FACTUAL-ANCHORED COUNTERFACTUALS FOR CAUSAL-DIAGNOSTIC EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A counterfactual explanation of an observed case requires two choices: how alternative values are coupled to the values that actually occurred, and how much of the observed case the alternative world inherits downstream. We introduce *factual-anchored counterfactuals* (FACs), which encode these choices through natural-value-dependent laws. Scoped variables are revised by stochastic rules explicitly anchored at their realized factual values, while downstream propagation proceeds either through the original unit-level mechanisms (unit-persistent, u-FAC) or by sequential resampling from endogenous cross-world laws (resampling, r-FAC). We study diagnostic queries that condition simultaneously on the factual case and on a desired alternative outcome. Both variants admit the same Bayes decomposition into a factual-anchored revision law and a target-success score. However, they differ in their need for latent-background abduction. The u-FAC diagnostic law is generally not identified by the causal graph, revision law, and observed data. Conversely, r-FAC is point identified when its resampling laws are defined from the observed distribution, whilst giving up interventionist unit-level semantics. Finally, the revision law can be specified as an exponential tilt of the observed law, revealing connections between FAC-based explanations and feature-space, disturbance-based, and optimal-transport approaches to counterfactual explanation.

1 INTRODUCTION

An applicant whose loan application was denied asks an auditor a retrospective question: *given everything that actually happened in this case, what could plausibly have been different for the decision to have gone the other way?* Such *diagnostic* questions condition on two things at once, the factual record and a desired alternative outcome. These are the questions some explanation tools are intended to answer. For instance, feature-space methods answer them by searching for nearby inputs that flip a predictor, subject to plausibility or actionability constraints (Wachter et al., 2018; Ustun et al., 2019; Poyiadzi et al., 2020; Guidotti, 2024; Verma et al., 2024). Once causal structure is taken seriously, however, revising the explanatory variables is only half of the specification (Karimi et al., 2020; 2021; 2022; Beckers, 2022).

The other half is the cross-world law itself, and it involves two choices that are rarely separated. The first is the *coupling*: which alternative value is offered to *this* case. A method can be right about the population of alternative values and still attach the wrong alternative to the case at hand. The second is *downstream semantics*: does the alternative world retain the same unit-level disturbances, or are downstream variables resampled from cross-world laws anchored to the factual context?

Interventionist counterfactuals preserve the exogenous latent background across worlds while replacing the structural assignments at intervened variables, providing a same-unit semantics that generally requires abduction for factual-conditioned cross-world inference (Pearl, 2009; Balke & Pearl, 1994). Modified treatment policies, shift interventions and related natural-value-dependent interventions instead define intervened values as transformations of the values that would arise *naturally* under a specified regime (Haneuse & Rotnitzky, 2013; Young et al., 2014; Sani et al., 2020; Díaz et al., 2023), and have been developed for effect and policy evaluation rather than for retrospective explanation of an observed case. Backtracking counterfactuals take a different route: they preserve the structural assignments and generate cross-world variation by revising the exogenous background through a cross-world kernel (von Kügelgen et al., 2023; Kladny et al., 2024; Hao et al., 2024; de Aguas &

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Jullum, 2026). In each of these frameworks, the rule used to construct the alternative world also determines how that world propagates downstream.

We introduce *factual-anchored counterfactuals* (FACs) by decoupling the rule that generates the revision from the semantics governing its downstream propagation. At every variable selected for revision, the alternative value is drawn from a law explicitly anchored at its factual value. A *unit-persistent* FAC (u-FAC) then propagates the revision through the original causal mechanisms, whereas a *resampling-based* FAC (r-FAC) regenerates downstream counterfactual variables from endogenous cross-world laws conditional on their factual value, factual Markov pillow, and evolving counterfactual Markov pillow.

The choice between the two downstream propagation semantics becomes operational for diagnostic and explanatory queries of the form $P_{A^\star}(a^\star \mid Z = z, A = a, Y = y, Y^\star = y^\star)$, where A denotes the revised variable, Z the pre-revision factual information, and Y the outcome of interest, while starred ($\star$) variables and values refer to the alternative counterfactual world. Crucially, the downstream semantic choice governs the identification status of these queries: under r-FAC with observed-law-defined resampling kernels they can be expressed as functionals of the observed data law, whereas under u-FAC they generally require additional untestable cross-world assumptions. This echoes a familiar distinction in mediation analysis, where randomized-interventional analogues are introduced in place of natural effects involving cross-world counterfactuals in order to obtain identified targets (Didelez et al., 2006; VanderWeele et al., 2014; Naimi et al., 2014; Moreno-Betancur & Carlin, 2018).

We make the following contributions:

- **A semantics that separates revision from downstream propagation.** We define FACs for arbitrary revision scopes, with unit-persistent and resampling variants in semi-Markovian models. We prove that u-FAC is well posed and reproduces the factual world under a null revision, whereas r-FAC need not reproduce factual descendants because they are resampled. Both variants are general enough to recover hard and shift interventions on the treated.
- **A positioning result.** We show that factual anchoring is the retrospective counterpart of the recursive natural values used in longitudinal policies. Unlike the backtracking counterfactual framework, FACs couple alternative worlds at the *endogenous* level.
- **An identification dichotomy for diagnostic queries.** We derive a Bayes decomposition common to u-FAC and r-FAC that is explicitly modulated by the revision law. We provide identification functionals estimable from observed data and prove that the u-FAC diagnostic law is not a functional of the observed law, even under full causal knowledge.

2 Notation and Graphical Preliminaries

Let $V = (V_1, \dots, V_p)$ be random variables with joint law P_V, all taking values in standard Borel spaces. With a slight abuse of notation, we use $\mathrm{d}P_V$ both for the density/mass of P_V with respect to a dominating measure μ_{dom} and for the corresponding probability element times $\mathrm{d}\mu_{\mathrm{dom}}$. The law P_V is induced by a semi-Markovian *structural causal model* (SCM) (Pearl, 2009), so the observed causal graph $\mathcal{G}$ is an *acyclic directed mixed graph* (ADMG) (Richardson, 2003), with directed edges encoding observed causal ordering and bidirected edges representing latent common causes. We write $\mathrm{pa}(\cdot;\mathcal{G})$, $\mathrm{de}(\cdot;\mathcal{G})$, $\mathrm{an}(\cdot;\mathcal{G})$ for directed parents, descendants and ancestors, and $\mathcal{G}\{\!\{A\}\!\}$ for the subgraph induced on $A \subseteq V$. A *district*, or max-c-component, is a maximal set connected by bidirected paths (Tian & Pearl, 2002; Bhattacharya et al., 2022); we write $\mathrm{dis}(V_j;\mathcal{G}')$ for the district of V_j in $\mathcal{G}'$.

Fix a topological order $\prec$ compatible with the directed edges of $\mathcal{G}$, let $\mathrm{pre}_\prec(V_j) = \{V_k : V_k \prec V_j\}$ and $\mathcal{G}_{\preceq j} = \mathcal{G}\{\!\{\mathrm{pre}_\prec(V_j) \cup \{V_j\}\}\!\}$, and define

$$C_j := \mathrm{dis}(V_j;\mathcal{G}_{\preceq j}), \qquad M_j := \mathrm{mp}_\prec(V_j;\mathcal{G}) := (C_j \setminus \{V_j\}) \cup \mathrm{pa}(C_j;\mathcal{G}_{\preceq j}). \tag{1}$$

We call M_j the *Markov pillow* of V_j under $\prec$. Because $M_j \subseteq \mathrm{pre}_\prec(V_j)$, the ordered local Markov property for ADMGs (Richardson, 2003) gives a version of the observed conditional law satisfying $P_{V_j}(v_j \mid V_{\prec j} = v_{\prec j}) = P_{V_j}(v_j \mid M_j = m_j)$, so that P_V can be generated sequentially from the ordered laws $\{P_{V_j}(v_j \mid m_j)\}_{j=1}^p$. If $\mathcal{G}$ is a *directed acyclic graph* (DAG), then $C_j = \{V_j\}$ and $M_j = \mathrm{pa}(V_j;\mathcal{G})$, recovering the usual parent factorization (Pearl, 1988).

We use lowercase letters for realizations of uppercase random variables, δ for a Dirac measure, and reserve $\star$ for alternative-world quantities.

3 Factual-Anchored Counterfactuals

A Motivating Example. Consider the chain $X \to A \to Y$: X is an applicant's recorded financial history, A a revisable exposure such as the debt-to-income ratio, Y the decision, and the observed case (x, a, y) ends in denial. A diagnostic explanation asks: *relative to the ratio a that actually occurred, which alternative values $A^\star$ are plausible for this applicant and could have produced approval $Y^\star = y^\star$?* FACs encode the first part through a revision law $A^\star \sim \Gamma_{A^\star}(\cdot \mid A = a, X = x)$, in which the realized a is an explicit argument, so the explanation is attached to this case. The second part is a separate semantic choice. One may ask what would have happened to the *same applicant*, so that $Y^\star$ is generated by the original structural equation using the factual disturbance U_Y, which makes a strong claim about an unobserved quantity. Alternatively, $Y^\star$ may be coupled directly to $(Y, A, A^\star)$ through a resampling law $\Gamma_{Y^\star}(\cdot \mid Y, A, A^\star)$, in the simplest case by drawing $Y^\star$ from $P_Y(\cdot \mid A^\star)$, yielding a weaker claim stated entirely in observable terms. Panels (h) and (d) of Figure 1 depict the two readings; the definitions below extend the construction to arbitrary scopes in semi-Markovian graphs.

Definition 1 (Unit-persistent factual-anchored counterfactual law)**.** Let P_V, $\mathcal{G}$, $\prec$, and the Markov pillows M_j be as in Section 2, and let $\mathcal{M}$ be a semi-Markovian SCM compatible with $\mathcal{G}$ and inducing P_V, with structural assignments $V_j = f_{V_j}(\mathrm{pa}(V_j; \mathcal{G}), U_j)$ almost surely.

Let $\mathcal{S} \subseteq V$ be a *revision scope* and $\mathcal{D} := \mathcal{S} \cup \mathrm{de}(\mathcal{S}; \mathcal{G})$ its *affected region*. Variables outside $\mathcal{D}$ are shared across worlds, while each $V_j \in \mathcal{D}$ has an alternative counterpart $V_j^\star$. For $V_j \in \mathcal{D}$, define the *counterfactual Markov pillow* by replacing affected members of M_j with their starred counterparts and retaining unaffected members at their factual values:

$$M_j^\star := \{V_k^\star : V_k \in M_j \cap \mathcal{D}\} \cup \{V_k : V_k \in M_j \setminus \mathcal{D}\}. \tag{2}$$

For every scoped node $V_j \in \mathcal{S}$, let $\Gamma_{V_j^\star}(v_j^\star \mid v_j, m_j, m_j^\star)$ be a *revision law*: a regular conditional distribution specifying how its factual value may be stochastically revised as a function of factual and counterfactual context. For every affected descendant $V_j \in \mathcal{D} \setminus \mathcal{S}$, define the *counterfactual parent configuration*

$$R_j^\star := \{V_k^\star : V_k \in \mathrm{pa}(V_j; \mathcal{G}) \cap \mathcal{D}\} \cup \{V_k : V_k \in \mathrm{pa}(V_j; \mathcal{G}) \setminus \mathcal{D}\}. \tag{3}$$

The *unit-persistent factual-anchored counterfactual* (u-FAC) law $P^{\Gamma,\mathcal{M}}$ is the joint law of $(V, \mathcal{D}^\star)$ obtained by first drawing U from $\mathcal{M}$ and generating the factual world V, and then generating the starred affected region in the restriction of $\prec$ to $\mathcal{D}$ according to

$$\mathrm{d}P^{\Gamma,\mathcal{M}}_{V_j^\star}(v_j^\star \mid V, U, V^\star_{\prec j}) = \begin{cases} \mathrm{d}\Gamma_{V_j^\star}(v_j^\star \mid V_j, M_j, M_j^\star), & V_j \in \mathcal{S}, \\ \delta_{v_j^\star}(f_{V_j}(R_j^\star, U_j)), & V_j \in \mathcal{D} \setminus \mathcal{S}, \end{cases} \tag{4}$$

where all draws use randomization independent of (V, U). Revision thus enters only on the scope: every affected but unscoped descendant evolves through its original mechanism with the factual exogenous realization, preserving downstream unit-level variation and latent dependence.

One particular specification of Γ is a *discrepancy-regularized* family,

$$\mathrm{d}\Gamma_{V_j^\star}(v_j^\star \mid v_j, m_j, m_j^\star) \propto \mathrm{d}P_{V_j}(v_j^\star \mid m_j^\star) \exp\{-\beta\, \mathrm{dis}((v_j, m_j), (v_j^\star, m_j^\star))\}, \tag{5}$$

where $\mathrm{dis}(\cdot, \cdot)$ is a nonnegative cross-world discrepancy and $\beta \geq 0$ controls the strength of persistence toward the factual realization. The observed conditional law $P_{V_j}(\cdot \mid m_j^\star)$ provides a reference distribution that keeps candidate counterfactual values within the conditional support of the data, rather than allowing revisions in unconstrained regions of the state space. This reflects the motivation behind on-manifold counterfactual explanation (Dominguez-Olmedo et al., 2023; Tsiourvas et al., 2024). The exponential factor then reweights these candidate values according to their discrepancy from the factual configuration. This is analogous to backtracking kernels defined through distances in the exogenous-noise space (von Kügelgen et al., 2023).

Appendix B shows that this Gibbs specification can be written as an exponential tilt of the reference distribution, placing it within the broader class of stochastic interventions that reweight an observed law rather than replace it. The appendix also develops its close connections to other approaches for constructing counterfactual explanations, including **feature-space and disturbance-based explanations** and **optimal-transport explanations**.

Definition 2 (Resampling factual-anchored counterfactual law)**.** Retain the setup, revision scope $\mathcal{S}$, affected region $\mathcal{D}$, counterfactual Markov pillows $M_j^\star$, and scoped revision laws of Definition 1.

For each affected descendant $V_j \in \mathcal{D}\setminus\mathcal{S}$, let $\Gamma_{V_j^\star}(v_j^\star \mid v_j, m_j, m_j^\star)$ be a *resampling law*, specifying directly how $V_j^\star$ is regenerated after upstream revisions while conditioning on the factual realization V_j, the factual Markov pillow M_j, and the counterfactual Markov pillow $M_j^\star$.

The *resampling factual-anchored counterfactual* (r-FAC) law P^Γ is obtained by drawing $V \sim P_V$ and then generating the starred affected region sequentially under $\prec$ according to

$$P^\Gamma_{V_j^\star}(v_j^\star \mid V, V^\star_{\prec j}) = \Gamma_{V_j^\star}(v_j^\star \mid V_j, M_j, M_j^\star), \qquad V_j \in \mathcal{D}. \tag{6}$$

For $V_j \in \mathcal{S}$, $\Gamma_{V_j^\star}$ has a revision role. For $V_j \in \mathcal{D}\setminus\mathcal{S}$, it has a downstream resampling role. Unlike u-FAC, affected descendants reuse neither their factual disturbance nor their structural assignment: downstream propagation is specified directly through Γ, and no SCM enters the construction.

At the limiting choice $\beta = 0$ of equation 5, the discrepancy term disappears and the variable is simply resampled from its observed conditional law,

$$\mathrm{d}\Gamma_{V_j^\star}(v_j^\star \mid v_j, m_j, m_j^\star) := \mathrm{d}P_{V_j}(v_j^\star \mid m_j^\star), \qquad V_j \in \mathcal{D}\setminus\mathcal{S}. \tag{7}$$

This is the least persistent resampling completion in the family: it preserves the graphical ordering and evolving counterfactual context, but retains no unit-specific information from the factual realization.

Both constructions are well posed, but only u-FAC is consistent with the factual world under a null revision. Proofs are in Appendix A.

Proposition 1 (Well-posedness and null-revision consistency)**.** *Let Γ be a family of revision and resampling laws as above. Equations equation 4 and equation 6 define unique joint laws of $(V, U, \mathcal{D}^\star)$ and $(V, \mathcal{D}^\star)$, respectively, and under both constructions the marginal law of V is P_V.*

If $\mathrm{d}\Gamma_{V_j^\star} = \delta_{v_j}$ for every $V_j \in \mathcal{S}$, then $\mathcal{D}^\star = \mathcal{D}$ almost surely under u-FAC.

Under r-FAC the same null revision does not in general imply $\mathcal{D}^\star = \mathcal{D}$, because affected descendants are still regenerated from their resampling laws.

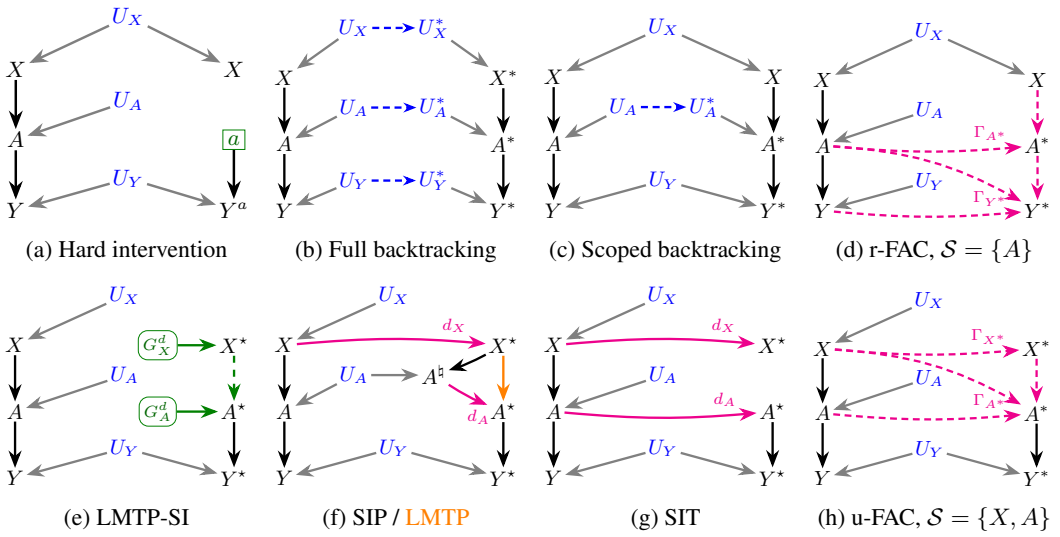
Proposition 2 (Reductions)**.** *Fix $\mathcal{S}$ and $\mathcal{D}$ as above, let $W := V \setminus \mathcal{D}$ collect the factual variables outside the affected region, and let $s^\star$ be a realized scoped revision.*

Suppose either that $\mathrm{d}\Gamma_{V_j^\star} = \delta_{s_j^\star}$ is constant for every $V_j \in \mathcal{S}$ under u-FAC, or that $\mathcal{G}$ is a DAG and every resampling law takes the form equation 7 under r-FAC. Then one obtains the ordinary conditional interventional law $P_{(\mathcal{D}\setminus\mathcal{S})^\star}(\cdot \mid \mathcal{S}^\star = s^\star, W = w) = P_{\mathcal{D}\setminus\mathcal{S}}(\cdot \mid \mathrm{do}(\mathcal{S} = s^\star), W = w)$. However, for a general ADMG, the r-FAC completion need not coincide with the law induced by any node intervention.

If instead $\mathrm{d}\Gamma_{V_j^\star} = \delta_{d_j(v_j, m_j)}$ for deterministic maps d_j that depend only on factual arguments, then u-FAC reproduces the shift-intervention-on-the-treated rollout of Sani et al. (2020).

Proposition 1 formalizes a basic distinction between the two downstream semantics. Under u-FAC, if the scoped variables are left unchanged, then the same structural mechanisms are evaluated with the same factual disturbances, so every affected downstream variable also reproduces its factual value. Under r-FAC, by contrast, affected descendants are regenerated from their resampling laws, and hence may differ from their factual values even when nothing is revised at the scope.

Proposition 2 shows that FACs reduce to familiar causal constructions under particular choices of the revision and resampling laws. In simple terms, if the scoped variables are fixed to externally chosen values, the resulting rollout coincides with an ordinary intervention: automatically for u-FAC, and for



(a) Hard intervention (b) Full backtracking (c) Scoped backtracking (d) r-FAC, $\mathcal{S} = \{A\}$

(e) LMTP-SI (f) SIP / LMTP (g) SIT (h) u-FAC, $\mathcal{S} = \{X, A\}$

Figure 1: Cross-world representations for the chain $X \to A \to Y$, with the factual world on the left. Solid arrows denote unit-level deterministic or structural relations, whereas broken arrows denote sampling or statistical cross-world dependence. **Top row**: only A is revised. Panel (a) shows a hard intervention, (b) full backtracking, (c) scoped backtracking restricted to U_A, and (d) r-FAC with $\mathcal{S} = \{A\}$, using Γ-laws for revision and downstream resampling. **Bottom row**: both X and A are revised. Panel (e) shows LMTP-SI, which samples from policies G^d; (f) SIP, which recursively modifies the natural value $A^\natural$ arising after the earlier modification of X, with the orange edge indicating the additional revised-history dependence allowed by a general LMTP; (g) SIT, which instead modifies each corresponding factual value; and (h) u-FAC, which samples the scoped variables through factual-anchored Γ-laws while propagating Y^* through the original unit-level SCM mechanism with shared U_Y.

r-FAC in a DAG when downstream variables are resampled from their observed conditional laws. If instead the scoped values are obtained by deterministic transformations of their own factual values, u-FAC recovers the shift-intervention-on-the-treated construction.

4 Positioning

Figure 1 summarizes the semantic contrasts. Hard interventions replace a target structural assignment while retaining the remaining mechanisms and exogenous background; soft or stochastic interventions replace a target mechanism or distribution (Pearl, 2009; Balke & Pearl, 1994; Díaz Muñoz & van der Laan, 2012; Correa & Bareinboim, 2020). Natural-value-dependent policies additionally use the value that would naturally occur as an input to the modification. *Modified treatment policies* (MTPs) operate on this natural treatment (Haneuse & Rotnitzky, 2013; Young et al., 2014). *Shift intervention policies* (SIPs) and *longitudinal MTPs* recursively define later natural values under the previously modified history (Sani et al., 2020; Díaz et al., 2023; Hoffman et al., 2024); *shift interventions on the treated* (SITs) stay anchored to the original factual values; and *stochastic LMTPs* (LMTP-SI) sample from policy-induced distributions with no unit-level pairing. Panels (e)–(g) display these distinctions.

Factual anchoring is the retrospective counterpart of these constructions. When several variables are scoped, FAC generates the factual vector and then rolls out the starred affected region under $\prec$, and at every scoped node the natural-value input remains the original factual V_j, even after upstream scoped variables have been revised. That is the anchoring idea of SIT in panel (g), but with stochastic rather than deterministic revision and with downstream propagation specified separately. Recursive natural values, as in SIPs and LMTPs, guide a prospective question, *what a policy would do going forward*; FAC guides a retrospective one, *what could have been different for this case*.

Backtracking changes a different object: it preserves endogenous structural assignments and couples the factual and counterfactual exogenous backgrounds through a kernel $\mathrm{d}B(U^* \mid U)$ (von Kügelgen et al., 2023), with deep and necessity-constrained variants (Kladny et al., 2024; Hao et al., 2024; Fatemi et al., 2025) and scope-restricted versions that fix which exogenous coordinates may be revised (de Aguas & Jullum, 2026). Panels (b) and (c) make this latent coupling explicit. FAC instead couples

worlds at the *endogenous* level: the scoped law $\Gamma_{V_j^\star}(\cdot \mid V_j, M_j, M_j^\star)$ records directly which observed factual values are paired with which alternatives. The distinction can vanish after marginalization: a distributional FAC may induce the same alternative-only law as an ordinary stochastic intervention. It remains consequential for diagnosis, however, because a marginal law does not generally identify the factual–alternative pairing (Proposition 6).

The two FAC variants then differ only in downstream propagation, and the trade-off they expose is a familiar one. In u-FAC, panel (h), unscoped descendants reuse their factual disturbances and original mechanisms, giving a same-unit reading at the cost of the abduction burden of cross-world quantities (Pearl, 2009; Avin et al., 2005; Shpitser & Pearl, 2008). In r-FAC, panel (d), downstream variables are sequentially resampled from endogenous cross-world laws, so once P_V and Γ are specified, no latent disturbance need be recovered, and by Proposition 1 consistency then holds only at the level of the resampling laws. Replacing a cross-world quantity by a randomized draw in order to obtain identification is what distinguishes randomized-interventional effects from natural direct and indirect effects in mediation analysis (Didelez et al., 2006; VanderWeele et al., 2014; Naimi et al., 2014; Moreno-Betancur & Carlin, 2018).

5 Diagnostic and Retrospective Explanation Queries

The distinction between the variants becomes consequential when the explanatory target is itself cross-world. Let $A \subseteq V$ be the revision scope, so $\mathcal{S} = A$, where A may be a single variable or an ordered collection; let Z collect observed factual information upstream of the scope, chosen to contain the factual arguments required by the scoped revision laws; and let $Y \in \mathrm{de}(A; \mathcal{G})$ be an outcome. Write $H = (Z, A, Y)$, $h = (z, a, y)$, and let $y^\star \neq y$ be a desired alternative outcome. Our target is the conditional law

$$\mathrm{d}P_{A^\star}(a^\star \mid H = h, Y^\star = y^\star), \tag{8}$$

which asks: *among alternative worlds compatible with the observed factual history h and the desired outcome $y^\star$, how are the candidate revisions of A relatively weighted?*

This is not a single-world intervention effect. It conditions simultaneously on factual information, including the realized $A = a$ and the outcome $Y = y$, and on a desired alternative outcome $Y^\star = y^\star$. In this sense, it follows the logic of the *probabilities of causation*, which are cross-world and typically only bounded by the observed law (Tian & Pearl, 2000; Halpern & Pearl, 2005).

5.1 A Common Bayes Decomposition

Write the revision scope as $\mathcal{S} = \{A_1, \ldots, A_m\}$, ordered so that $A_1 \prec \cdots \prec A_m$, and let $A = (A_1, \ldots, A_m)$. Two properties allow the scoped rollout to be expressed as a law of $a^\star$ given (z, a) alone. First, each factual pillow M_{A_r} is determined by (z, a), since Z is chosen to contain $\cup_r M_{A_r} \setminus \mathcal{S}$. Second, by equation 2, the counterfactual pillow $M^\star_{A_r}$ consists of starred coordinates from $M_{A_r} \cap \mathcal{D}$ and factual ones from $M_{A_r} \setminus \mathcal{D}$. The starred coordinates require no additional marginalization when

$$M_j \cap \mathcal{D} \subseteq \mathcal{S} \qquad \text{for every } V_j \in \mathcal{S}, \tag{9}$$

because every affected starred coordinate entering the revision law of a scoped variable is then itself a previously revised scoped variable.

Condition equation 9 holds automatically for a singleton scope. More generally, it holds whenever every affected variable entering the Markov pillow of a scoped variable is itself in the revision scope. For example, along the chain $A_k \to A_\ell \to A_m$, if both A_k and A_m are revised, then $A_\ell \in M_{A_m} \cap \mathcal{D}$, so condition equation 9 requires A_ℓ to be revised as well. By contrast, an affected variable that merely precedes a scoped variable but does not enter its Markov pillow need not belong to the scope.

Under this condition, no intermediate unscoped counterfactual variables need to be generated or marginalized out: the scoped rollout closes on $(z, a, a^\star)$, and the **revision law** is the ordered product

$$\mathrm{d}G_\Gamma(a^\star \mid z, a) = \prod_{r=1}^{m} \mathrm{d}\Gamma_{A_r^\star}\big(a_r^\star \mid a_r, m_{A_r}, m^\star_{A_r}\big), \qquad m^\star_{A_r} = m^\star_{A_r}(z, a^\star_{\prec r}), \tag{10}$$

which reduces to the single scoped law $\mathrm{d}\Gamma_A$ when A is a singleton. If condition equation 9 fails, some scoped law reads a starred unscoped descendant; that coordinate has to be integrated out,

and the result depends also on the factual values of those descendants under r-FAC and on the exogenous state U under u-FAC. Appendix A records this together with the conditional independence $A^\star \perp\!\!\!\perp (V, U) \mid (Z, A)$ implied by condition equation 9, which lets us condition on the factual outcome $Y = y$ without disturbing equation 10.

The second ingredient is the **target-success score** $\mathrm{d}P_{Y^\star}(y^\star \mid A^\star = a^\star, H = h)$: the chance that a proposed revision reaches the desired outcome once the downstream semantics has been applied. The two combine in the same way under either construction.

Proposition 3 (Bayes decomposition)**.** *Under condition equation 9, under Definition 1 and under Definition 2 alike,*

$$\mathrm{d}P_{A^\star}(a^\star \mid h, y^\star) \;\propto\; \mathrm{d}P_{Y^\star}(y^\star \mid A^\star = a^\star, H = h)\, \mathrm{d}G_\Gamma(a^\star \mid z, a), \tag{11}$$

where $\propto$ denotes proportionality up to a finite normalizing constant and $\mathrm{d}G_\Gamma$ is given by equation 10.

The decomposition separates *which revisions are plausible for the observed case* from *which of those revisions are likely to achieve the target*. The revision law G_Γ plays the first role: given the factual information (z, a), it assigns relative weight to the candidate values $a^\star$ according to the chosen factual-anchored revision mechanism, before imposing the requirement $Y^\star = y^\star$. The target-success score $P_{Y^\star}(y^\star \mid A^\star = a^\star, H = h)$ plays the second role: for each proposed revision $a^\star$, it measures how compatible that revision is with the desired counterfactual outcome after downstream propagation. Bayes' rule multiplies these two quantities and renormalizes, so the diagnostic law favors revisions that are simultaneously plausible departures from the factual case and effective at producing the target outcome. The decomposition is identical for u-FAC and r-FAC; what differs is how the target-success score is defined and evaluated under their respective downstream semantics.

5.2 u-FAC: Same-Unit Success Requires Abduction

Under u-FAC, once $A^\star = a^\star$ is drawn, every affected but unscoped descendant is propagated through its original structural assignment with the factual unit's exogenous realization. Let $\Phi_Y^{\mathcal{M}}(a^\star, z, u)$ denote the value of $Y^\star$ obtained by recursively evaluating the u-FAC rollout in $\mathcal{M}$ with scoped values fixed at $a^\star$ and exogenous state u.

Proposition 4 (Abduction representation)**.** *Under Definition 1 and condition equation 9,*

$$\mathrm{d}P_{Y^\star}^{\Gamma,\mathcal{M}}(y^\star \mid A^\star = a^\star, H = h) = \int \delta_{y^\star}(\Phi_Y^{\mathcal{M}}(a^\star, z, u))\, \mathrm{d}P_U^{\mathcal{M}}(u \mid H = h), \tag{12}$$

and equation 11 holds with this score.

Equation equation 12 is an abduction, revision and prediction calculation (Pearl, 2009; Balke & Pearl, 1994): infer the posterior of the factual unit's disturbances $P_U^{\mathcal{M}}(u \mid H = h)$, revise A through G_Γ, and propagate the same disturbances downstream. The posterior is the problem. Observationally indistinguishable SCMs can agree on P_V while inducing different posteriors, and the disagreement changes which revision an explanation query would recommend.

Theorem 1 (The u-FAC diagnostic law is not identified)**.** *There exist a causal graph $\mathcal{G}$, a revision scope A, a revision law Γ, a factual case h, and two semi-Markovian SCMs $\mathcal{M}_1, \mathcal{M}_2$ compatible with $\mathcal{G}$ and inducing the* same *observed law P_V, such that $\mathrm{d}P_{A^\star}^{\Gamma,\mathcal{M}_1}(\cdot \mid h, y^\star) \neq \mathrm{d}P_{A^\star}^{\Gamma,\mathcal{M}_2}(\cdot \mid h, y^\star)$. Consequently no functional of $(\mathcal{G}, P_V, \Gamma)$ equals the u-FAC diagnostic law for all compatible SCMs.*

A non-identification example for Theorem 1 is given in Appendix A.6. In that example, the r-FAC answer coincides with the u-FAC answer under the compatible SCM in which the arm-specific potential responses are cross-world independent, so conditioning on the factual response does not alter the alternative-arm response distribution. Resampling semantics is therefore not assumption-free, but its assumptions are expressed through observable conditional laws and can be varied explicitly in sensitivity analyses. When the SCM itself is available, for example through a known or learned invertible generative mechanism (Pawlowski et al., 2020; Nasr-Esfahany et al., 2023), the u-FAC

formulation may be preferable. Otherwise, quantities such as equation 12 should be accompanied by partial-identification bounds (Tian & Pearl, 2000).

5.3 r-FAC: Endogenous Completion Without Latent Abduction

Where u-FAC relies on the exogenous posterior in Proposition 4, leaving the diagnostic law unidentified from $(\mathcal{G}, P_V, \Gamma)$ alone by Theorem 1, r-FAC abandons the interventionist same-unit interpretation in favor of a construction defined without reference to the underlying SCM $\mathcal{M}$. This does not make every r-FAC law learnable from P_V, since a scientifically chosen resampling law may still encode cross-world restrictions that no experiment could reveal. The gain is modularity: r-FAC states those restrictions at the endogenous level, where they can be inspected and varied, and where families built from observable conditionals make them estimable rather than assumed. Under r-FAC, downstream variables are resampled directly through the endogenous Γ-laws.

Write $C := \mathcal{D} \setminus A$ for the affected descendants and $L := V \setminus (Z \cup A \cup \{Y\})$ for the factual variables not included in H, and let

$$P_C^{\Gamma}(c^\star \mid v, a^\star) := \prod_{V_j \in C}^{\prec} \Gamma_{V_j^\star}(v_j^\star \mid v_j, m_j, m_j^\star) \tag{13}$$

be the sequential starred completion law induced by Definition 2, with the scoped values fixed at $a^\star$.

Proposition 5 (r-FAC identification). *Under Definition 2 and condition equation 9, the target-success score is then*

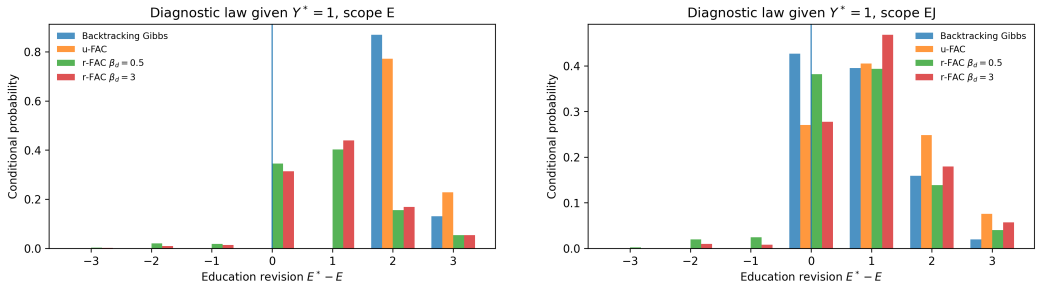
$$\mathrm{d}P^{\Gamma}_{Y^\star}(y^\star \mid A^\star = a^\star, H = h) = \int \mathrm{d}P^{\Gamma}_{Y^\star}(y^\star \mid h, \ell, a^\star)\, \mathrm{d}P_{L|H}(\ell \mid h), \tag{14}$$

and equation 11 holds with this score; no posterior over exogenous variables appears, and the outer integration is unnecessary when H contains all factual arguments of the resampling laws.

If in addition every $\Gamma_{V_j^\star}$, $V_j \in C$, is a functional of P_V, as for the families equation 5 and equation 7, then, subject to positivity of the conditioning event and of the normalizing constant, the diagnostic law equation 8 is a point-identified functional of $(P_V, \mathcal{G}, G_\Gamma)$.

Computation. Proposition 5 yields a direct simulation-based recipe for evaluating the r-FAC diagnostic law. Given fitted observable-law components and a candidate revision mechanism G_Γ, each proposed $a^\star$ is completed forward through the starred system, scored according to its induced probability of the target event, and then reweighted according to the Bayes decomposition in equation 11. Concretely:

1. Fit the conditional law $P_{L|H}$ whenever the factual variables in L are not already contained in H, together with all observable laws required to construct the chosen resampling kernels $\Gamma_{V_j^\star}$.
2. For the observed diagnostic state $(Z = z, A = a)$, draw candidate revisions
$$a^\star \sim G_\Gamma(\cdot \mid z, a).$$
3. For each candidate $a^\star$, if needed draw
$$\ell \sim P_{L|H}(\cdot \mid h),$$
and reconstruct the corresponding factual arguments entering the downstream resampling laws.
4. Generate the starred descendants sequentially in the order $\prec$ using
$$V_j^\star \sim \Gamma_{V_j^\star}\big(\cdot \mid v_j, m_j, m_j^\star\big), \qquad V_j \in C,$$
thereby producing a draw from the completion law $P_C^{\Gamma}(\cdot \mid v, a^\star)$.
5. Estimate the candidate-specific target-success score
$$P^{\Gamma}_{Y^\star}(y^\star \mid A^\star = a^\star, H = h)$$
by Monte Carlo averaging over the draws in the previous steps, or by direct evaluation when the relevant conditional law is available analytically.
6. Combine the resulting scores with the candidate-generation law $G_\Gamma(a^\star \mid z, a)$ according to equation 11, and normalize over $a^\star$ to obtain the diagnostic distribution.



(a) Scope $\mathcal{S}_E = \{E\}$.

(b) Scope $\mathcal{S}_{EJ} = \{E, J\}$.

Figure 2: Diagnostic laws of the education revision centered at the factual value ($E^\star - E$) and conditional on $Y^\star = 1$ (high class). Under the singleton scope in panel (a), downstream propagation sharply changes which revisions survive conditioning on the target. Once hours/week is also revisable, the education distributions become similar. See Figure 3 in Appendix D.4 for diagnostic distributions of the hours/week revision.

6 Application: Adult Benchmark

We illustrate FACs on the UCI Adult benchmark (Becker & Kohavi, 1996). Following the causal graph in Plečko & Meinshausen (2020), let $D = (S, G, N)$ collect sex, age, and native country, let M be marital status, and set $X = (E, R, J, O)$ for education level, workclass, hours/week, and occupation. We fit random forest causal mechanisms $D \xrightarrow{f_\theta} M$, $(D, M) \xrightarrow{f_\theta} E$, $(D, M, E) \xrightarrow{f_\theta} (R, J, O)$, with $Y = f_\theta(X)$ modeling income class as the outcome. We detail the causal modeling in Appendix D.

For the target $Y^\star = 1$ (high class), we estimate the diagnostic law under two scopes, $\mathcal{S}_E = \{E\}$ and $\mathcal{S}_{EJ} = \{E, J\}$. We compare scope-matched backtracking with a Gibbs kernel on abducted latent ranks, u-FAC with the endogenous exponential-tilt revision law in equation 5, and two r-FAC completions using the same revision law but downstream tilt parameters $\beta_d \in \{0.5, 3\}$. Thus the two r-FAC curves differ only in how strongly downstream resampling is anchored to the factual case.

The selected factual case lies almost exactly at the decision boundary, with fitted probability 0.499 for $Y = 1$. Figure 2 shows that, under the narrow scope $\mathcal{S}_E$, the diagnostic law depends strongly on downstream semantics. Every successful backtracking and u-FAC draw increases education, with mean revisions $E^\star - E = 2.13$ and 2.23, respectively, in units on the Adult education-number scale. In contrast, the two r-FAC laws require much smaller direct education changes, with means 0.81 and 0.90 for $\beta_d = 0.5$ and 3, and assign probabilities of 34.5% and 31.4% to $E^\star = E$ among successful worlds. The corresponding induced target probabilities are 2.95%, 5.58%, 22.14%, and 22.25%. Since u-FAC and r-FAC use the same revision law for $E^\star$, their different conditional distributions isolate the role of downstream propagation in the target-success factor of the diagnostic law.

When hours/week is also included in the scope, $\mathcal{S}_{EJ}$, the education distributions become substantially closer: the mean extra-education revisions are 0.77, 1.13, 0.72, and 0.97. Thus, allowing J to carry part of the revision reduces the dependence of the education explanation on downstream semantics. Differences remain in how the successful worlds are completed: Appendix D.4 shows that weaker r-FAC persistence produces substantially more downstream variation.

7 Discussion and Conclusion

FACs separate the ingredients of counterfactual explanations: the variables allowed to change, the stochastic rule pairing their factual values with alternative values, and the rule governing how downstream variables propagate across worlds. These choices are crucial for the identifiability of FAC diagnostic explanations. While u-FAC inherits the identification burden of cross-world quantities, r-FAC relocates the cross-world modeling choice to endogenous resampling laws that may be estimated without invoking interventionist semantics. Future work will focus on the remaining dependence of both FAC variants on scientifically chosen scopes and revision laws, as well as on the specification of the downstream resampling law for r-FACs. These choices are not determined by the observed data alone and should be justified through assumptions assessed by sensitivity analysis.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and with drafting the structure and wording of selected sections of the paper, refining aspects of the presentation, and literature search and synthesis. All AI-assisted material was reviewed, edited, and checked by the authors, including mathematical claims, code, empirical results, and cited sources. The authors take full responsibility for the final content of the work, including text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work is methodological and involves no new collection of human-subject data. The empirical illustration uses the publicly available UCI Adult benchmark, which contains demographic and socioeconomic attributes. The dataset is used solely to illustrate differences between counterfactual semantics; we draw no substantive causal, fairness, or policy conclusions about income or about individuals or demographic groups.

REPRODUCIBILITY STATEMENT

The supplementary material contains the files required to reproduce the example presented in the paper. The assumptions, definitions, and theoretical arguments underlying the proposed framework are provided in the main text and appendices. A dedicated public GitHub repository containing the corresponding code and reproducibility materials will be released upon acceptance.

AUTHOR CONTRIBUTIONS

Author contributions will be reported in the final version upon acceptance.

ACKNOWLEDGMENTS

Acknowledgments, including funding information where applicable, will be included in the final version upon acceptance.

REFERENCES

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 357–363. Morgan Kaufmann, 2005.

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

Barry Becker and Ronny Kohavi. Adult, 1996.

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

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23(295):1–76, 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, pp. 10093–10100, 2020. doi: 10.1609/aaai.v34i06.6567.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975. doi: 10.1214/aop/1176996454.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

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, pp. 36654–36662, 2026. doi: 10.1609/aaai.v40i43.40989.

Johan de Aguas and Martin Jullum. Scope-restricted backtracking counterfactuals. In *Proceedings of the 13th International Conference on Probabilistic Graphical Models*, 2026. URL https://openreview.net/forum?id=BEb4fBP6LI. To appear in Proceedings of Machine Learning Research; available on OpenReview.

Lucas de Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25, 2024.

Iván Díaz, Nicholas Williams, Katherine L. Hoffman, and Edward J. 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.

Iván Díaz Muñoz and Mark J. van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012. doi: 10.1111/j.1541-0420.2011.01685.x.

Vanessa Didelez, Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pp. 138–146. AUAI Press, 2006.

Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, Georgios Arvanitidis, and Bernhard Schölkopf. On data manifolds entailed by structural causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8188–8201. PMLR, 2023.

Fredrik Esscher. On the probability function in the collective theory of risk. *Skandinavisk Aktuarietidskrift*, 15(3):175–195, 1932.

Pouria Fatemi, Ehsan Sharifian, and Mohammad Hossein Yassaee. A new approach to backtracking counterfactual explanations: A unified causal framework for efficient model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16273–16285. PMLR, 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.

Joseph Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005. doi: 10.1093/bjps/axi148.

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.

Katherine L. Hoffman, Diego Salazar-Barreto, Nicholas T. Williams, Kara E. Rudolph, and Iván Díaz. Studying continuous, time-varying, and/or complex exposures using longitudinal modified treatment policies. *Epidemiology*, 35(5):667–675, 2024. doi: 10.1097/EDE.0000000000001764.

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, pp. 265–277, 2020.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445899.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. doi: 10.1145/3527848.

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.

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

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019. doi: 10.48550/arXiv.1912.03277.

Margarita Moreno-Betancur and John B. Carlin. Understanding interventional effects: A more natural approach to mediation analysis? *Epidemiology*, 29(5):614–617, 2018. doi: 10.1097/EDE.0000000000000866.

Ramaravind K. 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*, pp. 607–617. Association for Computing Machinery, 2020. doi: 10.1145/3351095.3372850.

Ashley I. Naimi, Jay S. Kaufman, and Richard F. MacLehose. Mediation misgivings: Ambiguous clinical and public health interpretations of natural direct and indirect effects. *International Journal of Epidemiology*, 43(5):1656–1661, 2014. doi: 10.1093/ije/dyu107.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25733–25754. PMLR, 2023.

Nick Pawlowski, Daniel Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, pp. 857–869, 2020.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

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*, pp. 344–350. Association for Computing Machinery, 2020. doi: 10.1145/3375627.3375850.

Thomas S. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003. doi: 10.1111/1467-9469.00323.

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

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1–4):287–313, 2000. doi: 10.1023/A:1018912507879.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pp. 567–573. AAAI Press, 2002.

Asterios Tsiourvas, Wei Sun, and Georgia Perakis. Manifold-aligned counterfactual explanations for neural networks. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3763–3771. PMLR, 2024.

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

Mark J. van der Laan, Alexander R. Luedtke, and Iván Díaz. Discussion of identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data, by jessica young, miguel hernán, and james robins. *Epidemiologic Methods*, 3 (1):21–31, 2014. doi: 10.1515/em-2014-0012.

Tyler VanderWeele, Stijn Vansteelandt, and James Robins. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*, 25(2):300–306, 2014. doi: 10.1097/EDE.0000000000000034.

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, Abdirisak Mohamed, and Sander Beckers. Backtracking counterfactuals. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 177–196. PMLR, 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.

Jessica G. Young, Miguel A. Hernán, and James M. Robins. Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic Methods*, 3(1):1–19, 2014. doi: 10.1515/em-2012-0001.

A Technical Results and Proofs

Throughout, all variables take values in standard Borel spaces, so regular conditional distributions exist. We repeatedly use the construction convention of Definitions 1 and 2: the starred affected region is generated after the factual world, sequentially in the restriction of $\prec$ to $\mathcal{D}$, and every Γ-draw uses randomization independent of (V, U) and of previous draws given its arguments.

A.1 Anchoring Independence

The following lemma is what licenses equation 10 and the use of G_Γ inside equation 11: conditioning on the factual outcome $Y = y$, which lies downstream of the scope, leaves the revision law untouched.

Lemma 1 (Anchoring independence)**.** *Assume condition equation 9 and that* $Z \supseteq \cup_r M_{A_r} \setminus A$. *Then, under either Definition 1 or Definition 2,* $A^\star \perp\!\!\!\perp (V, U) \mid (Z, A)$, *and* $\mathrm{d}P_{A^\star}(a^\star \mid H = h) = \mathrm{d}G_\Gamma(a^\star \mid z, a)$ *with* G_Γ *given by equation 10.*

Proof. Order the scope as $A_1 \prec \cdots \prec A_m$ and show by induction that $(A_1^\star, \ldots, A_r^\star) \perp\!\!\!\perp (V, U) \mid (Z, A)$ for every r. For $r = 1$, the law of $A_1^\star$ given (V, U) is $\Gamma_{A_1^\star}(\cdot \mid A_1, M_{A_1}, M_{A_1}^\star)$. Its factual arguments lie in (Z, A) by hypothesis, and by condition equation 9 the starred coordinates of $M_{A_1}^\star$ belong to $\mathcal{S}$ and precede A_1, so there are none; the remaining coordinates of $M_{A_1}^\star$ are factual members

of $M_{A_1} \setminus \mathcal{D}$, again functions of z. Since the randomization is independent of (V, U), the conditional law of $A_1^\star$ given (V, U) depends on (V, U) only through (Z, A). For the inductive step, the law of $A_{r+1}^\star$ given $(V, U, A_1^\star, \ldots, A_r^\star)$ is $\Gamma_{A_{r+1}^\star}(\cdot \mid A_{r+1}, M_{A_{r+1}}, M_{A_{r+1}}^\star)$, whose arguments are functions of (Z, A) and, again by condition equation 9, of the previously drawn scoped values $(A_1^\star, \ldots, A_r^\star)$ only. Combining with the inductive hypothesis gives the claim for $r + 1$, and the chain rule applied to this recursion is exactly equation 10. Since $H = (Z, A, Y)$ is a function of V, conditioning on $H = h$ gives $\mathrm{d}P_{A^\star}(a^\star \mid H = h) = \mathrm{d}G_\Gamma(a^\star \mid z, a)$. The argument uses only the construction convention and is therefore valid under both definitions. □

Remark 1 (When condition equation 9 fails). Suppose some scoped $V_j \in \mathcal{S}$ has $M_j \cap \mathcal{D} \not\subseteq \mathcal{S}$, so that $M_j^\star$ contains $V_k^\star$ for an unscoped affected $V_k \prec V_j$. That coordinate is generated before $V_j^\star$ and must be integrated out. Under r-FAC the integrand is $\Gamma_{V_k^\star}(\cdot \mid V_k, M_k, M_k^\star)$, so the resulting revision law depends on the factual value V_k, which need not belong to (Z, A); under u-FAC the integrand is the Dirac law at $f_{V_k}(R_k^\star, U_k)$, so the revision law depends on the exogenous state U. In both cases the conclusion of Lemma 1 fails, and equation 11 must be restated with the revision law conditioned on the additional factual information, or on U, that the scoped laws actually read.

A.2 Proof of Proposition 1

Proof. For the first claim, enumerate the affected variables as $\mathcal{D} = \{V_{j_1}, \ldots, V_{j_d}\}$ according to the fixed topological order, with $j_1 \prec \cdots \prec j_d$. At each step, the conditional law of $V_{j_r}^\star$ given the factual variables and the previously generated counterfactual variables is a regular conditional distribution. For a scoped node, this is $\Gamma_{V_j^\star}(\cdot \mid V_j, M_j, M_j^\star)$, whose arguments are measurable functions of the factual variables and of previously generated starred variables because $M_j^\star$ contains only counterfactual coordinates preceding V_j in the order. For an unscoped affected node, u-FAC uses the Dirac law induced by $f_{V_j}(R_j^\star, U_j)$, while r-FAC uses the corresponding resampling law $\Gamma_{V_j^\star}$. Hence finite sequential composition of these conditional laws with the factual base law determines a unique joint law. Equivalently, this follows from the Ionescu–Tulcea construction.

For u-FAC the base law is the joint law of (V, U) induced by $\mathcal{M}$, whereas for r-FAC the base law is P_V. In either case, all starred variables are subsequently generated conditionally on the factual variables. Integrating out the starred coordinates therefore leaves the factual marginal unchanged, so the marginal law of V is P_V under both constructions.

For the second claim, suppose that $\Gamma_{V_j^\star} = \delta_{v_j}$ for every $V_j \in \mathcal{S}$. We proceed inductively along the topological order restricted to $\mathcal{D}$. At every scoped node, $V_j^\star = V_j$ almost surely by construction. Now let $V_j \in \mathcal{D} \setminus \mathcal{S}$ and suppose that $V_k^\star = V_k$ almost surely for every affected variable $V_k \prec V_j$. Every affected parent of V_j therefore takes its factual value in the counterfactual world, while parents outside $\mathcal{D}$ are shared across worlds by construction. Consequently, $R_j^\star$ coincides with the factual parent configuration and

$$V_j^\star = f_{V_j}(R_j^\star, U_j) = f_{V_j}(R_j, U_j) = V_j \qquad \text{a.s.}$$

The induction yields $\mathcal{D}^\star = \mathcal{D}$ almost surely.

For the third claim, the same conclusion does not follow under r-FAC. The null revision forces $V_j^\star = V_j$ only at scoped nodes. An affected descendant $V_j \in \mathcal{D} \setminus \mathcal{S}$ is instead generated from $\Gamma_{V_j^\star}(\cdot \mid V_j, M_j, M_j^\star)$. In particular, consider the first unscoped affected node in the topological order for which resampling occurs. All affected variables preceding it have so far retained their factual values, and hence $M_j^\star = M_j$. Its conditional law is therefore $\Gamma_{V_j^\star}(\cdot \mid V_j, M_j)$, which need not be concentrated at V_j. Whenever this law assigns positive probability away from the factual value, the descendant may differ from its factual counterpart despite the null scoped revision. Thus a null revision of the scope does not in general imply $\mathcal{D}^\star = \mathcal{D}$ under r-FAC. Under the distributional choice equation 7 the law is $P_{V_j}(\cdot \mid M_j)$, which is degenerate at V_j only if V_j is a deterministic function of its Markov pillow. □

A.3 Proof of Proposition 2

Proof. First claim, u-FAC. Suppose that $\Gamma_{V_j^\star} = \delta_{s_j^\star}$ for every $V_j \in \mathcal{S}$. Then the u-FAC recursion fixes the scoped variables at $s^\star$ and, for every $V_j \in \mathcal{D} \setminus \mathcal{S}$, evaluates $V_j^\star = f_{V_j}(R_j^\star, U_j)$ using the

same exogenous realization as in the factual world. Pointwise in U, this is exactly the structural recursion obtained from $\mathcal{M}$ after the intervention $\mathrm{do}(\mathcal{S} = s^\star)$, so $(\mathcal{D} \setminus \mathcal{S})^\star = (\mathcal{D} \setminus \mathcal{S})(s^\star)$ almost surely. Because $W = \mathcal{V} \setminus \mathcal{D}$ contains no descendant of $\mathcal{S}$, the structural assignment of every variable in W is unaffected by the intervention. Conditioning on $W = w$ therefore gives

$$P_{(\mathcal{D}\setminus\mathcal{S})^\star}(\cdot \mid W = w) = P_{\mathcal{M}}((\mathcal{D} \setminus \mathcal{S})(s^\star) \in \cdot \mid W = w) = P_{\mathcal{D}\setminus\mathcal{S}}(\cdot \mid \mathrm{do}(\mathcal{S} = s^\star), W = w),$$

and the conditioning event $\mathcal{S}^\star = s^\star$ is almost sure under this revision law. Notice that the equality concerns conditioning only on pre-revision information. Conditioning instead on factual variables within the affected region would produce a genuinely cross-world interventionist conditional law and would generally require abduction, which is the content of Theorem 1.

First claim, r-FAC. Suppose that $\mathcal{G}$ is a DAG. Then the Markov pillow of each node reduces to its graphical parents, so $M_j^\star = \mathrm{pa}(V_j; \mathcal{G})^\star$. Fix a realized scoped revision $\mathcal{S}^\star = s^\star$ and factual values v_j for every $V_j \in W$. Under the distributional choice equation 7, each affected descendant is generated according to $\mathrm{d}P_{V_j}(v_j^\star \mid \mathrm{pa}(V_j; \mathcal{G})^\star)$, $V_j \in \mathcal{D} \setminus \mathcal{S}$, and in particular its law does not depend on the factual values of the affected descendants themselves. Consequently, conditional on $W = w$ and on $\mathcal{S}^\star = s^\star$, the r-FAC law of the affected descendants has density or probability element

$$\prod_{V_j \in \mathcal{D}\setminus\mathcal{S}} \mathrm{d}P_{V_j}\left(v_j^\star \mid \mathrm{pa}(V_j; \mathcal{G})^\star\right),$$

with the scoped coordinates fixed at $s^\star$ and all parent values outside $\mathcal{D}$ fixed at their factual values. Since $\mathcal{D} = \mathcal{S} \cup \mathrm{de}(\mathcal{S}; \mathcal{G})$, every variable in W is a non-descendant of $\mathcal{S}$. Moreover, no variable in $\mathcal{D}$ can be a parent of a variable in W, since that would make the latter a descendant of $\mathcal{S}$. Thus, after conditioning on the factual values of all variables in W, the truncated DAG factorization under $\mathrm{do}(\mathcal{S} = s^\star)$ reduces exactly to the product above, which proves the stated equality. For a general ADMG this argument no longer applies: ordered observed-law conditionals of the form $P_{V_j}(\cdot \mid M_j^\star)$ do not in general form the truncated causal factorization associated with a node intervention in the presence of latent confounding. Interventional distributions are instead governed, when identified, by the appropriate district-level or nested-factorization functionals (Tian & Pearl, 2002; Bhattacharya et al., 2022). Hence sequential resampling from the observed Markov-pillow conditionals still defines a valid r-FAC completion, but that completion need not coincide with the distribution generated by any ordinary node intervention.

Second claim. Suppose that $\Gamma_{V_j^\star} = \delta_{d_j(v_j, m_j)}$ for $V_j \in \mathcal{S}$, where each d_j depends only on factual arguments. Thus the value assigned to every scoped node is obtained from its naturally realized factual value and factual context, rather than from a value recursively recomputed after earlier scoped revisions. This is precisely the distinction defining a shift intervention on the treated in Sani et al. (2020). After the scoped values have been modified, u-FAC propagates them through the original structural assignments using the factual exogenous realization, so its recursive rollout agrees with the corresponding unit-level SIT counterfactual. □

A.4 Proof of Proposition 3

Proof. Work on the standard Borel space carrying $(H, A^\star, Y^\star)$ and let ν dominate the relevant conditional laws. For fixed h, the chain rule for regular conditional distributions factorizes the joint conditional element of $(A^\star, Y^\star)$ given $H = h$ as

$$\mathrm{d}P(a^\star, y^\star \mid h) = \mathrm{d}P_{A^\star}(a^\star \mid h)\, \mathrm{d}P_{Y^\star}(y^\star \mid A^\star = a^\star, H = h).$$

By Lemma 1, which applies under condition equation 9 to both constructions, $\mathrm{d}P_{A^\star}(a^\star \mid h) = \mathrm{d}G_\Gamma(a^\star \mid z, a)$ with G_Γ as in equation 10. Bayes' rule applied to this factorization gives equation 11, the omitted normalizing constant being the conditional element of $Y^\star = y^\star$ given $H = h$, which is positive and finite by assumption; the resulting object is a version of the conditional law of $A^\star$ given $(H, Y^\star) = (h, y^\star)$, unique up to null sets. For a target event $Y^\star \in T$ replace the score by $P(Y^\star \in T \mid A^\star = a^\star, H = h)$ throughout. □

A.5 Proof of Proposition 4

Proof. Under Definition 1, conditionally on $(U = u, A^\star = a^\star)$ and on the factual variables, every affected unscoped descendant is a deterministic function of $a^\star$, of the factual values of unaffected

variables, and of u; composing these assignments along $\prec$ gives $Y^\star = \Phi_Y^{\mathcal{M}}(a^\star, z, u)$ almost surely. Hence the score equals $\int \delta_{y^\star}(\Phi_Y^{\mathcal{M}}(a^\star, z, u))\, \mathrm{d}P_U^{\mathcal{M}}(u \mid A^\star = a^\star, H = h)$. By Lemma 1, $A^\star \perp\!\!\!\perp U \mid (Z, A)$ and hence $\mathrm{d}P_U^{\mathcal{M}}(u \mid A^\star = a^\star, H = h) = \mathrm{d}P_U^{\mathcal{M}}(u \mid H = h)$, which gives equation 12. The decomposition statement is Proposition 3 specialized to $P^{\Gamma,\mathcal{M}}$. □

A.6 Proof of Theorem 1

Proof. It suffices to exhibit one such configuration; we verify each claim. Let $\mathcal{G}$ be the DAG $A \to Y$ with $A \in \{0, 1, 2\}$, $Y \in \{0, 1\}$, $\mathcal{S} = A = \{A\}$, $Z = \emptyset$. Condition equation 9 holds trivially for a singleton scope. Both models take $A = f_A(U_A)$ with U_A uniform on $\{0, 1, 2\}$ and $U_A \perp\!\!\!\perp U_Y$, so P_A is uniform in both. Write the response function of Y as $t = (t(0), t(1), t(2)) \in \{0, 1\}^3$, that is, $Y = f_Y(A, U_Y) = t(A)$ with $t = t(U_Y)$; every semi-Markovian SCM on this graph with $U_A \perp\!\!\!\perp U_Y$ is equivalent to a distribution over t (Balke & Pearl, 1994).

Models. In $\mathcal{M}_1$, $t(0), t(1) \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and $t(2) \equiv 1$. In $\mathcal{M}_2$, $t(1) \sim \text{Bernoulli}(1/2)$, $t(0) = 1 - t(1)$, and $t(2) \equiv 1$.

Same observed law. Since $A \perp\!\!\!\perp t$, $P(Y = 1 \mid A = a) = E[t(a)]$. Under both models $E[t(0)] = E[t(1)] = 1/2$ and $E[t(2)] = 1$, and P_A is uniform in both, so the joint law $P_{A,Y}$ is the same; both models are compatible with $\mathcal{G}$.

Positivity. $P(A = 1, Y = 0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6} > 0$ under both.

Revision law. Let $\Gamma_{A^\star}(\cdot \mid A = 1)$ be uniform on $\{0, 2\}$; it is the same law in both models and it is factual-anchored, since it reads the realized $A = 1$.

Scores. With $h = (a, y) = (1, 0)$ and $y^\star = 1$, Proposition 4 gives a score equal to $P^{\mathcal{M}}(t(a^\star) = 1 \mid t(1) = 0)$, since $\{A = 1, Y = 0\} = \{A = 1\} \cap \{t(1) = 0\}$ and $A \perp\!\!\!\perp t$. Under $\mathcal{M}_1$, independence gives $P(t(0) = 1 \mid t(1) = 0) = 1/2$ and $P(t(2) = 1 \mid t(1) = 0) = 1$. Under $\mathcal{M}_2$, $t(1) = 0$ forces $t(0) = 1$, so the two scores are 1 and 1.

Diagnostic laws. By equation 11, $\mathrm{d}P_{A^\star}^{\Gamma,\mathcal{M}}(a^\star \mid h, y^\star) \propto P^{\mathcal{M}}(t(a^\star) = 1 \mid t(1) = 0) \cdot \frac{1}{2}$ on $\{0, 2\}$. Normalizing, $\mathcal{M}_1$ gives $(1/3, 2/3)$ and $\mathcal{M}_2$ gives $(1/2, 1/2)$. These differ, which proves the first claim. If some functional $\Psi(\mathcal{G}, P_V, \Gamma)$ returned the u-FAC diagnostic law for every compatible SCM, it would have to equal both values, a contradiction; hence no such functional exists.

r-FAC comparison. Under distributional r-FAC, $C = \mathcal{D} \setminus \mathcal{S} = \{Y\}$ and equation 7 gives a score equal to $P(Y = 1 \mid A = a^\star)$, which is $1/2$ at $a^\star = 0$ and 1 at $a^\star = 2$ in both models; the diagnostic law is $(1/3, 2/3)$, a functional of P_V alone, in agreement with Proposition 5. It coincides with the u-FAC law of the model $\mathcal{M}_1$, in which the arm-specific responses are cross-world independent. □

A.7 Proof of Proposition 5

Proof. By Definition 2, conditionally on $V = v$ and $A^\star = a^\star$ the starred affected region is generated by the sequential law equation 13, so $P_{Y^\star}^{\Gamma}(y^\star \mid A^\star = a^\star, V = v) = P_C^{\Gamma}(y^\star \mid v, a^\star)$, the marginal of equation 13 at the coordinate $Y^\star$. Averaging over the factual variables not contained in H and using Lemma 1, so that the conditional law of L given $(H, A^\star)$ equals $P_{L|H}$, yields equation 14. Substituting into Proposition 3 gives the stated form of the diagnostic law. If H contains all factual arguments of the resampling laws, then $P_{Y^\star}^{\Gamma}(y^\star \mid v, a^\star)$ is $\sigma(H)$-measurable and the outer integral is redundant.

For identification, note that each ingredient is a functional of $(P_V, \mathcal{G}, G_\Gamma)$: $P_{L|H}$ is a conditional of P_V; the ordered laws $P_{V_j}(\cdot \mid m_j^\star)$ entering equation 5–equation 7 are conditionals of P_V evaluated at the starred pillow values, with dis and β fixed by the analyst; the Markov pillows M_j are determined by $(\mathcal{G}, \prec)$; and G_Γ is specified. Hence the score, and therefore equation 11, is a fixed functional of $(P_V, \mathcal{G}, G_\Gamma)$ whenever the conditioning event has positive probability, or positive density, and the normalizing constant is positive and finite. Since no object in the construction refers to $\mathcal{M}$, any two SCMs inducing P_V give the same value, in contrast with Theorem 1. Support conditions enter only through the requirement that the conditionals $P_{V_j}(\cdot \mid m_j^\star)$ be well defined at the starred pillow values visited by the rollout; when G_Γ proposes revisions outside the conditional support of the data, the

functional is defined only up to that extrapolation, which is the on-manifold requirement discussed in Section 3. □

A.8 Marginal Insufficiency of the Alternative-Only Law

The diagnostic query also shows why FAC cannot be replaced by its alternative-only marginal. For a singleton A, define $\mathrm{d}\overline{G}_\Gamma(a^\star \mid z) := \int \mathrm{d}G_\Gamma(a^\star \mid z, a)\,\mathrm{d}P_A(a \mid z)$. Once the factual value is integrated out, $\overline{G}_\Gamma$ may coincide with an ordinary stochastic intervention law, as happens for several natural-value-dependent policies (Young et al., 2014; van der Laan et al., 2014). But equation 8 conditions on the observed $A = a$, and two couplings with identical marginals can assign completely different revisions to that value: Proposition 6 exhibits two revision laws with the same $\overline{G}_\Gamma$ whose diagnostic laws are point masses at different alternatives for the same case. What FAC retains is the joint law of $(A, A^\star)$ given $Z = z$ rather than the law of $A^\star$ alone, and equation 11 uses its conditional $G_\Gamma(\cdot \mid z, a)$ at the observed $A = a$.

Proposition 6 (Marginal insufficiency). *There exist two revision laws $G_\Gamma^1 \neq G_\Gamma^2$ with $\overline{G}_\Gamma^1 = \overline{G}_\Gamma^2$ whose diagnostic laws equation 11 differ for some factual case h with $P(H = h) > 0$, under either FAC variant and with all other components held fixed.*

Proof. Take $Z = \emptyset$ and the construction used in the proof of Theorem 1, so that $A \in \{0, 1, 2\}$ with P_A uniform, and fix any FAC completion whose score is strictly positive at $a^\star \in \{0, 2\}$ for the factual case $h = (a, y) = (1, 0)$ and target $y^\star = 1$; both $\mathcal{M}_1$ under u-FAC and the distributional r-FAC completion qualify, as computed in the proof of Theorem 1. Let

$$G_\Gamma^1(\cdot \mid a) = \delta_{(a+1) \bmod 3}, \qquad G_\Gamma^2(\cdot \mid a) = \delta_{(a+2) \bmod 3},$$

two deterministic factual-anchored revision laws. Both are bijections of $\{0, 1, 2\}$, so for each of them $\overline{G}_\Gamma(a^\star) = \sum_a G_\Gamma(a^\star \mid a) P_A(a) = 1/3$ for every $a^\star$, and the marginalized revision laws coincide. At the factual value $a = 1$, however, G_Γ^1 puts all its mass on $a^\star = 2$ and G_Γ^2 all its mass on $a^\star = 0$. Since the score is strictly positive at both points, equation 11 returns the point mass at $a^\star = 2$ under G_Γ^1 and the point mass at $a^\star = 0$ under G_Γ^2. These differ, and $P(H = h) = 1/6 > 0$. The construction used no property of the downstream semantics, so it applies to both variants. □

B Gibbs Revision Laws and Links to Other Explanation Methods

The Gibbs form in equation 5 is a well-motivated choice. Fix a node V_j, a factual configuration (v_j, m_j) and a counterfactual context $m_j^\star$, and ask for the law of $V_j^\star$ closest to the observed conditional $P_{V_j}(\cdot \mid m_j^\star)$ while keeping the counterfactual configuration near the factual one. Scoring the second requirement by the cross-world discrepancy on the pair $\big((v_j, m_j), (v_j^\star, m_j^\star)\big)$, the problem is to find Γ minimizing

$$\int \mathrm{dis}\big((v_j, m_j), (v_j^\star, m_j^\star)\big)\,\mathrm{d}\Gamma(v_j^\star \mid v_j, m_j, m_j^\star) + \beta^{-1}\,\mathbb{KL}\big(\Gamma(\cdot \mid v_j, m_j, m_j^\star) \,\|\, P_{V_j}(\cdot \mid m_j^\star)\big), \quad (15)$$

whose solution is exactly equation 5; Lemma 2 below gives the variational argument. The choice $\beta \to 0$ returns the reference and recovers the resampling law equation 7, while large β concentrates on configurations closest to the factual one. Thus, equation 5 is a cost-penalized information projection (Csiszár, 1975; de Aguas, 2026), an Esscher-type exponential tilt of the observed conditional law (Esscher, 1932): the support of $P_{V_j}(\cdot \mid m_j^\star)$ is preserved and only the relative weights change. Its natural neighbors are therefore the stochastic interventions that tilt an observed treatment law rather than replace it, as do *incremental propensity score interventions* (Kennedy, 2019).

B.1 Variational Characterization and Zero-Temperature Limit

We first record the variational characterization of the Gibbs family.

Lemma 2 (Cost-penalized information projection). *Let R be a reference law on a standard Borel space, $c \geq 0$ a measurable cost and $\theta > 0$ with $\int e^{-\theta c}\mathrm{d}R < \infty$. Then the unique minimizer over laws $G \ll R$ of $\int c\,\mathrm{d}G + \theta^{-1}\mathbb{KL}(G \,\|\, R)$ is $\mathrm{d}G^\star \propto \mathrm{d}R\, e^{-\theta c}$, with optimal value $-\theta^{-1} \log \int e^{-\theta c}\mathrm{d}R$.*

Two instances are used below. Taking $R = P_{V_j}(\cdot \mid m_j^\star)$, $c = \mathrm{dis}((v_j, m_j), (\cdot, m_j^\star))$ *and* $\theta = \beta$ *makes equation 5 the minimizer of equation 15; taking* $R = Q(\cdot \mid z)$, $c = c(a, \cdot)$ *and* $\theta = \lambda$ *gives equation 16.*

Proof. Let $Z_\theta := \int e^{-\theta c} \mathrm{d}R$ and $\mathrm{d}G^\star := Z_\theta^{-1} e^{-\theta c} \mathrm{d}R$. For any $G \ll R$ with finite objective, $\mathbb{KL}(G\|G^\star) = \mathbb{KL}(G\|R) + \theta \int c \, \mathrm{d}G + \log Z_\theta$, by substituting $\log(\mathrm{d}G^\star/\mathrm{d}R) = -\theta c - \log Z_\theta$. Dividing by θ gives $\int c \, \mathrm{d}G + \theta^{-1}\mathbb{KL}(G\|R) = \theta^{-1}\mathbb{KL}(G\|G^\star) - \theta^{-1} \log Z_\theta$, and $\mathbb{KL}(G\|G^\star) \geq 0$ with equality if and only if $G = G^\star$. □

Written at the level of the scope, with a general reference Q and tilt strength λ, the law of Lemma 2 reads

$$\mathrm{d}G_\Gamma(a^\star \mid z, a) \propto \mathrm{d}Q(a^\star \mid z) \exp\{-\lambda c(a, a^\star)\}, \tag{16}$$

and for a singleton scope this is exactly equation 5 under the dictionary $Q = P_A(\cdot \mid m_A^\star)$, $\lambda = \beta$ and $c = \mathrm{dis}$. The two statements below are proved for equation 16 and therefore apply verbatim to the revision laws of Appendix B.

Proposition 7 (Zero-temperature limit)**.** *Suppose* A *has countable support, the score is deterministic,* $\mathrm{d}P_{Y^\star}(y^\star \mid A^\star = a^\star, H = h) = \mathbb{I}\{a^\star \in \mathcal{A}^\star(h)\}$ *for a success set* $\mathcal{A}^\star(h)$*, and the revision law is equation 16. Then the modes of equation 11 are the minimizers over* $\mathcal{A}^\star(h)$ *of* $c(a, a^\star) - \lambda^{-1} \log \mathrm{d}Q(a^\star \mid z)$*. If moreover the infimum of* $c(a, \cdot)$ *over* $\mathcal{A}^\star(h)$ *is attained on a set of positive* $Q(\cdot \mid z)$*-mass, then as* $\lambda \to \infty$ *the diagnostic law concentrates on* $\arg\min_{a^\star \in \mathcal{A}^\star(h)} c(a, a^\star)$.

Proof. Substituting equation 16 and the indicator score into equation 11 gives, for $a^\star$ in the countable support,

$$P(a^\star \mid h, y^\star) = \frac{\mathbb{I}\{a^\star \in \mathcal{A}^\star(h)\} \, \mathrm{d}Q(a^\star \mid z) \, e^{-\lambda c(a, a^\star)}}{\sum_{\tilde{a} \in \mathcal{A}^\star(h)} \mathrm{d}Q(\tilde{a} \mid z) \, e^{-\lambda c(a, \tilde{a})}},$$

which is positive only on $\mathcal{A}^\star(h)$ and is maximized there by maximizing $\log \mathrm{d}Q(a^\star \mid z) - \lambda c(a, a^\star)$, that is by minimizing $c(a, a^\star) - \lambda^{-1} \log \mathrm{d}Q(a^\star \mid z)$; this is the objective of cost-minimizing counterfactual explanation with a plausibility penalty. Let $c^\star := \inf_{a^\star \in \mathcal{A}^\star(h)} c(a, a^\star)$ and $\mathcal{A}^{\star\star} := \{a^\star \in \mathcal{A}^\star(h) : c(a, a^\star) = c^\star\}$, nonempty with $Q(\mathcal{A}^{\star\star} \mid z) > 0$ by assumption. Multiplying numerator and denominator by $e^{\lambda c^\star}$, every term with $c(a, \tilde{a}) > c^\star$ carries a factor $e^{-\lambda(c(a,\tilde{a}) - c^\star)} \to 0$, while terms in $\mathcal{A}^{\star\star}$ are constant in λ; dominated convergence for the countable sums gives $P(\cdot \mid h, y^\star) \to Q(\cdot \mid z)$ restricted and renormalized to $\mathcal{A}^{\star\star}$, which is supported on the minimizers. If the minimizer is unique the limit is a point mass at it. □

B.2 Feature-Space and Disturbance-Based Explanations

Feature-space methods return a single nearest input that flips the predictor, subject to plausibility or actionability constraints (Wachter et al., 2018; Mothilal et al., 2020). They are related to the above family. Take a singleton scope, so that the revision law is $\Gamma_{A^\star}$ itself, and let the score be the indicator of a success set. Then the modes of equation 11 minimize $\mathrm{dis}\big((a, m_A), (a^\star, m_A^\star)\big) - \beta^{-1} \log \mathrm{d}P_A(a^\star \mid m_A^\star)$ over successful revisions, and as $\beta \to \infty$ the diagnostic law concentrates on the discrepancy-minimizing ones (Proposition 7 in Appendix B). A nearest-valid counterfactual is thus a point summary of a distributional object, with $-\beta^{-1} \log \mathrm{d}P_A$ the plausibility penalty such methods add by hand. Their causal counterparts revise features and re-evaluate the outcome with the disturbances recovered from the factual case (Karimi et al., 2020; Mahajan et al., 2019), which is u-FAC under a degenerate revision law.

B.3 Optimal-Transport Revision Couplings

Conditional on $Z = z$, a revision law induces the joint coupling $\gamma_{\Gamma,z}(\mathrm{d}a, \mathrm{d}a^\star) := P_A(\mathrm{d}a \mid z) \, G_\Gamma(\mathrm{d}a^\star \mid z, a)$, whose first marginal is the factual law of $A \mid Z = z$. Write $\Pi(P_A(\cdot \mid z), \rho_z)$ for the couplings with first marginal $P_A(\cdot \mid z)$ and second marginal ρ_z.

Proposition 8 (Transport couplings and local tilts)**.** *Fix* z *and an alternative marginal* ρ_z.

- *The revision laws whose induced coupling $\gamma_{\Gamma,z}$ has second marginal ρ_z are exactly the disintegrations of the elements of $\Pi(P_A(\cdot \mid z), \rho_z)$, up to $P_A(\cdot \mid z)$-null sets. Minimizing $\int c \, \mathrm{d}\gamma$ over that set therefore selects an optimal-transport revision coupling.*
- *The solution γ_ε of the entropic problem $\inf_{\gamma \in \Pi(P_A(\cdot|z),\rho_z)} \int c \, \mathrm{d}\gamma + \varepsilon \, \mathbb{KL}\big(\gamma \,\|\, P_A(\cdot \mid z) \otimes \rho_z\big)$ disintegrates into a revision law of the Gibbs form equation 16 with $\lambda = 1/\varepsilon$ and $\mathrm{d}Q(a^\star \mid z) \propto e^{g(a^\star)/\varepsilon} \mathrm{d}\rho_z(a^\star)$, where g is the second Schrödinger potential.*
- *Conversely, for a freely chosen reference Q the second marginal of $\gamma_{\Gamma,z}$ is whatever equation 16 induces. Matching a prescribed ρ_z requires solving the Schrödinger system for g, a global problem that a node-local specification such as equation 5 does not address.*

Proof. For the first claim, if the second marginal of $\gamma_{\Gamma,z}$ is ρ_z then $\gamma_{\Gamma,z} \in \Pi(P_A(\cdot \mid z), \rho_z)$ by construction. Conversely, any $\gamma \in \Pi(P_A(\cdot \mid z), \rho_z)$ on a standard Borel product space admits a disintegration $\gamma(\mathrm{d}a, \mathrm{d}a^\star) = P_A(\mathrm{d}a \mid z)\, G(\mathrm{d}a^\star \mid z, a)$ with G a regular conditional distribution, unique up to $P_A(\cdot \mid z)$-null sets; taking $G_\Gamma := G$ gives a factual-anchored revision law with the prescribed second marginal. The two sets therefore coincide up to factual-null modifications.

For the second claim, the entropic problem is strictly convex with a unique solution of the form $\mathrm{d}\gamma_\varepsilon(a, a^\star) = e^{\{f(a)+g(a^\star)-c(a,a^\star)\}/\varepsilon} \mathrm{d}P_A(a \mid z) \mathrm{d}\rho_z(a^\star)$ for potentials (f, g) solving the associated Schrödinger system (Cuturi, 2013; Peyré & Cuturi, 2019). Disintegrating with respect to the first marginal, the factor $e^{f(a)/\varepsilon}$ is absorbed into the normalization in a, leaving $\mathrm{d}\gamma_\varepsilon(a^\star \mid a) \propto e^{g(a^\star)/\varepsilon} e^{-c(a,a^\star)/\varepsilon} \mathrm{d}\rho_z(a^\star)$, which is equation 16 with $\lambda = 1/\varepsilon$ and Q as stated.

For the third claim, the second marginal induced by equation 16 is $\int \mathrm{d}G_\Gamma(a^\star \mid z, a) \, \mathrm{d}P_A(a \mid z)$, a determined functional of (Q, c, λ). Requiring it to equal a prescribed ρ_z is precisely the fixed-point condition defining the second Schrödinger potential, which couples all values of $a^\star$ simultaneously; a law specified node by node, as in equation 5, imposes no such condition. □

Relation to OT-based explanations. OT-based methods instead move the factual law onto a prescribed target law and read each case off the resulting map or plan (Cuturi, 2013; Peyré & Cuturi, 2019; de Lara et al., 2024). In FAC terms such a method is a choice of revision law, and Proposition 8 makes this exact: the revision laws whose coupling has second marginal ρ_z are precisely the disintegrations of $\Pi(P_A(\cdot \mid z), \rho_z)$, and the entropic plan turns into a tilt of exactly the shape equation 5, with the target marginal tilted by the second Sinkhorn potential in place of the observed conditional. The converse fails: a node-local tilt constrains no second marginal, so transport enters FAC only when one deliberately imposes ρ_z and solves for the potentials, which a local specification does not do.

Both families thus sit inside FAC as choices of revision law, leaving the downstream semantics as the separate decision the two variants make explicit (Appendix C). What factual anchoring adds is that the discrepancy reads the realized configuration, so the revision law is a cross-world coupling rather than a reweighting of a marginal law, a distinction that Proposition 6 shows is lost under marginalization.

C Comparison of the Two FAC Semantics

Table 1 summarizes the implications of the two downstream semantics for the diagnostic query equation 8, as established in Section 5.

D Adult Benchmark Application Details

This section gives the structural specification and simulation protocol underlying Section 6. Adult[1] is used only as a familiar benchmark on which a fitted causal model and a learned decision rule can be compared under controlled cross-world semantics; no causal, fairness, or policy conclusions about income are intended to be drawn from these results.

[1] UCI Adult: https://archive.ics.uci.edu/dataset/2/adult.

Table 1: Implications of the two FAC semantics for the diagnostic query in equation 8.

	u-FAC	**r-FAC**
Revision of A	Factual-anchored $\mathrm{d}G_\Gamma(a^\star \mid z, a)$	Factual-anchored $\mathrm{d}G_\Gamma(a^\star \mid z, a)$
Downstream propagation	Original structural assignments with the factual unit's disturbances	Endogenous resampling laws $\Gamma_{V_j^\star}$
Target-success score	Same-unit structural counterfactual probability equation 12	Sequential resampling probability equation 14
Consistency under a null revision	Same-unit: $\mathcal{D}^\star = \mathcal{D}$ a.s. (Prop. 1)	Resampling-law level only (Prop. 1)
SCM abduction	Generally required (Prop. 4)	Not required once the resampling laws are specified
Point identification from P_V	No in general, even with $\mathcal{G}$ and Γ known (Thm. 1)	Yes for observed-law-defined subclasses such as equation 7 (Prop. 5); otherwise the chosen Γ remains part of the estimand
Closest explanatory analogy	Feature revision with disturbance persistence; structural causal counterfactual explanation	Distributional and feature-space explanation with an explicit causal completion law
Role of tilting and transport	The tilt selects the revision coupling, but not the downstream structural response	The tilt selects the revision coupling and can also generate the resampling laws

D.1 Benchmark Graph and Fitted Models

After removing rows with missing values, the supplied script pools the complete records and uses a fixed stratified 70/30 train/test split, matching the earlier benchmark analysis. Write $D = (S, G, N)$ for sex, age, and native country; M for marital status; E for education number; R for workclass; J for hours/week; and O for occupation. The fitted benchmark graph is modeled as

$$D \xrightarrow{f_\theta} M, \qquad (D, M) \xrightarrow{f_\theta} E, \qquad (D, M, E) \xrightarrow{f_\theta} R, J, O,$$

and the outcome mechanism

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

Causal mechanisms $f_\theta(\cdot)$ are implemented as random forest models learned on the training data. Numerical targets are modeled using regression together with an empirical residual-rank representation of the exogenous latent variables. For categorical nodes, conditional probabilities are modeled using classification, with an inverse-CDF latent rank defining the node category. These invertible mechanistic representations allow for the abduction step needed for backtracking and u-FAC. The analysis selects, by a method-independent rule, the true-negative test case receiving the largest fitted probability below the 0.5 decision threshold among cases with $E \leq 13$.

D.2 Two Scopes and Four Cross-World Constructions

We compare

$$\mathcal{S}_E = \{E\}, \qquad \mathcal{S}_{EJ} = \{E, J\},$$

and condition on the complete factual case $H = h$ and the favorable alternative decision $Y^\star = 1$. For $\mathcal{S}_E$, the diagnostic target is the law of $E^\star$; for $\mathcal{S}_{EJ}$ it is the joint law of $(E^\star, J^\star)$. The plots show their marginal revision distributions.

Scope-matched backtracking. For each modeled node, the fitted latent rank $U_V \in (0, 1)$ is mapped to $Z_V = \Phi^{-1}(U_V)$. Abduction samples the factual rank from the fitted inverse-CDF interval compatible with the observed value. At scoped coordinates the Gibbs kernel is

$$\mathrm{d}K_{\beta_B}(z_V^\star \mid z_V) \propto \phi(z_V^\star) \exp\left\{-\frac{\beta_B}{2}(z_V^\star - z_V)^2\right\} \mathrm{d}z_V^\star, \tag{17}$$

with $\beta_B = 2$. Equivalently, $Z_V^\star \mid Z_V = z_V \sim \mathcal{N}(\beta_B z_V/(1+\beta_B), 1/(1+\beta_B))$. Unscoped latent ranks are pinned to their abducted factual values and the fitted SCM is regenerated forward.

u-FAC. At each scoped numerical node we use the fitted version of equation 5,

$$\mathrm{d}\Gamma_{V^\star}(v^\star \mid v, m, m^\star) \propto \mathrm{d}\widehat{P}_V(v^\star \mid m^\star) \exp\left\{-\frac{\beta_s}{2}\left(\frac{v^\star - v}{\widehat{s}_V}\right)^2\right\}, \tag{18}$$

with $\beta_s = 1.5$ and $\widehat{s}_V$ the training-sample standard deviation. Under $\mathcal{S}_{EJ}$, the reference law for $J^\star$ is evaluated at the already generated $E^\star$. All affected but unscoped descendants then reuse the latent ranks abducted from the factual case, implementing the unit-persistent semantics.

r-FAC. The scoped law is exactly equation 18; only downstream propagation changes. Numerical descendants use the same exponential-tilt form with a downstream parameter β_d, while categorical descendants use mismatch cost $\mathbb{I}\{v^\star \neq v\}$,

$$\Gamma_{V^\star}(v^\star \mid v, m, m^\star) \propto \widehat{P}_V(v^\star \mid m^\star) \exp\{-\beta_d \mathbb{I}(v^\star \neq v)\}.$$

We compare $\beta_d = 0.5$ and $\beta_d = 3$, representing weaker and stronger factual persistence in the endogenous completion. The final decision node is deterministic, $Y^\star = f_\theta(E^\star, R^\star, J^\star, O^\star)$, which is a degenerate r-FAC resampling law and keeps the explanatory target fixed across methods.

D.3 Monte Carlo Diagnostic Query and Reproducibility Outputs

For each scope–method pair, the script draws from the corresponding joint factual–alternative construction and retains the draws satisfying $Y^\star = 1$. These retained samples are Monte Carlo draws from the case-specific diagnostic law because the proposal mechanism is precisely the cross-world law before conditioning on the target event. The default run starts with 20,000 proposals per condition and continues, when needed, until at least 2,500 successful draws are obtained or 100,000 proposals have been generated.

Running `python app.py` creates the directory `adult_results/` with:

- `selected_factual_case.csv`, containing the factual case and classifier score;
- `adult_diagnostic_samples.csv.gz`, containing every Monte Carlo proposal and its success indicator;
- `adult_diagnostic_success_samples.csv`, containing the draws used for the conditional curves;
- `adult_diagnostic_summary.csv`, containing success rates and conditional revision summaries;
- `adult_diagnostic_E.png` and `adult_diagnostic_EJ.png`, used by the main-text figure once generated, plus supplementary hours/week and feature-distance plots.

D.4 Numerical Results

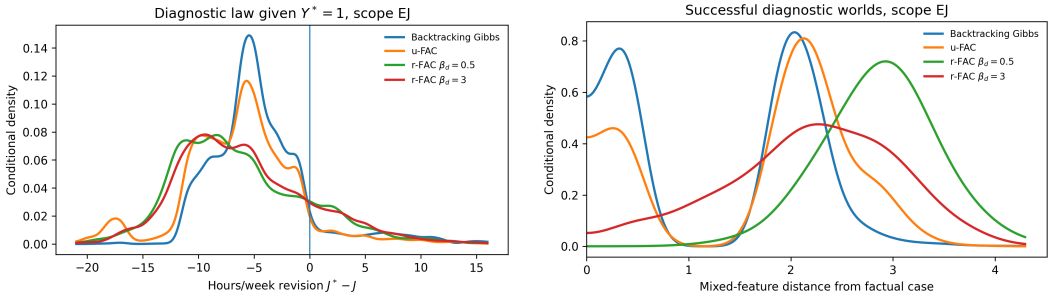
The complete-case data yield 31,655 training and 13,567 test observations. The fitted decision model has test accuracy 0.790 and ROC AUC 0.794. The fixed selection rule chooses a 29-year-old factual case with education number $E = 13$, corresponding to a bachelor's degree on the Adult ordinal education-number scale, local-government workclass R, $J = 56$ hours worked per week, and protective-service occupation O. The observed Adult income label is in the low-income class ($\leq$ \$50K per year), while the fitted probability of being classified into the high-income class ($>$ \$50K per year) is 0.499, placing the case close to the classifier boundary. In this application, the target event $\mathfrak{y} \equiv (Y^\star = 1)$ denotes that the counterfactual profile is classified into this high-income class.

Table 2 reports both the probability of reaching the target before conditioning and summaries of the resulting diagnostic law. Here $\Delta E := E^\star - E$ is measured in steps on the Adult education-number scale; for example, $\Delta E = 1$ means moving one category upward from the factual education code. Likewise, $\Delta J := J^\star - J$ is measured in hours worked per week, so $\Delta J = -5$ represents five fewer hours per week than in the factual case. The variables R and O denote the categorical workclass and occupation, respectively.

The quantity $d_{\text{feat}} = |E^\star - E|/\widehat{s}_E + |J^\star - J|/\widehat{s}_J + \mathbb{I}\{R^\star \neq R\} + \mathbb{I}\{O^\star \neq O\}$ is the dimensionless mixed-feature distance used in Figure 3. The two numerical coordinates, E and J, are standardized

Table 2: Monte Carlo results for the selected Adult case. Conditional means and change probabilities are computed only among draws satisfying $\mathfrak{y} \equiv (Y^\star = 1)$. Under scope $\mathcal{S}_E$, changes in J, R, and O arise entirely through downstream propagation.

Scope	Method	n_{succ}/n	$\widehat{P}(\mathfrak{y})$	$\mathbb{E}[\Delta E\|\mathfrak{y}]$	$\mathbb{E}[\Delta J\|\mathfrak{y}]$	$\mathbb{E}[d_{\text{feat}}\|\mathfrak{y}]$	$P(R^\star \neq R\|\mathfrak{y})$	$P(O^\star \neq O\|\mathfrak{y})$
$\mathcal{S}_E$	Backtracking	$2506/85k$	2.95%	2.13	6.87	2.41	0.0%	100.0%
	u-FAC	$2513/45k$	5.58%	2.23	6.77	2.44	0.0%	100.0%
	r-FAC, $\beta_d = 0.5$	$4427/20k$	22.14%	0.81	−7.02	2.99	91.94%	97.99%
	r-FAC, $\beta_d = 3$	$4449/20k$	22.25%	0.90	−4.64	2.06	45.38%	71.79%
$\mathcal{S}_{EJ}$	Backtracking	$3135/20k$	15.68%	0.77	−4.35	1.35	0.0%	57.35%
	u-FAC	$3440/20k$	17.20%	1.13	−6.23	1.74	0.0%	72.94%
	r-FAC, $\beta_d = 0.5$	$5039/20k$	25.20%	0.72	−6.25	2.87	92.22%	98.13%
	r-FAC, $\beta_d = 3$	$4023/20k$	20.12%	0.97	−5.88	2.20	45.54%	73.88%



(a) Conditional distribution of $J^\star - J$ under $\mathcal{S}_{EJ}$.

(b) Mixed-feature distance under $\mathcal{S}_{EJ}$.

Figure 3: Additional diagnostics for successful worlds under scope $\mathcal{S}_{EJ}$. Left: smoothed conditional density of the hours/week revision. Right: smoothed density of the mixed-feature distance from the factual case. Lower β_d permits substantially more variation in the r-FAC downstream completion, while larger β_d more strongly anchors resampling to the factual realization.

by their training-sample standard deviations, while a change in either categorical variable contributes one unit. Thus, d_{feat} summarizes overall departure from the factual feature profile rather than having a direct physical unit. The target rates in the table should not be read as predictive-performance scores: they quantify how much probability each specified cross-world construction assigns to classification into the high-income class, $Y^\star = 1$, before the diagnostic conditioning is imposed.

Under the education-only scope $\mathcal{S}_E$, the distinction between unit-persistent and resampling semantics is especially pronounced. Successful backtracking worlds increase education by either two or three steps on the Adult education-number scale, with 87.0% of the conditional mass at $\Delta E = E^\star - E = 2$; u-FAC behaves similarly, with 77.2% at $\Delta E = 2$. Their median education revision is therefore two education-scale steps in both cases. Since the factual value $E = 13$ corresponds to a bachelor's degree, a revision to $E^\star = 15$, for example, corresponds to the Adult category `Prof-school`; these differences are ordinal coding steps rather than years of schooling. Because J is not included in $\mathcal{S}_E$, changes in hours worked per week arise entirely through downstream propagation. Conditional on classification into the high-income class, backtracking and u-FAC yield mean increases in J of 6.87 and 6.77 hours per week, respectively, with a median increase of 7 hours per week for both.

The r-FAC diagnostic law is qualitatively different even though its scoped $E^\star$ revision kernel is exactly the same one used by u-FAC. For $\beta_d = 0.5$, 34.5% of successful worlds leave the education code unchanged, $\Delta E = 0$, and the median revision is only one education-scale step; for $\beta_d = 3$, the corresponding probability is 31.4%, again with median $\Delta E = 1$. Successful r-FAC worlds instead exploit variability in the downstream completion. In particular, their mean changes in hours worked per week are $\Delta J = -7.02$ and -4.64 hours for $\beta_d = 0.5$ and 3, respectively, while workclass changes in 91.9% and 45.4% of successful worlds. This provides a direct empirical illustration of the decomposition in equation 11: holding the scoped revision law fixed does not hold the diagnostic law fixed, because the downstream semantics modify the probability that each candidate revision results in classification into the target high-income class.

Expanding the scope to $\mathcal{S}_{EJ}$ changes this pattern. Education and hours worked per week can now both carry the explicit revision, and all four methods have a median education revision of $\Delta E = 1$, that is, one step upward on the Adult education-number scale. Backtracking places 42.6% of its successful mass at $E^\star = E$, corresponding to no education revision, compared with 27.1% for u-FAC, 38.2% for r-FAC with $\beta_d = 0.5$, and 27.7% for r-FAC with $\beta_d = 3$. The hours/week distributions in Figure 3a are predominantly negative for every construction: the median changes are -5, -6, -7, and -7 hours worked per week, respectively. For the factual value $J = 56$, a change of -7, for example, corresponds to a counterfactual value of $J^\star = 49$ hours per week. These shifts characterize the fitted, case-specific diagnostic law conditional on classification into the high-income class and should not be interpreted as claims about the causal effect of working fewer hours on income.

The downstream persistence parameter remains visible even after the scope is enlarged. With $\beta_d = 0.5$, r-FAC changes workclass and occupation in 92.2% and 98.1% of successful worlds, respectively, producing a mean dimensionless mixed-feature distance of 2.87. Increasing persistence to $\beta_d = 3$ lowers these probabilities to 45.5% and 73.9% and reduces the mean distance to 2.20. Figure 3b shows the same effect distributionally: the weakly persistent completion is concentrated farther from the factual feature profile, whereas the strongly persistent completion shifts substantial mass toward smaller distances. Backtracking and u-FAC exhibit a different distributional shape, including a distinct low-distance component, because their downstream variables are generated from the factual latent background rather than resampled from an endogenous tilted law.