 86%. Hence high debt remains a plausible, and potentially dominant, counterfactual explanation state among approval worlds, but because its identified interval crosses 0.5, the available information does not determine whether high or low debt is the more likely explanation state.

Appendix D. Details for the Adult Benchmark Application

This appendix gives the structural specification, fitted learners and backtracking mechanism, and empirical analyses underlying Section 6. We focus on three questions: how scope changes a diagnostic explanation, when that explanation reduces exactly to a modified-value policy on observable features, and how much information is lost when the cross-world coupling is not specified.

D.1. Benchmark and structural specification

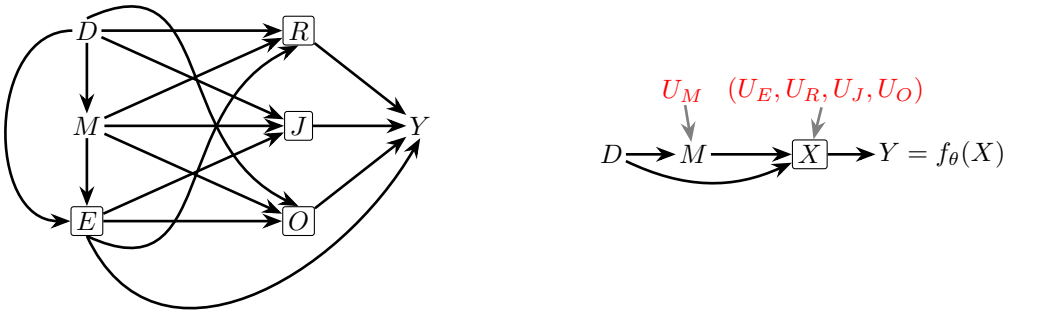
We use the UCI Adult benchmark (Becker and Kohavi, 1996), which contains 48,842 census-derived records with demographic and socioeconomic variables and is commonly used to predict whether annual income exceeds \$50K. Adult has also been widely used in the causal fairness and recourse literature, including by von Kügelgen et al. (2022). Its limitations as a substantive benchmark for fairness, however, are well documented (Ding et al., 2021). Accordingly, we use Adult here only for comparability with prior work and as a convenient fitted benchmark SCM for illustrating scope-restricted queries; we draw no substantive policy or fairness conclusions about income from the analysis.

We adopt the feature graph used in the Adult analysis of Plečko and Meinshausen (2020), whose causal structure has previously been studied and operationalized, and append a learned decision node Y; the resulting graph is displayed in Figure 3. After removing observations with missing values, the analysis contains 45,222 complete records, split into 31,655 training and 13,567 test observations. We define

$$Y = f_\theta(X), \qquad X = (E, R, J, O) = (\text{education}, \text{workclass}, \text{hours/week}, \text{occupation}),$$

Table 2: Learners used in the fitted Adult benchmark.

Component	Learner / representation
Numerical SCM node with parents	Random-forest regressor + empirical residual quantiles
Categorical SCM node with parents	Random-forest classifier + inverse-CDF latent ranks
Numerical SCM root	Empirical marginal distribution
Categorical SCM root	Empirical category probabilities
Decision $Y = f_\theta(X)$	Random-forest classifier



(a) Adult benchmark feature DAG

(b) Scoped revision and frontier $A = X$

Figure 3: Adult benchmark graph used in the application. Here S is sex, G age, N native country, M marital status, E education, R workclass, J hours per week, and O occupation; panel (a) groups the three root variables as $D = (S, G, N)$ to display the literature graph compactly. The explanatory vector is $X = (E, R, J, O)$ and the learned decision is $Y = f_\theta(X)$; boxed nodes indicate the frontier $A = X$. Thus, taking the candidate endogenous frontier in Section 4 to be $A = X$, every directed path from a scoped upstream revision to Y intersects the frontier. Panel (b) emphasizes that scopes are imposed on exogenous coordinates, while the corresponding modified-value representation acts on the endogenous variable X.

where f_θ is a random-forest classifier trained to predict the observed income label. Its test accuracy and ROC-AUC are 0.789 and 0.797, respectively; the observed positive rate in the test sample is 0.248, while the fitted classifier predicts a positive decision for 0.113 of test observations. The graph is used as a benchmark structural specification rather than as a claim that the Adult observational distribution identifies the underlying causal structure.

Table 2 lists the learners and distributional representations used in the fitted benchmark. All random forests use 200 trees and minimum leaf size 20. For f_θ, numerical components of X enter directly and categorical components are one-hot encoded; the binary decision is obtained by thresholding the fitted probability at 0.5.

Fitted structural model and backtracking kernel. We fit the structural assignments in the adopted DAG sequentially on the training sample. For a numerical non-root node, a random-forest regressor estimates its conditional mean given its parents, and the fitted structural disturbance is represented through the empirical quantile of the corresponding residual. Numerical roots instead use their empirical marginal distributions. For categorical non-root nodes, a random-forest classifier estimates conditional category probabilities and

Table 3: Scope-sensitive diagnostic explanations. Columns 2–4 give the mean probability of a successful revision across 200 factual negative cases. The final columns give equal-case averages of the case-specific explanation laws conditional on $Y^* = 1$ at $\rho = 0$.

	Success probability			Conditional on $Y^* = 1$, $\rho = 0$		
Scope	$\rho = 0$	$\rho = 0.5$	$\rho = 0.9$	$\mathbb{P}(E^* > E)$	$\mathbb{P}(J^* > J)$	$\mathbb{P}(O^* \neq O)$
$\mathcal{E}_{\text{edu}}$	0.076	0.058	0.032	0.996	0.766	0.633
$\mathcal{E}_X$	0.082	0.066	0.041	0.916	0.772	0.865
$\mathcal{E}_{\text{full}}$	0.089	0.076	0.051	0.915	0.797	0.877

the structural assignment is represented through the corresponding inverse conditional CDF; categorical roots use empirical marginal category probabilities.

For categorical nodes, the observed value identifies an interval of latent ranks rather than a unique U_V. Abduction therefore samples U_V uniformly within the fitted conditional-CDF interval associated with the observed category. A fixed lexicographic ordering of categories is used to define the inverse CDF. These choices are part of the fitted benchmark SCM rather than identification claims about the Adult population. As a numerical audit, abduction followed by regeneration reconstructed all eight modeled variables exactly on 1,500 held-out observations; for the numerical nodes, both the mean and maximum absolute reconstruction errors were zero.

For each scoped coordinate $U_V \in \mathcal{E}$, we use the Gaussian-copula sensitivity kernel

$$\zeta_V = \Phi^{-1}(U_V), \qquad \zeta_V^* = \rho\zeta_V + \sqrt{1-\rho^2}\,\epsilon_V, \qquad U_V^* = \Phi(\zeta_V^*),$$

where $\epsilon_V \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ independently across scoped coordinates and independently of the abducted factual ranks; unscoped coordinates satisfy $U_V^* = U_V$. Thus $\rho = 0$ gives an independent marginal redraw of the scoped disturbance, whereas $\rho \to 1$ approaches the factual latent background. We consider $\rho \in \{0, 0.5, 0.9\}$. The same abducted ranks and Gaussian innovations are reused across scope–ρ conditions to improve paired comparability.

Task 1: Scope-sensitive diagnostic explanations. Let $D = (S, G, N)$ denote the grouped demographic root variables, let $Z = (D, M) = (S, G, N, M)$ denote the remaining pre-X factual variables, and let $H = (X, Y, Z)$ denote the factual evidence. For individuals receiving the unfavorable prediction $Y = 0$, we study

$$p_{X^*}\big(x^* \,|\, Y^* = 1, H = h\big) \quad \text{under } (\mathcal{E}, \rho), \tag{24}$$

using Monte Carlo abduction, scoped revision, and structural propagation. The final analysis uses 200 factual negative cases and 1,000 revised backgrounds per case. We analyze

$$\mathcal{E}_{\text{edu}} = \{U_E\}, \qquad \mathcal{E}_X = \{U_E, U_R, U_J, U_O\}, \qquad \mathcal{E}_{\text{full}} = \mathcal{U}.$$

For each factual case, $\mathbb{P}(Y^* = 1 \mid H = h)$ is estimated by the proportion of its 1,000 revised worlds attaining $Y^* = 1$. The conditional explanation summaries are first computed within each case among successful draws and are then averaged equally over cases with at least one successful draw. Table 3 reports the resulting macro averages.

In these fitted results, broader scopes assign more probability to the favorable counterfactual at every value of ρ, whereas increasing ρ reduces this probability for every scope. More importantly, $\mathcal{E}_{\text{edu}}$ exhibits the specifically structural behavior motivating scope-restricted backtracking: at $\rho = 0$, successful case-specific explanation laws almost always satisfy $E^* > E$, but they also increase hours with probability 0.766 and change occupation with probability 0.633, despite U_J and U_O being pinned. Restricting which exogenous disturbances may change therefore does not freeze downstream endogenous variables.

The high-similarity regime also illustrates that conditional explanation summaries and success probability capture different aspects of the query. At $\rho = 0.9$, at least one successful Monte Carlo draw is observed for only 42.5% of factual cases under $\mathcal{E}_{\text{edu}}$, compared with 66.5% under $\mathcal{E}_X$ and 80.5% under $\mathcal{E}_{\text{full}}$. Thus the corresponding conditional summaries at large ρ describe the subset of case-specific laws for which the simulation observes positive success mass. Consistently with the increasing similarity, the mean distance between the factual and revised feature vectors among successful case-specific laws decreases under $\mathcal{E}_{\text{edu}}$ from 3.07 at $\rho = 0$ to 1.82 at $\rho = 0.9$. Here distance is the mixed-feature sum of standardized absolute differences over the numerical components of X and indicators of disagreement over its categorical components.

Task 2: From diagnostic explanation to recourse. The same construction gives a direct application of Proposition 2. Since

$$Y = f_\theta(X), \qquad X = (E, R, J, O),$$

taking the endogenous frontier to be $A = X$ intercepts every directed route through which a scoped upstream revision can affect Y. Moreover, for $C = (X, Y)$,

$$C^{\tilde{x}} = (\tilde{x}, f_\theta(\tilde{x})).$$

Although f_θ is fitted by random forest, once trained it is a fixed deterministic function of X. Hence $C^{\tilde{x}}$ is fixed once $\tilde{x}$ is specified and query–response compatibility holds automatically. The diagnostic law in (24) therefore admits the exact modified-value representation

$$p_{X^*}\big(x^* \,|\, Y^* = 1, H = h\big) = p_{X^\nu}\big(x^* \,|\, Y^\nu = 1, H = h\big),$$

where, because $H = (X, Y, Z)$, $X^\natural = X$, and the factual decision is deterministic given X, the generic policy ν_h of Proposition 2 simplifies here to

$$\nu(\cdot \,|\, z, x^\natural) = P_{X^*}\big(\cdot \,|\, Z = z, X^\natural = x^\natural\big).$$

Thus the same explanation generated through revisions of latent exogenous coordinates can be reproduced exactly by a randomized modified-value policy acting on observable X. The induced policy uses only pre-frontier Z and the natural frontier value $X^\natural$, giving the representation a prospective randomized-policy interpretation; interpreting it specifically as recourse additionally requires the substantive assumption that the components of X are jointly manipulable or actionable.

Table 4: Partial identification for $\mathbb{P}(E^* \geq 13 \mid Y = 0, Y^* = 1, S)$ under $\mathcal{E}_X$ and $\rho = 0$.

Factual sex	Fitted query	Marginal only	Known η
Female	0.949	$[0, 1]$	$[0.947, 1]$
Male	0.913	$[0, 1]$	$[0.895, 1]$

Task 3: Information supplied by the cross-world coupling. We finally ask how much of the diagnostic explanation remains if the fitted cross-world coupling is discarded. Within each factual-sex stratum, define

$$F = \{Y = 0\}, \qquad T^* = \{Y^* = 1\}, \qquad I^* = \{E^* \geq 13, Y^* = 1\},$$

so the fitted SCM and kernel determine the benchmark query under the fitted model

$$\mathbb{P}(E^* \geq 13 \,|\, Y = 0, Y^* = 1, S).$$

The partial-identification analysis uses 5,000 factual test observations and 10 revised worlds per observation. Within each factual-sex stratum we retain only

$$\phi = \mathbb{P}(F \,|\, S), \qquad \tau = \mathbb{P}(T^* \,|\, S), \qquad \gamma = \mathbb{P}(I^* \,|\, S),$$

and apply Proposition 4.

Across the three primary scopes, three values of ρ, and two factual-sex strata, there are 18 configurations. In 16 of them, $r_S = \max\{0, \phi + \tau - 1\} = 0$, so the margins permit a zero-probability conditioning event, the positivity hypothesis of Proposition 4 fails, and the marginal-only envelope is the trivial interval $[0, 1]$ rather than a sharp bound in the sense of that proposition. The only two non-vacuous marginal-only configurations occur for factual females under full backtracking:

$$\begin{aligned} \rho = 0 : &\qquad 0.925 \in [0.679, 1], \\ \rho = 0.5 : &\qquad 0.926 \in [0.435, 1]. \end{aligned}$$

Their widths are 0.321 and 0.565, respectively. Providing only the cross-world denominator

$$\eta = \mathbb{P}(F \cap T^* \,|\, S)$$

substantially sharpens the result. Across all 18 configurations, the median width of the refined interval is 0.122 and the mean width is 0.177. The median widths are 0.246, 0.108, and 0.104 under $\mathcal{E}_{\text{edu}}$, $\mathcal{E}_X$, and $\mathcal{E}_{\text{full}}$, respectively. Every refined interval contains the corresponding fully specified fitted-SCM benchmark, and all empirical Fréchet and subset-consistency checks are satisfied.

Table 4 illustrates the result under the X-local scope at $\rho = 0$. Although the marginal information alone is vacuous for both factual-sex strata, supplying η produces narrow intervals around the fitted benchmark.

The comparison makes explicit that the separate factual and counterfactual margins generally contain too little information even to guarantee positive mass for the conditioning event in this benchmark. Knowledge of its overlap probability removes this denominator ambiguity and can sharply contract the identified set, while the remaining width captures uncertainty about the unresolved cross-world association.