Improving Causal Explanations

Kai-Zhan Lee
Causal AI Lab
Columbia University
kl2792@columbia.edu

Drago Plečko
Causal AI Lab
Columbia University
drago@stat.ucla.edu

Elias Bareinboim
Causal AI Lab
Columbia University
eb@cs.columbia.edu

Abstract

Explaining model predictions in high-stakes settings requires knowing *why* a prediction was made, a central challenge in modern AI. Yet existing methods lack a principled causal framework for explanation, leading to a disagreement problem: explanation methods produce conflicting attributions with no way to determine which is correct. Based on the literature in cognitive psychology, law, and philosophy, and using the language of causality, we formalize four desiderata for explanation methods that assign numerical credit to variables. We show that no prior explanation method including SHAP and LIME satisfies all desiderata, and prove the Explanatory Impossibility Theorem, which shows that even in principle, no method can satisfy all desiderata without causal assumptions. We introduce *counterfactual Shapley values*, the aggregation of a novel counterfactual quantity we call the *natural total effect*, and prove they satisfy all desiderata. We develop an algorithm for computing tight bounds on counterfactual Shapley values. Experiments corroborate our theory.

1 Introduction

Science has long relied on falsifiable mathematical models to explain data, from Kepler's planetary laws to Newton's mechanics. Modern causal inference continues this tradition with structural causal models (SCMs) [1]. A similar challenge arises in explainable AI (XAI), where demands for explainability are central to trust, autonomy, recourse, and debugging [2]. Yet XAI lacks a causally principled, computable definition of explanation [3–5], leading to a disagreement problem where different methods, including SHAP [6] and LIME [7], produce conflicting explanations with no principled basis for comparison [8, 9]. In this work, we address this gap with a formal framework grounded in SCMs.

One might consider using the full SCM itself as an explanation, but this approach faces two barriers. The first is epistemic: the *Causal Hierarchy Theorem* [10, Thm. 1] establishes that SCMs are almost never identifiable from observational and interventional data alone. The second is communicative: even if recoverable, SCMs are generally too complex for human interpretation. We therefore focus on explaining specific events, rather than summarizing or recovering the full data-generating process.

Following Lewis [11]'s view that "to explain an event is to provide some information about its causal history," we argue that **an explanation of an event is a summary of its causes**. This thesis raises two core questions: (1) What is a cause? and (2) How do we summarize causes well?

To address the first question, we illustrate the celebrated *but-for test*[1] of causation [12–14] with two examples. We provide causal diagrams and structural equations for these examples in Fig. 1.

Example 1.1 (Betting game)**.** Alice bets heads ($X = 1$) on Bob's coin flip and wins her bet ($Y = 1$).

[1] The but-for test is sufficient but not necessary for causation; see Sec. 2 for discussion.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: Two scenarios with identical observational and interventional distributions, where $\mathbf{1}[\cdot]$ denotes an indicator variable. The sampled settings are $U_X = U_Y = 1$ in both examples. Notably, X causes Y in Ex. 1.1, and not in Ex. 1.2. A good explanation method should distinguish these scenarios, or know it lacks the information to.

If Alice had not bet on heads, she would not have won ($Y_{X=0} = 0 \neq Y$); thus, the but-for test identifies her bet as a cause of her win. Alice has a half chance to win regardless of her bet; formally, $X \perp Y$ and $X \perp \{Y_{X=0}, Y_{X=1}\}$, where $Y_{X=x}$ denotes the outcome had X been set to x. □

Example 1.2 (Coin flips)**.** Alice and Bob each flip a fair coin; each coin lands heads ($X = 1$, $Y = 1$).

If Alice changed her coin flip to tails, Bob's flip would be unaffected ($Y_{X=0} = 1 = Y$); thus, the but-for test fails, and Alice's coin flip is not a cause of Bob's. Like in Ex. 1.1, Bob's coin flips are statistically independent from Alice's, and the observational and interventional distributions of X and Y are identical between the two examples. Thus, SHAP, LIME, and every prior method that relies on data alone (which is drawn from an observational or interventional distribution) fails to yield the correct answer in one of the two scenarios, because the correct answer is underdetermined by the data.[2] Distinguishing these scenarios requires one to reason about what would have happened under alternative scenarios. This kind of reasoning is encoded by counterfactuals in the SCM framework. □

To address the second question of how to summarize causes well, we focus on methods that assign a numeric attribution to each variable, which we call *explanatory variable attributions* (EVAs).[3] We organize prior methods by the level of causal information they require, following the Pearl Causal Hierarchy [10]: observational (L_1) methods [17, 7, 6, 18–20], interventional (L_2) methods [21–23], and counterfactual (L_3) methods [13, 24, 15, 16, 25]; see App. E for extended discussion. As shown in Examples 1.1–1.2, observational and interventional methods fail to distinguish causes from non-causes. Counterfactual methods can determine causes in theory, but in practice, existing methods have limitations: methods based on probabilities of causation [13, 24] or actual causation [26, 16] do not generally reflect effect magnitude, and CF-Shapley [25] assumes additive noise and the absence of unobserved confounders. No existing method quantitatively summarizes causes for SCMs under general assumptions.

We address this gap with three contributions (Fig. 2).

1. **Desiderata (Sec. 2):** We introduce four formal desiderata for causal explanations (Def. 2.4) grounded in philosophical, legal, and cognitive science literature. We show that no existing method satisfies all desiderata (Table 1), and that even in principle, no $L_{1:2}$ method can without causal assumptions (Thm. 2.5).
2. **Quantity (Sec. 3):** We introduce *counterfactual Shapley values* (L_3 SVs, Def. 3.2), the Shapley aggregation of a new counterfactual quantity we call the *natural total effect* (NTE, Def. 3.1). We prove L_3 SVs satisfy all desiderata (Thm. 3.3).
3. **Algorithm (Sec. 4):** We develop an algorithm that computes tight bounds on L_3 SVs from a causal diagram and $L_{1:2}$ data (Alg. 1, Thm. 4.2).

Experiments in Sec. 5 corroborate our theory. Extended discussion with examples appears in Apps. A and B. Experimental details and extended results appear in App. C. All proofs appear in App. D.

[2] Specifically, under the classifier $f(X, Y) = Y$, SHAP and LIME assign zero attribution to X in both scenarios: SHAP because $X \perp Y$, and LIME because f does not depend on X. However, these approaches clearly give a poor explanation in the betting game scenario, because Alice's bet (X) always causes her to win or to lose (Y).

[3] We chose the term EVA for two reasons: to unify feature attribution methods [7, 6] and counterfactual methods [13, 15, 16], so that we can compare them against formal desiderata; and to delineate these variable attributions from subset attributions (Sec. B.2.4).

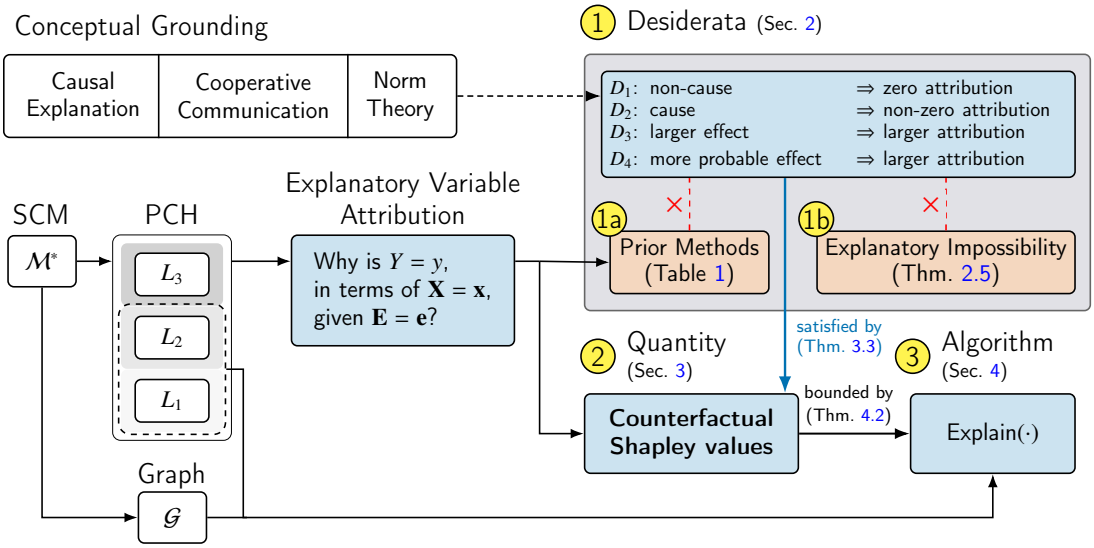


Figure 2: **Contribution overview.** (1) We introduce four formal desiderata $D_{1:4}$ for causal explanations; we show that (a) no prior method satisfies all four, and (b) no method based on observational and interventional distributions alone can. (2) We prove that counterfactual Shapley values satisfy all four. (3) We develop EXPLAIN (Sec. 4), an algorithm for bounding these values from a causal diagram and $L_{1:2}$ data. Blue: our contributions. Orange: prior methods and impossibility results.

Preliminaries. Capital letters X denote random variables, with lowercase letters x denoting values. Sets of random variables $\mathbf{X}$ are bolded; we write $x \in \text{Dom}(X)$ for a value x in the domain of X. A structural causal model (SCM) [1, 10] $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ contains endogenous variables $\mathbf{V}$, exogenous variables $\mathbf{U}$, functions $\mathcal{F}$ where each $V_i \leftarrow f_{V_i}(\mathbf{pa}(V_i), \mathbf{U}_{V_i})$ with parents $\mathbf{pa}(V_i) \subseteq \mathbf{V}$ and noise $\mathbf{U}_{V_i} \subseteq \mathbf{U}$, and probability measure $P(\mathbf{u})$. Each SCM has an associated causal diagram $\mathcal{G}$ [1, 10] where $V_i \to V_j$ if V_i is an argument of f_{V_j}, and $V_i \leftrightarrow V_j$ if the corresponding U_{V_i}, U_{V_j} are not independent. A *unit* $\mathbf{U} = \mathbf{u}$ is a specific realization of the exogenous variables. Potential outcome $Y_x(\mathbf{u})$ denotes the value of Y under intervention $X = x$ for unit $\mathbf{u}$, computed as the solution for Y in submodel $\mathcal{M}_x$ where all equations for X are replaced by $X = x$. We assume observations are generated by a *causal world* $(\mathcal{M}, \mathbf{u})$, a tuple consisting of an SCM $\mathcal{M}$ and specific unit $\mathbf{u}$ [1, Def. 7.1.8].

2 Framework

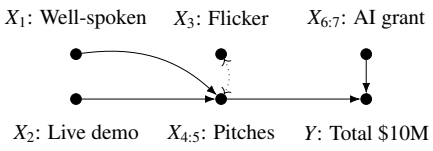
We study *explanatory variable attributions* (EVAs), functions assigning numerical attributions to variables explaining an outcome in a structural causal model (SCM). First, we formalize four causal desiderata for a good EVA, grounded in philosophy, law, and cognitive psychology (Sec. 2.1). Then, we show that no prior EVA satisfies all four desiderata, and prove the Explanatory Impossibility Theorem, which states that even in principle, no EVA can satisfy all four without counterfactual information (Sec. 2.2).

2.1 Desiderata

What makes a good explanation? In this section, we formalize four causal desiderata to answer this question. We first motivate our discussion with an example, illustrated in Fig. 3.

Example 2.1 (Startup funding)**.** Alice, a well-spoken entrepreneur ($X_1 = 1$), runs a live technical demo ($X_2 = 1$) during a startup pitch. Unexpectedly, the lights in the room flicker ($X_3 = 1$) while her laptop crashes mid-demo. Luckily, due to the quality of her pitch, Alice secures Series A ($X_4 = 1$) and angel funding ($X_5 = 1$) worth \$8M and \$1M, and a \$1M AI grant contingent on winning a competition ($X_6 = 1$) and passing a background check ($X_7 = 1$), for a total of $Y = 10$ (in \$M). Alice wonders, "Why did I receive \$10M?" □

To formalize Alice's question, we introduce the *why-query* $\text{Why}(Y \mid \mathbf{e}; \mathbf{X})$, composed of three parts:



$$U_i \sim \text{Bern}(0.05), \quad U_7 \sim \text{Bern}(0.95)$$
$$U_j^2 \sim \text{Bern}(0.05), \quad U_j^3 \sim \text{Bern}(0.95)$$
$$X_i = U_i \tag{1}$$
$$X_j = X_1 \wedge ((X_2 \oplus U_3) \vee U_j^2) \wedge U_j^3 \tag{2}$$
$$Y = 8X_4 + X_5 + X_6X_7 \tag{3}$$

Figure 3: Alice's startup funding scenario (Ex. 2.1). *Left:* causal diagram. *Right:* structural equations, where $i \in \{1, 2, 3, 6, 7\}$, $j \in \{4, 5\}$, and $\oplus$ denotes XOR. All U_i are exogenous; $X_{1:7}$ and Y are endogenous. $X_{4:5}$ and $X_{6:7}$ are clustered for concision.

(1) the event being explained, or *explanandum*, $Y = y$, inducing alternate values, or *foils*, $Y = y' \neq y$;
(2) the *evidence* $\mathbf{E} = \mathbf{e}$, the observed context used to infer causes; and
(3) the *explanatory variables* $\mathbf{X} \subseteq \mathbf{V}$, which prescribe the subset of variables to attribute to.[4]

We write $\text{Why}(y \mid \mathbf{e})$ as shorthand for $\text{Why}(Y{=}y \mid \mathbf{e}; \mathbf{X})$ when $\mathbf{X}$ is clear from context.

Given a why-query, an EVA assigns a numerical attribution to each explanatory variable in $\mathbf{X}$.

Definition 2.2 (Explanatory Variable Attribution)**.** An explanatory variable attribution (EVA) is a function $\phi : \Omega \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}$, where Ω is the space of SCMs over $\mathbf{V}$ and $\mathbb{W}$ is the space of why-queries $\text{Why}(Y \mid \mathbf{e}; \mathbf{X})$ asking why $Y = y \in \mathbf{e}$, given evidence $\mathbf{E} = \mathbf{e}$, in terms of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$. □

With $w = \text{Why}(Y{=}10 \mid X_{1:7}{=}1, Y{=}10;\ X_{1:7})$ as Alice's why-query, an EVA will assign attribution $\phi_{1:7}$ to each of $X_{1:7}$. What should those attributions track? Following the structural account of causality [1], we use functional dependence to quantify how much each variable causally influenced Y.

Definition 2.3 (Functional Dependence)**.** The *functional dependence* of Y on X is defined as

$$c(\mathcal{M}, \mathbf{u}, X, Y, b) = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}', x'}(\mathbf{u}), \tag{4}$$

given causal world $(\mathcal{M}, \mathbf{u})$, candidate cause X, outcome Y, and baseline $b = (x', \mathbf{z}')$, for adjustment set $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$ and baseline values $x' \in \text{Dom}(X)$ and $\mathbf{z}' \in \text{Dom}(\mathbf{Z})$. We abbreviate this expression as $c_{\mathcal{M}}(b)$ or $c(x', \mathbf{z}')$ where the rest is unambiguous. □

This quantity answers the question: "In the world $(\mathcal{M}_{\mathbf{z}'}, \mathbf{u})$, where we have set $\mathbf{Z} = \mathbf{z}'$, what is the effect on Y of allowing X to take its actual value $X = X_{\mathbf{z}'}(\mathbf{u})$, rather than setting it to baseline $X = x'$?" We argue that non-zero functional dependence under some baseline is *necessary* for causation: for X to be a cause of Y, changing X must change Y under *some* setting of other variables $\mathbf{Z} = \mathbf{z}'$.[5] Notably, when adjustment set $\mathbf{Z}$ is empty, non-zero functional dependence corresponds exactly to but-for causation: in the actual world $(\mathcal{M}, \mathbf{u})$, changing X will change Y. Following the literature, we argue that but-for causation is a *sufficient* condition for causation [27, 28] (see Sec. A.1.1 for discussion).

We now ask what makes a good explanation. We identify four qualitative desiderata grounded in the philosophy [11, 29], law [12], and cognitive psychology [30] literature (see Sec. A.3 for discussion).

D_1: **Admissibility:** non-causes receive zero attribution.
D_2: **Power:** causes receive non-zero attribution.
D_3: **Effect scaling:** variables with larger effects receive larger attribution.
D_4: **Normality:** variables with effects that are more likely to occur receive larger attribution.

In particular, D_4 follows norm theory as introduced in Kahneman and Miller [30]: causes that "could easily have been otherwise" should receive higher attribution. To encode this first notion, we introduce a distribution P_B over baselines $b = (x', \mathbf{z}')$, induced by $\mathbf{u}' \sim P(\mathbf{U})$ and a uniform distribution over the powerset of $\mathbf{X} \setminus \{X\}$.

Definition 2.4 (Explanatory Desiderata)**.** Given SCM $\mathcal{M} \in \Omega$ and EVA ϕ, we use three forms of functional dependence c: $c_{\mathcal{M}}(\cdot)$ as a function on baselines, $c_{\mathcal{M}}(b)$ as a scalar at fixed b, and $c_B := c_{\mathcal{M}}(B)$ as a random variable under $B \sim P_B$. Write $\mathbf{z}'_{\varnothing}$ for the but-for baseline $\mathbf{Z} = \varnothing$; δ_0 for the

[4] Explicitly selecting the set of explanatory variables as a subset of observed variables allows us to exclude irrelevant or unobserved causes such as "Alice arriving" or "the Big Bang", even though it is true that if Alice did not arrive, or the Big Bang did not occur, the funding outcome would be different!

[5] We denote functional dependence by c as a necessary condition for causation.

Dirac measure at zero; $P \equiv P'$ for distributional equality over $B \sim P_B$ and $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$; and $P_+^{\mathcal{M}}(b)$ for $\text{sign}(c_{\mathcal{M}}(b))\, P^{\mathcal{M}}(b)$.

$$
\begin{aligned}
&D_1\colon c_B \equiv \delta_0 && \implies \phi = 0\\
&D_2\colon P^{\mathcal{M}}(c_B \mid \mathbf{z}'_{\varnothing}) \not\equiv \delta_0 && \implies \phi \neq 0\\
&D_3\colon c_{\mathcal{M}} \geq c_{\mathcal{M}'} \wedge P^{\mathcal{M}}(B) \equiv P^{\mathcal{M}'}(B) \wedge P^{\mathcal{M}}(c_B|\mathbf{z}'_{\varnothing}) \not\equiv P^{\mathcal{M}'}(c_B|\mathbf{z}'_{\varnothing}) && \implies \phi_{\mathcal{M}} > \phi_{\mathcal{M}'}\\
&D_4\colon c_{\mathcal{M}} = c_{\mathcal{M}'} \wedge P_+^{\mathcal{M}}(B) \not\equiv P_+^{\mathcal{M}'}(B) \wedge P_+^{\mathcal{M}}(b) \geq P_+^{\mathcal{M}'}(b) && \implies \phi_{\mathcal{M}} > \phi_{\mathcal{M}'}
\end{aligned}
$$

All premises are universally quantified over all $\mathcal{M}, \mathcal{M}'$ sharing $\mathbf{V}$, $\mathbf{U}$ and $P(\mathbf{U} \mid \mathbf{e})$, all w, $X \in \mathbf{X}$, actual units $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$ (i.e., units consistent with the evidence), and baselines b in the support of P_B. □

In plain terms, D_1 requires that a variable with zero functional dependence for all baselines receives zero attribution.[6] D_2 requires that but-for causes (with $c(x', \varnothing) \neq 0$) receive nonzero attribution. D_3 and D_4 govern the magnitude of attributions rather than their sign: D_3 requires that among but-for causes with the same baseline distribution, larger effects yield larger attributions; D_4 requires that among variables with the same effects, those with more probable baselines yield larger attributions. We apply each to Alice's startup pitch to provide intuition for the desiderata:

(1) D_1 **(Admissibility):** D_1 requires $\phi_2 = \phi_3 = 0$ because the live demo (X_2) and lights flickering (X_3) are non-causes. X_3 has no directed path to Y (Fig. 3), so $c(x'_3, \mathbf{z}') = 0$ for all baselines. For X_2: since $X_3 = U_3$ (Eq. (1)), observing $X_3 = 1$ gives $U_3 = 1$ by abduction. Since $X_j = 1$ implies $U_j^{2,3} = 1$ by abduction, the demo term satisfies $(X'_2 \oplus U_3) \vee U_j^2 = (X'_2 \oplus 1) \vee 1 = 1$ for any counterfactual X'_2; thus $X_j = X_1 \wedge 1 \wedge 1 = X_1$ (Eq. (2)), independent of X'_2, so $c(x'_2, \mathbf{z}') = 0$ for all baselines.

(2) D_2 **(Power):** D_2 requires $\phi_1, \phi_4, \phi_5, \phi_6, \phi_7 \neq 0$ because eloquence (X_1) and funding sources ($X_{4:7}$) are but-for causes. For X_1: absent eloquence, pitch investors do not fund ($X_4 = X_5 = 0$), so $c(X_1{=}0, \varnothing) = 10 - 1 = 9 \neq 0$ (the grant $X_6X_7 = 1$ persists). Substituting each $X_k = 0$ into Eq. (3) gives $Y_{X_4=0} = 0 + 1 + 1 = 2$, $Y_{X_5=0} = 8 + 0 + 1 = 9$, and $Y_{X_6=0} = Y_{X_7=0} = 8 + 1 + 0 = 9$; thus, $c(X_4{=}0, \varnothing) = 8 \neq 0$ and $c(X_{5:7}{=}0, \varnothing) = 1 \neq 0$.

(3) D_3 **(Effect Scaling):** D_3 requires $\phi_4 > \phi_5$ because Series A (X_4, \$8M) has a larger causal effect than angel funding (X_5, \$1M), though losing either is equally likely. Substituting into Eq. (3) gives $Y_{X_4=0}(\mathbf{u}) = 0+1+1 = 2$ and $Y_{X_5=0}(\mathbf{u}) = 8+0+1 = 9$, so $c(X_4{=}0, \varnothing) = 8 > 1 = c(X_5{=}0, \varnothing)$.

(4) D_4 **(Normality):** D_4 requires $\phi_6 > \phi_7$ because winning the AI grant is a better cause of receiving the \$1M in funding than passing the background check: *not* winning the grant is far more probable than failing the background check, though both baselines would reduce funding by the same amount, \$1M. Substituting into Eq. (3) gives $Y_{X_6=0}(\mathbf{u}) = Y_{X_7=0}(\mathbf{u}) = 8 + 1 + 0 = 9$, so $c(X_6{=}0, \varnothing) = c(X_7{=}0, \varnothing) = 1$. But $P(X_6{=}0) = 0.95 \gg 0.05 = P(X_7{=}0)$: winning the grant "could easily have been otherwise" [30], while passing the background check could not.

Additional worked examples are provided in Sec. A.4.

2.2 Prior Methods

We conclude that no existing method satisfies all four desiderata; Table 1 shows the attributions each method assigns and which desiderata are violated; App. B provides formal counterexamples for each cell. One might surmise that a better $L_{1:2}$-based method could satisfy $D_{1:4}$; the next result shows this is not the case.

Theorem 2.5 (Explanatory Impossibility Theorem)**.** *There exists no EVA ϕ that satisfies $D_{1:4}$ for all $\mathcal{M} \in \Omega$ and depends only on $L_{1:2}$ data: i.e., there is no function g with $\phi_X(\mathcal{M}, w) = g(\mathcal{P}^{L_{1:2}}(\mathcal{M}), w, X)$ for all $X \in \mathbf{X}$, where $\mathbf{V}_{\mathbf{z}}$ denotes $\mathbf{V}$ under* $\text{do}(\mathbf{Z}{=}\mathbf{z})$ *and $\mathcal{P}^{L_{1:2}}(\mathcal{M}) = \{P^{\mathcal{M}}(\mathbf{V}_{\mathbf{z}}) : \mathbf{Z} \subseteq \mathbf{V}, \mathbf{z} \in \text{Dom}(\mathbf{Z})\}$.* □

Proof sketch. We construct two SCMs that induce identical $L_{1:2}$ distributions but differ in functional dependence: one in which X is a but-for cause of Y, requiring $\phi_X \neq 0$ by D_2; one in which X and Y are functionally independent, requiring $\phi_X = 0$ by D_1. Any $L_{1:2}$-only EVA must assign the same value in both, violating at least one of D_1, D_2. Full construction in Sec. D.3. □

[6] $D_{1:4}$ extend to arbitrary HP witness sets $\mathbf{W} \subseteq \mathbf{X} \setminus \{X\}$ in place of $\mathbf{z}'_{\varnothing}$ [28]; see Sec. A.1.1.

Table 1: No existing method satisfies all four desiderata on Ex. 2.1 (Def. 2.4). The *Expected* row indicates the behavior required by the desiderata; **bold** values match it. Violations of desiderata are indicated by ×; if no violation is found on this example, the cell is marked –. Extended results and discussion appear in App. B.

Method	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	D_1	D_2	D_3	D_4
Expected	$\neq 0$	0	0	$\neq 0; > \phi_5$	$\neq 0; < \phi_4$	$\neq 0; > \phi_7$	$\neq 0; < \phi_6$				
LIME (L_1)	0.00	**0.00**	**0.00**	**7.58**	**0.98**	**0.80**	0.00	–	×	–	–
SHAP (L_1)	**1.02**	–0.32	–0.32	**5.85**	**2.71**	**0.93**	**0.03**	×	–	–	–
ASV-prox (L_1)	0.00	**0.00**	0.33	**5.20**	**3.41**	**0.93**	**0.03**	×	×	–	–
do-SHAP (L_2)	**1.67**	1.27	**0.00**	**5.33**	**0.67**	**0.93**	**0.03**	×	–	–	–
PN(x, y) (L_3)	**1.00**	1.00	**0.00**	1.00	1.00	1.00	1.00	×	–	×	×
PN($x, y \mid \mathbf{x}, y$) (L_3)	**1.00**	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	–	×	×
CF-Shapley (L_3)	0.00	**0.00**	**0.00**	**8.00**	**1.00**	0.50	0.50	–	×	–	×
Beckers-CF (L_3)	0.00	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	×	×	×

Thus, we conclude that satisfying the four desiderata requires counterfactual information, and no existing method has yet succeeded in doing so. We address this gap in Sec. 3.

3 Counterfactual Shapley Values

Having established a conceptual and formal framework for evaluating explanatory methods, we now introduce counterfactual Shapley values (L_3 SVs), the first EVA to satisfy all desiderata. Our approach introduces the natural total effect (NTE), a novel counterfactual quantity extending the unit-level total effect [1, 31], and decomposes it using Shapley values [32] to assign per-variable credit.

The NTE measures how much the outcome changes when a set of variables takes its actual rather than baseline values for a specific unit.

Definition 3.1 (Natural Total Effect)**.** Given SCM $\mathcal{M}$, why-query $w = \text{Why}(y|\mathbf{e})$, and variable subset $\mathbf{Z} \subseteq \mathbf{X}$, the natural total effect (NTE) is defined for actual and baseline units $\mathbf{u}, \mathbf{u}' \in \text{Dom}(\mathbf{U})$ as:

$$\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \tag{5}$$

where $\mathbf{Z}(\mathbf{u})$ denotes the value of $\mathbf{Z}$ in unit $\mathbf{u}$. Here $Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u})$ denotes the outcome in unit $\mathbf{u}$ when $\mathbf{Z}$ is set to its value in $\mathbf{u}$, and $Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})$ denotes the outcome in unit $\mathbf{u}$ when $\mathbf{Z}$ is set to its value in the baseline unit $\mathbf{u}'$. □

The NTE captures functional dependence (Def. 2.3) through three design choices.[7] First, it extends unit-level total effects [1, 31] to multiple variables simultaneously [28]. Second, treatment selection $\mathbf{Z}(\mathbf{u})$ uses the observed actual world to capture realized causal effects. Third, natural baseline selection $\mathbf{Z}(\mathbf{u}')$ uses the exogenous distribution $\mathbf{u}' \sim P(\mathbf{U})$ to incorporate causal normality (D_4) and reflect the intuition that causes that "could easily have been otherwise" [30] are more salient explanations.

Since the actual unit $\mathbf{u}$ is unobserved, we condition on the observed evidence $\mathbf{e}$ it induced,[8] generalizing the intuition of the probability of necessity (PN) [13], which asks whether the outcome would have differed had the cause not occurred, to the multivariate setting (see Sec. A.6). When $\mathcal{M}$ is known, the conditional expectation below may be estimated via Monte Carlo sampling:

$$\text{NTE}(\mathbf{Z}, Y|\mathbf{e}) = \mathbb{E}_{\substack{\mathbf{u}\sim P(\mathbf{U}|\mathbf{e}) \\ \mathbf{u}'\sim P(\mathbf{U})}}[\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})] \tag{6}$$

For example, consider Alice's AI grant (X_6): with probability 0.95 the baseline takes the value of no grant ($X_6 = 0$), reducing funding from \$10M to \$9M for a unit-level NTE of 1; otherwise the baseline matches the actual value, giving an NTE of 0. The expected NTE for X_6 is therefore 0.95. The background check (X_7) yields an expected NTE of only 0.05, since its absent baseline has probability only 0.05. The gap of 0.90 between the NTEs reflects normality (D_4): the same causal effect receives higher attribution when its counterfactual baseline is more probable.

[7] We prove an equivalence between the NTE and functional dependence in Sec. A.7.

[8] This conditioning corresponds to local feature attribution in the XAI literature.

Algorithm 1 Bounding L_3 Shapley Values (Def. 3.2)

Input: Dataset $\mathcal{T}$, causal diagram $\mathcal{G}$, why-query w, variable X
Output: Bounds $[\phi^-, \phi^+]$ where $\phi^{\pm} = \phi_X^{L_3}(\hat{\mathcal{M}}^{\pm}, w)$
$\hat{\mathcal{M}}^- \leftarrow \arg\min_{\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T},\mathcal{G})} \phi_X^{L_3}(\mathcal{M}_\theta, w)$
$\hat{\mathcal{M}}^+ \leftarrow \arg\max_{\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T},\mathcal{G})} \phi_X^{L_3}(\mathcal{M}_\theta, w)$
Return $[\phi_X^{L_3}(\hat{\mathcal{M}}^-, w), \phi_X^{L_3}(\hat{\mathcal{M}}^+, w)]$

While NTEs capture functional dependence, computing them for all $2^{|\mathbf{X}|}$ subsets $\mathbf{Z} \subseteq \mathbf{X}$ is exponentially complex and impractical for individual attribution. Shapley values [32] summarize these per-subset NTEs into one attribution $\phi_X^{L_3}$ per variable.

Definition 3.2 (Counterfactual Shapley Value)**.** Given SCM $\mathcal{M}$, why-query $w = \text{Why}(y|\mathbf{e})$ and variable $X \in \mathbf{X}$, the counterfactual Shapley value (L_3 SV) is the Shapley value of the value function $v(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{e})$:

$$\phi_X^{L_3}(\mathcal{M}, w) = \mathbb{E}_{\pi \sim \text{Unif}(\Pi_{\mathbf{X}})}\left[v(\pi_{\leq X}) - v(\pi_{<X})\right] \tag{7}$$

where π is a uniformly random permutation of variables $\mathbf{X}$, $\Pi_{\mathbf{X}}$ denotes the set of all permutations of $\mathbf{X}$, $\pi_{\leq X}$ denotes variables up to and including X in permutation π, and $\pi_{<X} = \pi_{\leq X} \setminus \{X\}$ is the set of variables preceding X. We write $\phi_X^{L_3}(w)$ when $\mathcal{M}$ is clear from context. □

As an expectation of counterfactual differences over $(\mathbf{u}, \mathbf{u}')$, $\phi_X^{L_3}(\mathcal{M}, w)$ is a functional of the L_3 distribution of $\mathcal{M}$: two SCMs that induce the same L_3 distribution assign the same counterfactual Shapley value.

L_3 SVs satisfy the four Shapley axioms (efficiency, symmetry, null player, additivity) for value function $v(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{e})$ (Corollary A.18), ensuring variables with zero marginal contribution to every coalition receive zero attribution.

Theorem 3.3 (L_3 SVs satisfy explanatory desiderata)**.** *L_3 SVs satisfy $D_{1,3,4}$ for all SCMs, and D_2 for all but a measure-zero set of structural parameters (Sec. A.9).* □

Proof sketch. By the equivalence between the NTE and functional dependence (Sec. A.7), each Shapley marginal $v(\pi_{\leq X}) - v(\pi_{<X})$ is an average of the unit-level functional dependence c at baseline $\mathbf{Z} = \pi_{<X} \subseteq \mathbf{X} \setminus \{X\}$ (matching Def. 2.3). D_1 follows because $c \equiv 0$ forces every marginal to zero. For D_2, $\phi_X^{L_3} = 0$ requires exact cancellation of NTE marginals; under the premise of D_2, $\phi_X^{L_3}$ is a linear function of the values of the structural equation f_Y with a positive coefficient on the observed outcome, so this equation fails on all but a measure-zero subset of structural parameters (Sec. A.9 and Lem. D.2). D_3 and D_4 follow because pointwise effect ordering and signed baseline dominance, respectively, transfer through the positive-weighted Shapley sum. Full proofs in Sec. D.5. □

This establishes L_3 SVs as the first method satisfying all explanatory desiderata $D_{1:4}$. The exception for D_2 arises when unit-level NTEs cancel exactly in the weighted Shapley sum, analogous to a random variable with nonzero variance having mean exactly zero (Sec. A.9); we introduce an extension of L_3 SVs that resolves this exception in Sec. A.9.2.

Connection to L_1 and L_2 SVs. L_1 SVs (SHAP, [6]) use value function $v^{L_1}(\mathbf{Z}) = \mathbb{E}[Y \mid \mathbf{Z}]$; L_2 SVs (do-SHAP, [22, 21]) use $v^{L_2}(\mathbf{Z}) = \mathbb{E}[Y_{\mathbf{Z}}]$. When the structural equation f_Y is deterministic in $\mathbf{X}$ and there are no causal relationships among $\mathbf{X}$, L_2 and L_3 SVs coincide: counterfactual queries reduce to interventional ones. If there is additionally no confounding within $\mathbf{X}$, $L_{1:3}$ SVs all coincide: observing $\mathbf{Z}$ has the same effect as intervening on $\mathbf{Z}$. We formalize this connection in Sec. A.12.

4 Bounding Algorithm

Given that we have shown that any EVA satisfying $D_{1:4}$ requires L_3 information (Thm. 2.5), and L_3 information is not in general identifiable from $L_{1:2}$ data (Causal Hierarchy Theorem, 10, Thm. 1), counterfactual Shapley values generally cannot be point-identified from $L_{1:2}$ data. We therefore

Table 2: L_3 SV oracle values and identification bounds on Ex. 2.1; non-causes (X_2, X_3) have $\phi = 0$ and bounds containing zero, while causes have $\phi \neq 0$ and the expected magnitude orderings ($\phi_4 > \phi_5$, $\phi_6 > \phi_7$).

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
Oracle L_3 SV	4.28	0.00	0.00	4.15	0.52	0.93	0.03
Bounds	[2.92, 4.28]	[0, 2.92]	[0, 0]	[2.75, 4.15]	[0.34, 0.52]	[0.93, 0.93]	[0.03, 0.03]

follow the approach of [33–35] to bound $\phi_X^{L_3}(\mathcal{M}, w)$ over the set of SCMs consistent with the causal diagram and the data.

We assume access to a causal diagram $\mathcal{G}$ and a dataset $\mathcal{T}$ containing observational (L_1) and interventional (L_2) samples: $\mathcal{T} = \{\mathbf{v}_{\mathbf{s}}^{1:n_{\mathbf{s}}} : \mathbf{S} \subseteq \mathbf{V},\ \mathbf{s} \in \text{Dom}(\mathbf{S})\}$ with $\mathbf{v}_{\mathbf{s}}^{1:n_{\mathbf{s}}} \sim P(\mathbf{V}_{\mathbf{s}})$ and $\mathbf{S} = \varnothing$ for L_1 samples. Let $\mathcal{S}(\mathcal{T}, \mathcal{G})$ denote the SCMs whose associated causal diagram is $\mathcal{G}$ and that reproduce $P(\mathbf{V}_{\mathbf{s}})$ for every $(\mathbf{S}, \mathbf{s})$ supported in $\mathcal{T}$, and let $\hat{\Omega}(\mathcal{T}, \mathcal{G}) \subseteq \mathcal{S}(\mathcal{T}, \mathcal{G})$ be its neural causal model subfamily (NCMs, 36), where each $\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T}, \mathcal{G})$ has structural equations parameterized as neural networks indexed by $\theta \in \Theta$.

Assumption 4.1 (NCM expressivity and regularity)**.** Θ is compact; the map $\theta \mapsto \phi_X^{L_3}(\mathcal{M}_\theta, w)$ is continuous; and for every $\mathcal{M}^* \in \mathcal{S}(\mathcal{T}, \mathcal{G})$, there exists $\theta \in \Theta$ such that $\mathcal{M}_\theta$ induces the same L_3 distribution as $\mathcal{M}^*$. □

By Assump. 4.1, $\hat{\Omega}(\mathcal{T}, \mathcal{G})$ covers every L_3 distribution in $\mathcal{S}(\mathcal{T}, \mathcal{G})$; optimizing $\phi_X^{L_3}(\mathcal{M}_\theta, w)$ over $\theta \in \Theta$ therefore recovers the supremum and infimum of $\phi_X^{L_3}(\mathcal{M}, w)$ over $\mathcal{S}(\mathcal{T}, \mathcal{G})$. Alg. 1 formalizes this optimization (Thm. 4.2).

Theorem 4.2 (Valid and Tight Bounds)**.** *Under Assump. 4.1, Alg. 1 returns* $[\phi^-, \phi^+]$ *such that:* (validity) $\phi_X^{L_3}(\mathcal{M}, w) \in [\phi^-, \phi^+]$ *for every* $\mathcal{M} \in \mathcal{S}(\mathcal{T}, \mathcal{G})$*;* (tightness) *each endpoint is attained at some* $\hat{\mathcal{M}}^\pm \in \hat{\Omega}(\mathcal{T}, \mathcal{G}) \subseteq \mathcal{S}(\mathcal{T}, \mathcal{G})$*, where* $\hat{\mathcal{M}}^-$ *and* $\hat{\mathcal{M}}^+$ *may differ.* □

Proof sketch. Since Θ is compact and $\phi_X^{L_3}(\mathcal{M}_\theta, w)$ is continuous in θ, the argmin/argmax over $\hat{\Omega}(\mathcal{T}, \mathcal{G})$ is attained; let $\hat{\mathcal{M}}^\pm$ be a minimizer/maximizer, so $\phi^\pm = \phi_X^{L_3}(\hat{\mathcal{M}}^\pm, w)$. By Assump. 4.1, every $\mathcal{M} \in \mathcal{S}(\mathcal{T}, \mathcal{G})$ shares its L_3 distribution with some $\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T}, \mathcal{G})$, so $\phi^- \leq \phi_X^{L_3}(\mathcal{M}, w) \leq \phi^+$ (validity). Each endpoint is attained at the witnessing $\hat{\mathcal{M}}^\pm \in \hat{\Omega}(\mathcal{T}, \mathcal{G}) \subseteq \mathcal{S}(\mathcal{T}, \mathcal{G})$ (tightness). See Sec. D.4. □

To allow for gradient optimization of surrogate models, following [36], we relax the optimization constraints of Alg. 1 into Lagrangian penalties.

$$\mathcal{L}^\pm(\theta; \mathcal{T}) = -\sum_{(\mathbf{s},\mathbf{v}) \in \mathcal{T}} \log P^{\mathcal{M}_\theta}(\mathbf{v}_{\mathbf{s}}) \mp \lambda\, \hat{\phi}_X^{L_3}(\mathcal{M}_\theta, w), \tag{8}$$

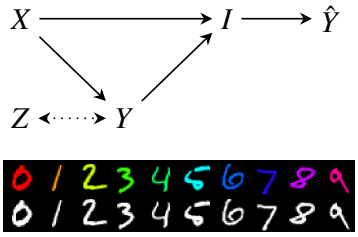
where $\hat{\phi}_X^{L_3}(\mathcal{M}_\theta, w)$ is the Monte Carlo estimator of Def. 3.2 and coefficient $\lambda > 0$ determines the strength of the penalty. See Sec. C.1 for implementation details.

Startup bounds. Table 2 applies Alg. 1 to Ex. 2.1 and reports the oracle counterfactual Shapley values alongside the computed bounds. The bounds are consistent with $D_{1:4}$: they contain zero for non-causes X_2, X_3; they exclude zero for but-for causes X_1, X_4, X_5, X_6, X_7; and they preserve $\phi_4 > \phi_5$ and $\phi_6 > \phi_7$, as required by D_3 and D_4.

5 Experiments

We empirically corroborate L_3 SVs' satisfaction of admissibility (D_1) and power (D_2) in a simple shortcut learning setting using an adaptation of the Color MNIST dataset [37], where $L_{1:2}$ SVs violate D_1.

Setup. As shown in Fig. 4, top left, images I are colored with hue X and contain digit Y. Data source Z indicates whether the hue affects the digit: with 90% chance, $Z = 1$ and the hue determines



Method	Std	Rob	D_1	D_2
L_1 SVs	**1.000***	1.000*	✗	✓
L_2 SVs	**1.000***	1.000*	✗	✓
L_3 SVs	**1.000***	**0.000**	✓	✓
L_3 bounds	**1.000***	**0.000**	✓	✓

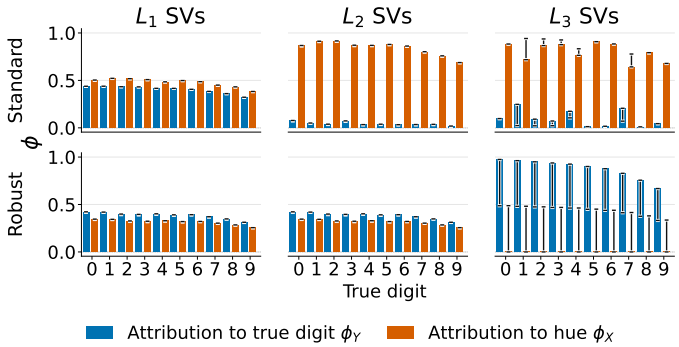


Figure 4: **Causal explanations on Color MNIST.** *Top left:* causal graph and sample digits. *Bottom left:* table showing the fraction of 1,000 queries in which each method attributes the prediction to hue under the standard (Std) and robust (Rob) classifiers, and whether the method satisfies D_1 and D_2. *Right:* hue (orange) and digit (blue) attributions for one query per digit; whiskers show bootstrap SEM for $L_{1:2}$ SVs, and identification bounds for L_3 SVs.

the digit, inducing a substantial correlation between hue and digit. Two LeNet classifiers [38] are trained on 60,000 sampled images to predict the digit of the image, producing prediction $\hat{Y}$: a standard classifier, which learns to predict the digit based on the image's hue, and a robust classifier, which receives only a grayscale version of the image and therefore cannot use the hue to make predictions. Because the handwritten digit is noisy and differs from Y with probability 0.17, the standard classifier learns to rely on hue as a shortcut to learn the digit. The robust classifier does not have access to the hue and therefore does not learn this shortcut. For each digit $d \in \{0, \ldots, 9\}$ we take 100 correctly classified images from source $Z = 0$ with $X = Y = d$ and pose the why-query $\text{Why}(\hat{Y}{=}d \mid Z{=}0, X{=}d, Y{=}d, I{=}i, \hat{Y}{=}d;\ \{X, Y\})$. Exact $L_{1:3}$ SVs are computed from the SCM, and bounds are computed from observational data and the graph, and the SCM is shown in Sec. C.3.1.

Hypotheses. Hue affects the prediction through two paths: directly, $X \to I \to \hat{Y}$, and through the digit, $X \to Y \to I \to \hat{Y}$. Following the query, every queried image comes from source $Z = 0$, where hue does not affect the digit, so the path through the digit is inactive; the robust classifier has no access to the hue, so the direct path is inactive as well, and D_1 requires $\phi_X = 0$ for the robust classifier. Under the standard classifier, the direct path is active and hue is a but-for cause of the prediction, so D_2 requires $\phi_X \neq 0$ and bounds that exclude zero. We expect $L_{1:2}$ SVs to violate D_1, because they marginalize over Z, and under $Z = 1$, hue does cause the digit, which activates the path through the digit for both classifiers. We expect L_3 SVs and their bounds to satisfy D_1 and D_2, because they condition on $Z = 0$.

Results. The table in Fig. 4 reports the fraction of queries in which the prediction $\hat{Y} = d$ is attributed to hue ($\phi_X \neq 0$, or $0 \notin [\phi_X^-, \phi_X^+]$), for each method and classifier. $L_{1:2}$ SVs violate D_1 by attributing the robust classifier's prediction to hue in every query. L_3 SVs satisfy D_1 and D_2, as they attribute the standard classifier's prediction to hue in every query and the robust classifier's in none. Consistent with the table, the attribution panels show that $L_{1:2}$ SVs assign positive attribution to hue for the robust classifier (second row, in orange), in violation of D_1, while L_3 SVs assign zero. The bounds for L_3 SVs include zero for hue under the robust classifier, and exclude zero for hue under the standard classifier and for the digit (in blue) under both. The bounds contain the exact L_3 SVs in every shown query, which validates their soundness, and they exclude zero for the digit under the robust classifier, which supports their informativeness: the digit's nonzero effect is inferred from observational data and the graph alone. Under the robust classifier, the upper bound for hue and the lower bound for the digit meet, because efficiency fixes $\phi_X + \phi_Y$ and the digit's response to other hues in source 0 is not identified from $P(Z, X, Y)$; the bounds therefore cannot order the two attributions, and separating them requires assumptions beyond the graph.

In summary, exact L_3 SVs and their bounds satisfy D_1 and D_2 in a shortcut learning setting where $L_{1:2}$ SVs violate D_1.

6 Conclusions

We developed a normative framework for evaluating explanatory variable attributions through four desiderata grounded in philosophy, law, and cognitive psychology: admissibility, power, effect scaling, and normality. We traced the disagreement problem [8, 9] to existing methods' failure to satisfy the desiderata, and to the fact that, even in principle, methods without counterfactual information cannot satisfy them, as shown by the Explanatory Impossibility Theorem. We introduced counterfactual Shapley values, which satisfy all desiderata, and developed an algorithm to produce tight bounds on counterfactual Shapley values. Experiments corroborate our theory.

Limitations. Our desiderata assume that functional dependence is a necessary condition for causation and that the but-for test is sufficient; if these assumptions fail, non-causes may receive nonzero attribution, or causes zero. Our method requires a known causal diagram. Computing L_3 SVs exactly requires exponential time in the number of variables, so a tractable method must approximate them. Bound tightness depends on convergence of the underlying optimization. Our experiments use synthetic and semi-synthetic data, and scaling to settings with more variables remains a direction for future research.

Acknowledgments and Disclosure of Funding

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

[1] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2000.

[2] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion*, 106(C), June 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2024.102301. URL https://doi.org/10.1016/j.inffus.2024.102301.

[3] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[4] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[6] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[8] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*, 2023.

[9] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[10] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition, 2022.

[11] David Lewis. Causal explanation. In David Lewis, editor, *Philosophical Papers, Volume II*, pages 214–240. Oxford University Press, 1986.

[12] Herbert Lionel Adolphus Hart and Tony Honoré. *Causation in the Law*. OUP Oxford, 1985.

[13] Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1-2):93–149, 1999. doi: 10.1023/a:1005233831499.

[14] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York, 2003.

[15] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2005.

[16] Sander Beckers. Causal explanations and xai. In *Conference on causal learning and reasoning*, pages 90–109. PMLR, 2022.

[17] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[18] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[19] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

[20] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

[21] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Neural Information Processing Systems*, 2020.

[22] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[23] Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing, P. Blobaum, and E. Bareinboim. On measuring causal contributions via do-interventions. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 10476–10501. PMLR, Jul 2022.

[24] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

[25] Amit Sharma, Hua Li, and Jian Jiao. The counterfactual-shapley value: Attributing change in system metrics. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022.

[26] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 2005.

[27] David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973. ISSN 0022362X. URL http://www.jstor.org/stable/2025310.

[28] Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

[29] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[30] Daniel Kahneman and Dale T Miller. Norm theory: Comparing reality to its alternatives. *Psychological review*, 93(2):136, 1986.

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

[32] Lloyd S Shapley et al. A value for n-person games. *Annals of Mathematics Studies*, 28:307–318, 1953.

[33] Charles F Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80 (2):319–323, 1990.

[34] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pages 46–54. Elsevier, 1994.

[35] Junzhe Zhang, Jin Tian, and Elias Bareinboim. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pages 26548–26558. PMLR, 2022.

[36] Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2022.

[37] Yushu Pan and Elias Bareinboim. Counterfactual image editing. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39087–39101. PMLR, 21–27 Jul 2024.

[38] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[39] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach: Part 1: Causes. In Jack S. Breese and Daphne Koller, editors, *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 194–202. Morgan Kaufmann, 2001. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=100&proceeding_id=17.

[40] Joseph Y Halpern. A modification of the halpern-pearl definition of causality. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3022–3033, 2015.

[41] Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, 195(2):835–862, 2018. doi: 10.1007/s11229-016-1247-1.

[42] Sander Beckers. The counterfactual ness definition of causation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6210–6217, May 2021. doi: 10.1609/aaai.v35i7.16772. URL https://ojs.aaai.org/index.php/AAAI/article/view/16772.

[43] Tomasz Wysocki. Conjoined cases. *Synthese*, 201(6):1–19, 2023. doi: 10.1007/s11229-023-04101-w.

[44] Marc Denecker, Bart Bogaerts, and Joost Vennekens. Explaining actual causation in terms of possible causal processes. In *Logics in Artificial Intelligence: 16th European Conference, JELIA 2019, Rende, Italy, May 7–11, 2019, Proceedings 16*, pages 214–230. Springer, 2019.

[45] Joseph Y Halpern and Christopher Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 2015.

[46] Mark Hopkins and Judea Pearl. Clarifying the usage of structural models for commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, pages 83–89. AAAI Press Menlo Park, CA, 2003.

[47] Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015.

[48] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

[49] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. ISBN 9780691120348. URL http://www.jstor.org/stable/j.ctvcm4j72.

[50] Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

[51] Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

[52] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[54] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[55] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[56] Dominik Janzing, Patrick Blöbaum, Atalanti A Mastakouri, Philipp M Faller, Lenon Minorics, and Kailash Budhathoki. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2024.

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

[58] 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*, pages 607–617, 2020.

[59] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), dec 2022. ISSN 0360-0300. doi: 10.1145/3527848. URL https://doi.org/10.1145/3527848.

[60] Sam Baron. Explainable ai and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33(2):347–377, 2023.

[61] Margot E Kaminski. The right to explanation, explained. In *Research handbook on information law and governance*, pages 278–299. Edward Elgar Publishing, 2021.

[62] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, page 3730–3738. IEEE, December 2015. doi: 10.1109/iccv.2015.425. URL http://dx.doi.org/10.1109/iccv.2015.425.

[63] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Appendix

Contents

A. Theoretical Foundations 14
- A.1 Functional Dependence 14
- A.2 Oracle-Proxy Framework 16
- A.3 Gricean Grounding of Desiderata 18
- A.4 Worked Examples of the Desiderata 19
- A.5 Extended Why Queries 20
- A.6 From Probability of Necessity to Natural Total Effects 20
- A.7 NTE-Functional Dependence Equivalence 21
- A.8 Normality Theory 21
- A.9 Measure on SCMs 23
- A.10 Startup Pitch SCM 25
- A.11 Shapley Axioms 25
- A.12 When L_1, L_2, and L_3 SVs Coincide 26
- A.13 Causal Explained Variance 26

B. Method Violations and Counterexamples 28
- B.1 Simple Causal Structures for D2 Power 29
- B.2 Extended Results Table and Discussion of Methods 30
- B.3 Computation of L_3 Shapley Values 32
- B.4 Examples 35

C. Computational Methods and Extended Experiments 38
- C.1 Computational Methodology 38
- C.2 Startup Experiment 39
- C.3 Color MNIST Experiment 39
- C.4 CelebA-HQ Experiment 43

D. Proofs 46
- D.1 Proof of Desiderata Lifting Theorem 47
- D.2 Proof of NTE-Functional Dependence Equivalence 47
- D.3 Proof of Explanatory Impossibility Theorem 48
- D.4 Proof of Validity and Tightness 49
- D.5 Proof that L_3 SVs Satisfy All Desiderata 50
- D.6 Proof of Shapley Axioms 53

E. Extended Related Work 53

F. Frequently Asked Questions 54

G. Broader Impacts 56

NeurIPS Paper Checklist 57

A Theoretical Foundations

A.1 Functional Dependence

Functional dependence (Def. 2.3) measures counterfactual differences in outcomes when changing a variable. We now establish its relationship to actual causation theories.

A.1.1 Necessity for Causation

Functional dependence appears as a necessary condition across all major definitions of actual causation. All causation theories using SCMs require that changing X can possibly affect Y under some contingency. The HP definition [39, 15, 40] requires AC2(a) or AC2(a^m), which asserts existence of a contingency where changing X changes Y (we focus on singleton candidate causes; the same necessity argument extends to set-valued candidates by applying AC2 to the set as a whole). Beckers' definition [41, 42] includes an analogous structural dependence condition AC2(a^c), which

requires that X makes a difference to Y under some permissible contingency, directly implying functional dependence. Lewis' *counterfactual dependence* [27] directly requires that changing the cause changes the outcome, which is functional dependence without contingency (i.e., with $\mathbf{Z} = \emptyset$). The convergence of major actual causation definitions [40, 42, 27] on the requirement that changing X can change Y provides strong evidence that functional dependence captures a necessary property of causation, which we adopt as our proxy.

Wysocki [43] argues that no current normality-augmented counterfactual theory of causation handles all conjoined cases; we sidestep this debate by remaining agnostic to the specific actual-causation definition (Sec. A.1.2).

Formally, for HP actual causation: if $X = x$ is an actual cause of $Y = y$ in $(\mathcal{M}, \mathbf{u})$, then there exists a witness $\mathbf{W}$ and setting $\mathbf{w}$ such that changing X from x to some x' under contingency $\mathbf{W} = \mathbf{w}$ changes Y from y to some $y' \neq y$. This directly implies functional dependence: $\exists x', \mathbf{z}' : Y_{\mathbf{z}'}(\mathbf{u}) \neq Y_{\mathbf{z}',x'}(\mathbf{u})$, so $c(\mathcal{M}, \mathbf{u}, X, Y, (x', \mathbf{z}')) \neq 0$. Therefore HP-causation implies functional dependence. Taking the contrapositive: if functional dependence is zero for all witnesses ($c(\mathcal{M}, \mathbf{u}, X, Y, z) = 0$ for all z), then HP-causation is zero. This establishes that functional dependence is necessary for HP actual causation.

Proposition A.1 (HP-AC implies FD)**.** *Let $\mathbf{X} = \mathbf{V}$ (explanatory variables equal all endogenous variables). If $X = x$ is an HP-actual cause of $Y = y$ in $(\mathcal{M}, \mathbf{u})$, then $c(\mathcal{M}, \mathbf{u}, X, Y, z) \neq 0$ for some baseline $z \in \mathcal{Z}$.* □

Proof. HP's AC2 requires a witness $\mathbf{W} \subseteq \mathbf{V} \setminus \{X\}$ where changing X changes Y. When $\mathbf{X} = \mathbf{V}$, we have $\mathbf{W} \subseteq \mathbf{X} \setminus \{X\}$, so the HP witness is a valid FD baseline in $\mathcal{Z}$. □

Note that functional dependence restricts contingencies to $\mathbf{X} \setminus \{X\}$, matching EVA's scope (Def. 2.2). When $\mathbf{X} \subsetneq \mathbf{V}$, a variable may be an HP-actual cause via a witness outside $\mathbf{X}$, but such pathways lie outside the why-query's scope: the EVA framework attributes only to user-specified explanatory variables. This scope restriction is intentional: it answers "is X_i a cause of Y among explanatory variables $\mathbf{X}$?" rather than "is X_i a cause of Y in the full model $\mathbf{V}$?"

The desiderata D_2, D_3 are stated using the but-for baseline $\mathbf{z}'_\emptyset$, but each generalizes to any HP witness set $\mathbf{W} \subseteq \mathbf{X} \setminus \{X\}$ [28]: the antecedent "$c(\mathcal{M}, \mathbf{u}, X, Y, (x', \mathbf{z}'_\emptyset)) \neq 0$" may be replaced by "there exist $\mathbf{w}, x'$ with $c(\mathcal{M}, \mathbf{u}, X, Y, (x', \mathbf{w})) \neq 0$", and analogously for D_3 on conditional effect distributions. For L_3 Shapley attributions and all but a measure-zero set of structural parameters (Sec. A.9), the two forms are equivalent at baselines of positive probability: ϕ satisfies both. We use the but-for form in the main text for readability.

A.1.2 Why we don't define actual causation

The actual causation literature [39, 15, 40–42] provides formal definitions of causation, but no definition has been proven correct [44], and counterexamples exist for all current definitions [43, 45]. Following Lewis [11]'s separation of explanation from causation, we develop a framework agnostic to the precise definition of causation.

Rather than resolve this debate, we use the *functional dependence* criterion, a necessary condition for major definitions of actual causation: under some intervention on the remaining variables, changing a cause must change the outcome. This generalizes the but-for test ("but for X, would Y have occurred?") central to jurisprudence, philosophy, and causal inference [12, 14, 15, 13].

We use functional dependence as the necessary condition (every major SCM-based AC definition requires it; Prop. A.1) and the but-for test as the sufficient condition [12]. This sandwiches the chosen AC definition, regardless of which definition: any EVA satisfying D_2 or violating D_1 under functional dependence c thereby does so under any AC definition that respects both halves (Thm. A.7). Functional dependence is also computable from an SCM without resolving ongoing debates about the correct definition of actual causation.

A.1.3 Non-Sufficiency of Functional Dependence

The converse fails: functional dependence does not imply causation. We illustrate with a classic example from Hopkins and Pearl [46]:

Example A.2 (Prisoner)**.** Suppose that a prisoner dies either if A loads B's gun and B shoots, or if C loads and shoots his own gun. The structural equation is $D = (A \wedge B) \vee C$, where D represents the prisoner's death.

In the actual world $\mathbf{u}$, we have $A = 1$ (A loads B's gun), $B = 0$ (B does not shoot), and $C = 1$ (C loads and shoots his own gun), so $D = 1$ (prisoner dies). Here $C = 1$ is a cause of $D = 1$, since the prisoner died due to C shooting. According to standard intuitions, the loader of the unshot gun $A = 1$ is not a cause of $D = 1$.

Yet D exhibits functional dependence on A in the contingent world where $B = 1$ and $C = 0$:

$$D_{B=1,C=0,A=0}(\mathbf{u}) = (0 \wedge 1) \vee 0 = 0 \tag{9}$$

$$D_{B=1,C=0,A=1}(\mathbf{u}) = (1 \wedge 1) \vee 0 = 1 \tag{10}$$

As we can see, there exists a contingency where changing A changes D. Thus $A = 1$ exhibits functional dependence despite not being an actual cause of $D = 1$. □

This example illustrates that functional dependence is a conservative proxy: it captures all actual causes but may include some non-causes (preemption, overdetermination cases). Our framework accommodates this gap through the conservative proxy assumption (Sec. A.2), which requires only that $c' \neq 0 \Rightarrow c \neq 0$ (conservativity), not the converse. This asymmetry is intentional: false positives from functional dependence are acceptable when the goal is to evaluate explanation methods rather than identify actual causes directly.

Together, conservativity (functional dependence is necessary for causation) and effect preservation (when causation holds, effect sizes match) provide a principled foundation for method evaluation. Functional dependence is computable from an SCM without requiring resolution of ongoing debates about the correct definition of actual causation.

A.1.4 Relationship to Shapley Axioms

L_3 SVs satisfy the four Shapley axioms (efficiency, symmetry, null player, marginality) for value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{e})$ (Corollary A.18). The null player axiom requires $\phi_X = 0$ when X has zero marginal contribution to every coalition. This is weaker than our admissibility desideratum (D_1): admissibility implies null player (non-causes have zero NTE contribution by definition), but null player does not imply admissibility (a variable could have zero attribution due to cancellation across coalitions while still being a cause). Our desiderata thus provide stronger guarantees than the Shapley axioms alone.

A.2 Oracle-Proxy Framework

We now present the full framework for evaluating explanatory methods under epistemic uncertainty about the true causation definition.

We introduce the *causal measure* as a generic interface to any causation definition.

Definition A.3 (Causal Measure)**.** A causal measure c is a mapping $c : \Omega \times \text{Dom}(\mathbf{U}) \times \mathbf{V} \times \mathbf{V} \times \mathcal{Z} \to \mathbb{R}$ that takes an (X, Y) pair, where Ω is the space of SCMs, $\text{Dom}(\mathbf{U})$ is the domain of the exogenous variables in $\mathcal{M}$, $(X, Y) \in \mathbf{V} \times \mathbf{V}$ are the cause-effect variables in $\mathcal{M}$, and $\mathcal{Z}$ is a baseline space equipped with probability measure $P^{\mathcal{M}}$, containing potential baselines for counterfactual comparison. □

This abstraction enables systematic evaluation without committing to any causation definition, generalizing beyond binary judgments to capture effect magnitudes. We assume objective causation exists and make our framework agnostic to its precise definition.

Assumption A.4 (Causation Oracle)**.** There exists an unknown causation oracle $c' \in \mathbb{C}$. □

This assumption allows us to: (1) formalize desiderata that must hold for good explanations given a causal measure we take to be the oracle c', and (2) infer violations (and in some cases, satisfactions) of oracle desiderata from functional dependence c under reasonable assumptions. Our approach will remain relevant as long as establishing correctness for causation definitions remains an open problem.

Since we lack c', we cannot directly evaluate oracle desiderata. For instance, we cannot determine whether Alice's live demo caused her funding, which if not would require methods to yield $\phi_2 = 0$

to satisfy D_1. We therefore use functional dependence c as a conservative measure that captures necessary conditions for causation.

Assumption A.5 (Conservative Proxy with Effect Preservation)**.** Under Assump. A.4, there exists a mapping $f : \mathcal{Z}' \to \mathcal{Z}$ (where $\mathcal{Z}'$ is the oracle's baseline space) such that for all $z' \in \mathcal{Z}'$: $c'(z') \neq 0 \implies c'(z') = c(f(z'))$. The mapping f maps but-for baselines to but-for baselines, and maps each set of oracle baselines with $c' \neq 0$ and positive probability to a set of baselines with positive probability. This ensures $c' \neq 0 \Rightarrow c \neq 0$. □

The assumption states that causation implies functional dependence (conservativity: $c' \neq 0 \Rightarrow c \neq 0$), and that effect sizes are correctly captured when causation holds (effect preservation: $c'(z') = c(f(z'))$). The converse need not hold, and is in fact known to fail for some examples (Sec. A.1.3). Conservativity is justified by Sec. A.1.1: functional dependence is necessary for causation across HP, Beckers, and Lewis. Effect preservation is reasonable because quantitative causal effects are defined as counterfactual differences across disciplines [1, 47–49]; functional dependence (c) directly captures these differences when there is causation ($c' \neq 0$). The conservative proxy assumption enables us to prove oracle desiderata violations on existing EVAs on the basis of functional dependence violations.

We now state the general form of explanatory desiderata, parameterized by an arbitrary causal measure.

Definition A.6 (General Explanatory Desiderata)**.** We define four desiderata as mappings $D_{1:4} : \Omega \times \mathbb{C} \times \Phi \to \{0, 1\}$, where $\mathbb{C}$ is the space of causal measures and Φ is the space of EVAs. Given SCM $\mathcal{M} \in \Omega$, causal measure $c \in \mathbb{C}$, and EVA $\phi \in \Phi$, we say $D_i(\mathcal{M}, c, \phi) = 1$ when the conditions in Def. 2.4 hold. This generalizes the main text definition, which fixes c to functional dependence. □

Theorem A.7 (Desiderata Lifting)**.** *Under Assump. A.5, violations of D_1, D_3, D_4 and satisfaction of D_2 lift from functional dependence c to causation c'. For D_1, violations lift unconditionally by conservativity. For D_2, satisfaction lifts unconditionally; violations additionally lift when there exists a set of unit-baseline pairs $\{(\mathbf{u}, z)\}$ with positive measure (where $\mathbf{u}$ induces observed evidence $\mathbf{e}$) satisfying $c = c' \neq 0$ and $\phi = 0$. For D_3, D_4, violations lift when $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$ for the SCMs in the comparison. Property P lifts from c to c' if $P(\mathcal{M}, c, \phi) \Rightarrow P(\mathcal{M}, c', \phi)$ for all SCMs $\mathcal{M}$.* □

A.2.1 When Lifting Works

The lifting theorem establishes when functional dependence violations imply causation violations. We clarify the conditions for each desideratum.

D1 (Admissibility) violations lift unconditionally. If a method assigns non-zero attribution when $c = 0$, this violates $D_1(c)$. By conservativity ($c = 0 \Rightarrow c' = 0$), we have $c' = 0$, so the method also violates $D_1(c')$. No additional conditions are needed.

D2 (Power) satisfaction lifts unconditionally; violations lift conditionally. If a method satisfies $D_2(c)$ (assigns non-zero attribution whenever $c \neq 0$), it also satisfies $D_2(c')$. This is because $c' \neq 0 \Rightarrow c \neq 0$ (by conservative proxy), so whenever $c' \neq 0$, we have $c \neq 0$, which by $D_2(c)$ means attribution is non-zero.

For D2 violations, showing a method violates $D_2(c)$ (assigns zero when $c \neq 0$) does not automatically imply violation of $D_2(c')$, because the case where $c \neq 0$ might have $c' = 0$ (false positive, as in the prisoner example). To establish $D_2(c')$ violation, we must show $c' = c$ for the specific instance. In Rem. B.1, we verify this for Alice's startup by exhibiting, for each variable, a unit with positive probability where the variable is a but-for cause.

D3, D4 (Effect Scaling, Normality) violations lift when $c = c'$ in the SCM comparison. These desiderata compare two SCMs $\mathcal{M}$ and $\mathcal{M}'$. If $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$ (functional dependence exactly matches causation in both SCMs), then violations under c imply violations under c'. For Alice's startup, we verify $c = c'$ for each relevant variable by the but-for test (Rem. B.1).

Asymmetry explanation. The asymmetry arises from conservativity: c is over-inclusive (captures all true causes plus some non-causes). Therefore: - Soundness desiderata (D1): Violations under over-inclusive c are violations under any c' - Completeness desiderata (D2): Satisfaction under c guarantees satisfaction under c', but violations require case-specific argument - Comparison desiderata (D3, D4): Violations under c are violations under c' when $c = c'$ in the comparison

This framework allows us to systematically evaluate methods under computable functional dependence while providing guarantees about performance under the unknown true causation measure.

A.3 Gricean Grounding of Desiderata

The desiderata in Def. 2.4 are grounded in Grice [29]'s conversational maxims. This grounding derives from communication theory rather than causation metaphysics, so it remains valid regardless of which causation definition is correct.

Grice's cooperative principle states: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." He decomposes this into four maxims, each with submaxims:

Quality. "Try to make your contribution one that is true."

(1) "Do not say what you believe to be false."
(2) "Do not say that for which you lack adequate evidence."

This grounds causal admissibility (D_1): attributing to non-causes would be saying something false about the causal history. The second submaxim also motivates providing bounds when exact ϕ_i values cannot be computed: bounds $[\phi_L, \phi_U]$ with documented evidence operationalize "adequate evidence."

Quantity.

(1) "Make your contribution as informative as is required (for the current purposes of the exchange)."
(2) "Do not make your contribution more informative than is required."

This grounds causal power (D_2): omitting actual causes would fail to be "as informative as is required." The second submaxim motivates scope limitation to explanatory variables $\mathbf{X}$ rather than all variables in the causal history.

Relation. "Be relevant." Grice explicitly acknowledged this as underspecified: "I find the treatment of such questions [about what different kinds and focuses of relevance there may be] exceedingly difficult." Causal effect scaling (D_3) and causal normality (D_4) specify two concrete "kinds and focuses of relevance" for causal explanation: causes with larger effects are more relevant (effect scaling), and causes with more probable baselines are more relevant (normality).

Manner. "Be perspicuous," with submaxims:

(1) "Avoid obscurity of expression."
(2) "Avoid ambiguity."
(3) "Be brief (avoid unnecessary prolixity)."
(4) "Be orderly."

This motivates the EVA framework itself (Def. 2.2): presenting causal contributions as a real-valued attribution vector $\phi \in \mathbb{R}^n$ provides a perspicuous format with unambiguous numerical values, brief $O(|\mathbf{X}|)$ output, and orderly ranking by magnitude.

Sufficiency, not minimality. We do not claim these four desiderata are minimal or complete. Additional desiderata mapping to relevance or other maxims may exist. However, the four desiderata are sufficient for our purpose: they provide systematic coverage of all four Gricean maxims and eliminate all existing methods (Table 1).

Relationship to prior axioms. Prior axiomatic frameworks for attribution methods characterize properties of value functions or distributions. The Shapley axioms (null player, efficiency, symmetry, marginality) constrain how a value function is decomposed into per-player contributions [32]. Integrated Gradients satisfies sensitivity and implementation invariance, which constrain gradient-based computations [18]. Interventional Shapley values (L_2 SVs) replace the observational value function with an interventional one [22, 21, 23]. Our desiderata $D_{1:4}$ are, to our knowledge, the first to be formalized using counterfactual quantities in the language of structural causal models. A method can satisfy all Shapley axioms while violating D_1, because the axioms constrain how marginal

contributions are combined, not whether the underlying value function responds to causal or spurious associations. For instance, SHAP satisfies all four Shapley axioms yet violates causal admissibility (D_1) because its L_1 value function captures spurious correlations. Similarly, L_2 SVs satisfy Shapley axioms over an interventional value function yet violate D_1 because interventional effects average over individuals rather than conditioning on the specific unit being explained. Finally, D_3 (effect scaling) and D_4 (normality) have no analogues in prior axiom sets: they concern how attribution *ranks* causes relative to each other based on effect magnitude and baseline probability, which no prior framework addresses.

A.4 Worked Examples of the Desiderata

Ex. 1.1 (Betting game). Alice bets heads ($X = 1$) on Bob's fair coin flip and wins her bet ($Y = 1$).

This example corresponds to the following SCM, which we refer to as $\mathcal{M}_1$:

$$\begin{aligned} U_X, U_Y &\sim \text{Bern}(0.5) \\ X &:= U_X \\ Y &:= \mathbf{1}[X = U_Y] \end{aligned}$$

where $\mathbf{1}[\cdot]$ denotes an indicator variable. The why-query is $\text{Why}(Y = 1 \mid X = 1, Y = 1; \{X\})$, and ϕ_X is computed in five steps:

1. **Abduction.** From $X = 1$ we may infer $U_X = 1$, and from $Y = 1$ together with $X = 1$, $U_Y = 1$. The observed evidence therefore identifies the unit $\mathbf{u} = \{U_X = 1, U_Y = 1\}$.
2. **Baselines.** X is the only explanatory variable, so a baseline is represented with $\mathbf{Z} = \varnothing$, and a value for an alternative bet $x' \in \{0, 1\}$. Each baseline $b = (x', \varnothing)$ has $P(b) = 0.5$.
3. **Functional dependence (Def. 2.3).** At $x' = 0$, $Y_{X=0}(\mathbf{u}) = \mathbf{1}[0 = 1] = 0$, so $c(x' = 0) = 1 - 0 = 1$: had Alice bet tails, she would have lost. Under the baseline $x' = 1$, which is the actual value of X, no change is made to the value of X, so $Y_{X=1}(\mathbf{u}) = Y(\mathbf{u}) = 1$ and $c(x' = 1) = 0$. With $\mathbf{Z} = \varnothing$ this is exactly the but-for test.
4. **Natural total effect (Def. 3.1).** Abduction identified a single unit, so the outer expectation in Eq. (6) collapses and $\text{NTE}(\{X\}, Y \mid e) = \mathbb{E}_{u' \sim P(\mathbf{U})}[Y_{X(\mathbf{u})}(\mathbf{u}) - Y_{X(u')}(\mathbf{u})] = 1 - (0.5 \cdot 0 + 0.5 \cdot 1) = 0.5$.
5. **Attribution (Def. 3.2).** With a single variable, $\phi_X = v(\{X\}) - v(\varnothing) = 0.5 - 0 = 0.5$.

We note that the example above illustrates the power desideratum (D_2). Specifically, we are able to infer that X is a cause of Y in this case, because it passes the but-for test: under the baseline $x' = 0$, we observe a non-zero effect of $c(x' = 0) = 1$ with baseline probability $P(b) = 0.5$. Therefore, the distribution of functional dependence over baselines is not degenerate at zero ($P^{\mathcal{M}}(c_B \mid \mathbf{z}'_\varnothing) \not\equiv \delta_0$), and D_2 thus requires $\phi_X \neq 0$. As we can see, L_3 SVs match this requirement, yielding $\phi_X = 0.5$.

We now introduce three variants of this example, to illustrate the remaining desiderata. Each variant changes either the distribution of U_X or the equation for Y, and is shown in the table below. The table lists the functional dependence c and the signed baseline measure $P_+(b) = \text{sign}(c(b))P(b)$ of Def. 2.4 at each of the two baselines.

Table 3: Functional dependence, signed baseline measure and attribution for $\mathcal{M}_1$ to $\mathcal{M}_4$.

SCM	Modification	$c(x'{=}0)$	$c(x'{=}1)$	$P_+(x'{=}0)$	$P_+(x'{=}1)$	ϕ_X
$\mathcal{M}_1$ (Fig. 1, Ex. 1.1)	$Y := \mathbf{1}[X = U_Y]$	1	0	0.5	0	0.5
$\mathcal{M}_2$ (Fig. 1, Ex. 1.2)	$Y := U_Y$	0	0	0	0	0
$\mathcal{M}_3$	$Y := 2 \cdot \mathbf{1}[X = U_Y] - 1$	2	0	0.5	0	1
$\mathcal{M}_4$	$P(U_X = 1) = 0.99$	1	0	0.01	0	0.01

Now, we discuss how each variant in the Table above exemplifies one desideratum:

1. $\mathcal{M}_2$ is identical to Ex. 1.2, in which Y is Bob's own coin flip rather than the settlement of a bet. Because X does not appear in the equation for Y, the effect is zero at every baseline, so the precondition of admissibility (D_1) is satisfied ($c_B \equiv \delta_0$), meaning that the attribution should yield $\phi_X = 0$. As we can see, L_3 SVs satisfy this requirement.
2. $\mathcal{M}_3$ differs from $\mathcal{M}_1$ in that Alice now loses a dollar if she loses, so her bet now makes twice the difference to the outcome (derived from the X_4 against X_5 comparison in Rem. B.1). Looking

at D_3, we evaluate the three criteria of the precondition: 1) Alice's bet never makes a smaller difference to the outcome ($c_{\mathcal{M}_3} \geq c_{\mathcal{M}_1}$, as $2 \geq 1$ at $x' = 0$ and $0 \geq 0$ at $x' = 1$), 2) the baseline distributions agree ($P^{\mathcal{M}_3}(B) \equiv P^{\mathcal{M}_1}(B)$, since $P(U_X)$ is unchanged), and 3) the effect distributions differ ($P^{\mathcal{M}_3}(c_B \mid \mathbf{z}'_\varnothing) \not\equiv P^{\mathcal{M}_1}(c_B \mid \mathbf{z}'_\varnothing)$). Notably, conditions (1) and (3) work together to imply that the effects that differed increased from $\mathcal{M}_1 \to \mathcal{M}_3$. In this case, and holding the baseline distribution fixed, effect scaling therefore requires $\phi_{\mathcal{M}_3} > \phi_{\mathcal{M}_1}$. As we can see, L_3 SVs satisfy this requirement, yielding $1 > 0.5$.

3. $\mathcal{M}_4$ differs from $\mathcal{M}_1$ in that Alice now almost always bets heads, as the equation for Y is unchanged and only $P(U_X)$ changes (derived from the X_6 against X_7 comparison in Rem. B.1). Looking at D_4, we evaluate the three criteria of the precondition: 1) the effects between the models are identical ($c_{\mathcal{M}_1} = c_{\mathcal{M}_4}$), 2) the signed baseline measures differ ($P_+^{\mathcal{M}_1}(B) \not\equiv P_+^{\mathcal{M}_4}(B)$), and 3) any bets that would have changed the outcome were not less probable under $\mathcal{M}_1$ ($P_+^{\mathcal{M}_1}(b) \geq P_+^{\mathcal{M}_4}(b)$, as $0.5 \geq 0.01$ at $x' = 0$ and $0 \geq 0$ at $x' = 1$). Notably, conditions (2) and (3) work together to imply that her bet was strictly more probable under $\mathcal{M}_1$ at some baseline. Normality (D_4) therefore requires $\phi_{\mathcal{M}_1} > \phi_{\mathcal{M}_4}$. As we can see, L_3 SVs satisfy this requirement, yielding $0.5 > 0.01$: a bet Alice could hardly have placed otherwise explains her win less than one she could as easily have placed on tails.

A.5 Extended Why Queries

Prior work introduces extended why-queries that allow for partial evidence about the world [26].

Definition A.8 (Extended Why Query)**.** Given SCM $\mathcal{M}$ with endogenous variables $\mathbf{V}$, an extended why-query is specified by an event explanandum $Y = y$ (the fact to be explained), explanatory variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$, and evidence $\mathbf{E} = \mathbf{e}$ where $\mathbf{E} \subseteq \mathbf{V}$ and $\{Y\} \subseteq \mathbf{E}$. We denote this query $\text{Why}(Y|\mathbf{e}; \mathbf{X})$, or simply $\text{Why}(Y|\mathbf{e})$ when $\mathbf{X}$ is clear from context. □

The evidence $\mathbf{e}$ determines the posterior distribution over exogenous variables $P(\mathbf{U}|\mathbf{e})$, identifying which units are consistent with what we observed. Two important special cases:

1. **Full evidence**: $\text{Why}(Y|\mathbf{x})$ where $\mathbf{e} = \{\mathbf{X} = \mathbf{x}, Y = y\}$. This is the standard query used throughout the main text, identifying a specific instance to explain.
2. **Outcome-only evidence**: $\text{Why}(Y|y)$ where $\mathbf{e} = \{Y = y\}$. This identifies all instances where $Y = y$ occurred, explaining the class rather than a specific instance.

The choice of evidence reflects different epistemic states of the explainee: full evidence queries ask why $Y = y$ occurred given observed $\mathbf{X} = \mathbf{x}$ (actual causation), while outcome-only queries ask why $Y = y$ occurred without knowing $\mathbf{X}$'s values (general causation), aggregating over uncertainty in the mechanism.

A.6 From Probability of Necessity to Natural Total Effects

Our approach builds on Pearl's probability of necessity [13], which formalizes the but-for intuition that X is a cause of Y if but for X, Y would not have occurred. Since the exogenous variables $\mathbf{U}$ are unobserved, PN conditions on observed evidence to induce a posterior distribution over $\mathbf{U}$ that enables counterfactual evaluation. For a single variable X and outcome Y, the probability of necessity is $\text{PN}(x, y|\mathbf{x}, y) := P(Y_{x'} \neq y|\mathbf{X} = \mathbf{x}, Y = y)$, where $x \in \mathbf{x}$ is the observed value of X, x' is a baseline value, and $\mathbf{x}$ denotes the observed context for all variables. This captures actual causation by asking whether $X = x$ was necessary for $Y = y$ in this instance given the observed evidence $\mathbf{X} = \mathbf{x}, Y = y$, not whether X generally affects Y. The flexibility of PN with respect to evidence distinguishes it from measures like average treatment effect or probability of causation, which condition on fixed evidence sets.

NTEs generalize PN to address three key limitations. First, PN is binary, considering only whether Y changes rather than by how much, while NTEs measure the expected magnitude $\mathbb{E}[Y - Y_{x'}|\mathbf{x}, y]$. Second, PN handles a single variable with a single baseline, while NTEs aggregate over all baselines $(x', \mathbf{z}')$ weighted by their probability under $P(\mathbf{U}|\mathbf{x}, y)$, and extend to coalitions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$ to enable Shapley value decomposition. Third, PN provides no mechanism to distribute credit among multiple causes, while L_3 SVs average over coalition-specific NTEs weighted by baseline probabilities, satisfying efficiency and symmetry.

The connection between PN and NTE is direct in binary settings. When X, Y are binary and observed values are $x = 1, y = 1$, we have:

$$\begin{aligned}
\text{NTE}(X, Y|\mathbf{x}, y) &= \mathbb{E}[Y|\mathbf{x}, y] - \mathbb{E}_{\mathbf{u}' \sim P(\mathbf{U})}[Y_{X(\mathbf{u}')}|\mathbf{x}, y] && (11)\\
&= 1 - (P(x) \cdot 1 + P(x')P(y_{x'}|\mathbf{x}, y)) && \text{[using } \mathbb{E}[Y|\mathbf{x}, y] = y = 1] \quad (12)\\
&= 1 - P(x) - P(x')P(y_{x'}|\mathbf{x}, y) && (13)\\
&= P(x') - P(x')P(y_{x'}|\mathbf{x}, y) && \text{[using } P(x') = 1 - P(x)] \quad (14)\\
&= P(x')(1 - P(y_{x'}|\mathbf{x}, y)) && (15)\\
&= P(x')\text{PN}(x, y|\mathbf{x}, y). && (16)
\end{aligned}$$

This shows that NTE is a probability-weighted version of PN for binary variables. Notably, this derivation reveals that PN measures total effects rather than merely direct effects, since the counterfactual $Y_{x'}$ captures all causal pathways from X to Y. This total effect property makes PN and NTEs appropriate for causal explanation, where we ask which variables are responsible for the outcome through any mechanism, not just through direct influence.

A.7 NTE-Functional Dependence Equivalence

Lemma A.9 (NTE-Functional Dependence Equivalence)**.** *Let $\mathcal{M}$ be a recursive SCM with discrete exogenous domain* $\text{Dom}(\mathbf{U})$*, and let $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$. For any baseline $b = (x', \mathbf{z}') \in \text{Dom}(X) \times \text{Dom}(\mathbf{Z})$ with $P_{X,\mathbf{Z}}(x', \mathbf{z}') > 0$, functional dependence c (Def. 2.3) equals the marginal contribution of X to an* NTE *of $\mathbf{Z} \cup \{X\}$ on Y under transition $\mathbf{u}' \to \mathbf{u}$:*

$$\begin{aligned}
c(\mathcal{M}, \mathbf{u}, X, Y, b) =& \text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u})\\
&- \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) \qquad (17)
\end{aligned}$$

for any $\mathbf{u}'$ such that $(X(\mathbf{u}'), \mathbf{Z}(\mathbf{u}')) = (x', \mathbf{z}')$ (the right-hand side does not depend on the choice). For continuous $\text{Dom}(\mathbf{U})$*, the equivalence holds $P_{X,\mathbf{Z}}$-almost everywhere via disintegration.* □

The proof is given in Sec. D.2.

A.8 Normality Theory

A.8.1 Abnormal Causes vs Normal Baselines

Kahneman and Miller [30] discuss two related but distinct concepts: (1) preference for abnormal causes, and (2) preference for causes that "could [more] easily have been otherwise." These concepts often correlate but are fundamentally distinct.

Example A.10 (Forest Fire)**.** A forest fire occurs ($Y = 1$) when both lightning strikes ($X_1 = 1$) and oxygen is present ($X_2 = 1$). Lightning strikes are rare ($X_1 = 1$ occurs 1% of the time) while oxygen is nearly always present ($X_2 = 1$ occurs 99% of the time). The structural equation is $Y = X_1 \wedge X_2$ where $X_1 = U_1 \sim \text{Bernoulli}(0.01)$, $X_2 = U_2 \sim \text{Bernoulli}(0.99)$, and $U_1 \perp U_2$.

We observe the event $\mathbf{v} = \{X_1 = 1, X_2 = 1, Y = 1\}$. Both X_1 and X_2 technically cause Y: had either lightning or oxygen been absent, the forest fire would not have occurred. However, human intuition identifies lightning as the cause of the forest fire, not oxygen.

Applying Kahneman and Miller's heuristics [30], the "abnormal causes" heuristic suggests X_1 receives higher attribution since $X_1 = 1$ is abnormal (occurs 1% of the time). Similarly, the "could easily have been otherwise" heuristic also favors X_1 since $X_1 = 0$ has probability 0.99. In this case, both heuristics agree: human intuition prefers X_1 as the cause. □

Here, the abnormal value ($X_1 = 1$) has a high-probability baseline ($X_1 = 0$ occurs with probability 0.99). This is typical for binary variables: if value 1 is abnormal, then value 0 is the normal (high-probability) baseline. Note that X_1 and X_2 have isomorphic counterfactuals relative to Y in this SCM, exhibiting identical functional dependence structure. Therefore our normality desideratum D_4 distinguishes them based solely on baseline probabilities in an intra-SCM comparison.

Example A.11 (Team Selection)**.** Bob applies for a competitive sports team that requires both athletic ability and prior experience. The selection process is modeled as:

$$\mathcal{F} := \begin{cases} X_1 & := U_1 \quad \text{(prior experience: 0=none, 1=amateur, 2=professional)} \\ X_2 & := U_2 \quad \text{(physical fitness: 0=poor, 1=average, 2=excellent)} \\ Y & := \mathbf{1}[X_1 > 0] \wedge \mathbf{1}[X_2 > 0] \quad \text{(selected for team)} \end{cases} \tag{18}$$

with applicant pool distributions:

$$P(\mathbf{U}) = \begin{cases} U_1 \sim \text{Categorical}([0.97, 0.02, 0.01]) & \text{(most have no experience)} \\ U_2 \sim \text{Categorical}([0.01, 0.01, 0.98]) & \text{(self-selection: most in excellent shape)} \end{cases} \tag{19}$$

where U_1, U_2 are independent. Bob has amateur experience ($X_1 = 1$) and average fitness ($X_2 = 1$), and was selected ($Y = 1$).

Both variables cause the outcome. Applying Kahneman and Miller's heuristics [30], the "abnormal causes" heuristic suggests amateur experience ($X_1 = 1$, probability 2%) receives slightly higher attribution than average fitness ($X_2 = 1$, probability 1%). However, the "could easily have been otherwise" heuristic strongly favors experience: 97% of applicants have no experience versus only 1% with poor fitness. The baseline probability dominates: Bob's amateur experience is the better explanation because lack of experience is far more typical in the applicant pool. □

This demonstrates that when the two heuristics conflict, baseline probability determines explanatory salience: despite average fitness being slightly more abnormal than amateur experience, the experience is the better explanation because it "could more easily have been otherwise."

Probability-weighted baselines. Our L_3 Shapley values use probability-weighted baselines from the natural distribution $P(\mathbf{U})$, formalizing the "could easily have been otherwise" intuition. For variable X with observed value x and evidence $\mathbf{v}$, the NTE weights each baseline $(x', \mathbf{z}')$ by $P(\mathbf{U} = \mathbf{u}')$ where $\mathbf{u}'$ induces those baseline values. This weighting prioritizes baselines with higher probability. In binary settings (Ex. A.10), the abnormal value has the normal value as its baseline, so both heuristics coincide. In ternary settings (Ex. A.11), the same pattern holds when the abnormal value has a highly probable baseline. This formalization explains empirical findings in Icard et al. [50] showing context-dependent normality judgments: humans weight both the abnormality of the actual cause and the probability of alternative baselines. The next section contrasts our probability-weighted approach with Halpern's single-baseline method [45].

A.8.2 Single vs Probability-Weighted Baselines

Halpern and Hitchcock [45] introduces graded causation, ranking causes by their "best witness" (or best baseline in our terminology), according to an unspecified abstract ranking criterion. This single-baseline approach selects the most normal baseline and uses only that baseline for grading causal strength. We illustrate how this approach fails to capture the full distribution of baseline probabilities when the ranking criterion is concretely specified (e.g., baseline probability), leading to counterintuitive rankings for many-valued variables.

Example A.12 (Guessing Game)**.** Alice, Bob, and Carol win money if they all guess 0 from numbers $\{0, \ldots, 9\}$. Consider SCM $\mathcal{M}$ with guesses $X_1, X_2, X_3 \in \{0, \ldots, 9\}$ and structural equations $\mathcal{F}$ and exogenous distributions $P(\mathbf{U})$ defined as:

$$\mathcal{F} := \begin{cases} X_1 = U_1,\ X_2 = U_2,\ X_3 = U_3 \\ Y = \mathbf{1}[X_1 = 0 \wedge X_2 = 0 \wedge X_3 = 0] \end{cases}, \quad P(\mathbf{U}) = \begin{cases} U_1 \sim \text{Cat}_{10}([0.9, 0.1, 0, \ldots]) & \text{(Alice)} \\ U_2 \sim \text{Unif}(\{0, \ldots, 9\}) & \text{(Bob)} \\ U_3 \sim \text{Cat}_{10}([0.19, 0.09, \ldots, 0.09]) & \text{(Carol)} \end{cases} \tag{20}$$

where $\text{Cat}_{10}([\cdot])$ denotes a categorical distribution over $\{0, \ldots, 9\}$. Alice's favorite number is 0 but sometimes likes 1 ($P(U_1 = 0) = 0.9$, $P(U_1 = 1) = 0.1$, others zero). Bob guesses uniformly at random ($P(U_2 = k) = 0.1$ for all k). Carol has slight preference for 0 ($P(U_3 = 0) = 0.19$, $P(U_3 = k) = 0.09$ for $k \geq 1$).

Observe $\mathbf{v} = \{X_1 = 0, X_2 = 0, X_3 = 0, Y = 1\}$ (they all guessed 0 and won). All three variables are counterfactually isomorphic: changing any from 0 to any other value changes Y from 1 to 0. However,

they differ in how easily they "could have been otherwise": Alice has only 10% chance of guessing otherwise (only alternative is $X_1 = 1$), Bob has 90% chance across nine equally-likely alternatives at 10% each, and Carol has 81% chance across nine alternatives at 9% each. Intuitively, Bob was very lucky to guess 0 (90% chance of guessing otherwise), Carol was quite lucky (81% chance), while Alice was fairly certain (only 10% chance of guessing 1). Thus Bob and Carol are better explanations than Alice.

Halpern's single-baseline method selects the best witness by baseline probability: Alice's best baseline is $X_1 = 1$ with $P = 0.1$, Bob's best baseline is any $X_2 \in \{1, \ldots, 9\}$ with $P = 0.1$ each, and Carol's best baseline is any $X_3 \in \{1, \ldots, 9\}$ with $P = 0.09$ each. Halpern ranks them Alice $\approx$ Bob $>$ Carol (by best witness probability), preferring Alice over Carol although Carol has a much higher probability (0.81 versus 0.1) of guessing another number.

To apply normality (D_4), compare Bob to Alice, for an EVA unchanged by relabeling variables, by constructing $\mathcal{M}'$ where we swap U_1 and U_2. Bob has strictly greater aggregate baseline probability than Alice (0.9 versus 0.1) with identical counterfactual structure, so D_4 requires $\phi_{\text{Bob}} > \phi_{\text{Alice}}$. Comparing Carol to Alice: Carol has strictly greater aggregate (0.81 versus 0.1), but $P(X_3 = 1) = 0.09$ is below $P(X_1 = 1) = 0.1$, so D_4, which compares each baseline, makes no claim. Between Bob and Carol, each baseline $k \geq 1$ has probability 0.1 for Bob and 0.09 for Carol, so D_4 requires $\phi_{\text{Bob}} > \phi_{\text{Carol}}$. □

Single-baseline ranking violates normality. Halpern's method ranks Alice $\approx$ Bob $>$ Carol by best witness probability alone, missing the aggregate distribution. This violates normality (D_4): Bob has a much higher probability than Alice of guessing another number and the same best witness probability, yet the method yields an equal attribution for each. The method also ranks Alice above Carol, although Carol has a much higher probability than Alice of guessing another number.

Our L_3 SVs resolve this by weighting all baselines by their probability in the Shapley decomposition. Computing the exact L_3 SVs yields:

$$\phi^{L_3}_{\text{Alice}} = 0.039 \tag{21}$$

$$\phi^{L_3}_{\text{Bob}} = 0.515 \tag{22}$$

$$\phi^{L_3}_{\text{Carol}} = 0.429 \tag{23}$$

These values are computed by enumerating all $2^3 = 8$ coalitions and averaging the marginal NTE contributions weighted by coalition probabilities, using the Shapley formula from Def. 3.2. Thus L_3 SVs correctly rank Bob $>$ Carol $>$ Alice, matching the ranking of baseline probabilities $P(X_2 \neq 0) = 0.90 > P(X_3 \neq 0) = 0.81 > P(X_1 \neq 0) = 0.10$. While the attributions do not equal the baseline probabilities directly due to Shapley's credit distribution across coalitions, they preserve the correct ranking and satisfy normality (D_4) by assigning higher attribution to variables with more probable baselines.

A.9 Measure on SCMs

We establish that D_2 (Power) has a measure-zero exception: it holds for all structural parameters outside a set of measure zero. Two cases must be distinguished. First, structural zeros: when diagram G contains no causal pathway from X to Y, functional dependence is structurally zero regardless of parameter values. Second, pathological cancellations: when G does contain an $X \to Y$ pathway, zero L_3 SV despite nonzero functional dependence requires all Shapley marginal contributions to cancel exactly. We show this cancellation occurs with measure zero, first for the structural parameters below (Lem. D.2) and for vectors of marginal contributions (Lem. A.14), and then for continuous parametric families (Lem. A.15).

Structural parameters. Fix a causal diagram G and a finite domain $\text{Dom}(\mathbf{U})$. In topological order of G, let an input of f_V consist of one input of each endogenous parent of V and one value of the exogenous parents of V. The structural parameters are $P(\mathbf{U})$ and one real number $f_V(i)$ for each $V \in \mathbf{V}$ and each input i of f_V. We use Lebesgue measure on these real numbers and on the probability simplex over $\text{Dom}(\mathbf{U})$. Outside a set of measure zero, f_V takes distinct values at distinct inputs, so these numbers define an SCM with finite domains, in which the value of V is f_V at its input.

NTE basis. Fix a causal diagram G, exogenous distribution $P(\mathbf{U})$, evidence $\mathbf{v}$, and variable X. For any SCM $\mathcal{M}$ compatible with G and $P(\mathbf{U})$, define the **NTE basis** as the indexed family

$$\mathcal{B}_X^{\mathrm{NTE}}(\mathcal{M}, \mathbf{v}) = (\mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}))_{(\mathbf{Z},\mathbf{u}',\mathbf{u}) \in I_{\mathcal{M}}} \tag{24}$$

where $I_{\mathcal{M}} = \{(\mathbf{Z}, \mathbf{u}', \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{u}', \mathbf{u} \in \mathrm{Dom}(\mathbf{U})^{\mathcal{M}}, \mathbf{V}_{\mathcal{M}}(\mathbf{u}) = \mathbf{v}\}$. When $\mathrm{Dom}(\mathbf{U})$ is discrete, $|I_{\mathcal{M}}|$ is finite, and the NTE basis is a vector in $\mathbb{R}^{|I_{\mathcal{M}}|}$. The NTE basis captures the counterfactual behavior of $\mathcal{M}$: the index set $I_{\mathcal{M}}$ fixes which exogenous values are consistent with the evidence (factual behavior), and varying the basis values varies only what would happen under interventions. Crucially, because $P(\mathbf{U})$ and the consistency set $\{\mathbf{u} : \mathbf{V}_{\mathcal{M}}(\mathbf{u}) = \mathbf{v}\}$ are both held fixed, the abduction weights $P(\mathbf{u}|\mathbf{e}) = P(\mathbf{u}) / \sum_{\mathbf{u}'':\mathbf{V}(\mathbf{u}'')=\mathbf{v}} P(\mathbf{u}'')$ are constant across the NTE basis space.

A.9.1 Cancellation Example

We illustrate the measure-0 exception to strong power with a concrete example.

Example A.13 (Shapley Cancellation)**.** Consider a linear SCM $\mathcal{M}$ with variables X_1, X_2, Y and structural equations

$$\mathcal{F} = \begin{cases} X_1 &= u_1, \\ X_2 &= \gamma x_1 + u_2, \\ Y &= \alpha x_1 + \beta x_2 + u_3, \end{cases} \tag{25}$$

where U_1, U_2, U_3 are independent exogenous noise variables with zero mean. The L_3 SV for X_1 is

$$\phi_{X_1}^{L_3} = \frac{1}{2}\left(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]\right) + \frac{1}{2}\left(\mathbb{E}[Y_{P(X_2)}|\mathbf{v}] - \mathbb{E}[Y_{P(X_1,X_2)}|\mathbf{v}]\right) \tag{26}$$

$$= \frac{1}{2}(\alpha + \beta\gamma)x_1 + \frac{1}{2}\alpha x_1 = \frac{2\alpha + \beta\gamma}{2} x_1. \tag{27}$$

When $\alpha = -\frac{\beta\gamma}{2}$, we have $\phi_{X_1}^{L_3} = 0$ despite X_1 having both nonzero total effect $(\alpha + \beta\gamma)$ and direct effect α on Y. This is a *cancellation*: the marginal contribution from coalition $\{X_1, X_2\}$ exactly negates the marginal contribution from coalition $\{X_1\}$. □

The cancellation condition $\alpha = -\frac{\beta\gamma}{2}$ defines a codimension-1 submanifold in the parameter space $(\alpha, \beta, \gamma) \in \mathbb{R}^3$, which has Lebesgue measure zero. This illustrates the general result:

Lemma A.14 (Measure-Zero Cancellation, Discrete Case)**.** *Fix a causal diagram G with a pathway from X to Y, a finite exogenous domain* $\mathrm{Dom}(\mathbf{U})$*, an exogenous distribution $P(\mathbf{U})$, and evidence $\mathbf{v}$ of positive probability. The cancellation set $\{\mathbf{b} \in \mathbb{R}^{|I_{\mathcal{M}}|} : \phi_X^{L_3}(\mathbf{b}) = 0$ and $\mathbf{b} \neq 0\}$ has Lebesgue measure zero in $\mathbb{R}^{|I_{\mathcal{M}}|}$ under the product Lebesgue measure.* □

Proof. Since $P(\mathbf{U})$ and the consistency set $\{\mathbf{u} : \mathbf{V}(\mathbf{u}) = \mathbf{v}\}$ are fixed, the abduction weights $P(\mathbf{u}|\mathbf{e})$ and prior weights $P(\mathbf{u}')$ are constants independent of the NTE basis $\mathbf{b}$. The L_3 Shapley value $\phi_X^{L_3}(\mathbf{b}) = \sum_S w_S \sum_{\mathbf{u},\mathbf{u}'} P(\mathbf{u}|\mathbf{e})\, P(\mathbf{u}')\, b_{S,\mathbf{u}',\mathbf{u}}$ is therefore a linear function of $\mathbf{b}$ with fixed coefficients. The condition $\phi_X^{L_3}(\mathbf{b}) = 0$ defines a hyperplane of codimension 1 in $\mathbb{R}^{|I_{\mathcal{M}}|}$, which has Lebesgue measure zero. □

As the values of f_Y vary with the other structural equations fixed, the marginal contributions range over a linear subspace of $\mathbb{R}^{|I_{\mathcal{M}}|}$, so Lem. D.2 states the corresponding result for structural parameters.

Lemma A.15 (Measure-Zero Cancellation, Parametric Case)**.** *Fix a causal diagram G with a pathway from X to Y, and consider a parametric SCM family $\{\mathcal{M}_\theta : \theta \in \Theta\}$ compatible with G, where Θ is a connected open subset of $\mathbb{R}^d$. If $\phi_X^{L_3}(\theta) := \phi_X^{L_3}(\mathcal{M}_\theta, w)$ is an analytic function of θ that is not identically zero, then the cancellation set $\{\theta \in \Theta : \phi_X^{L_3}(\theta) = 0$ and $c(\theta) \neq 0\}$ has Lebesgue measure zero in $\mathbb{R}^d$.* □

Proof. Since $\phi_X^{L_3}(\theta)$ is analytic and not identically zero, its zero set has Lebesgue measure zero in $\mathbb{R}^d$ [51, Theorem 6.3.3]. The cancellation set is a subset and therefore also has measure zero. □

Lem. A.14 covers vectors of marginal contributions over finite exogenous domains without any parametric assumption on the structural equations. Lem. A.15 extends to continuous exogenous domains, where the NTE basis is infinite-dimensional and we instead work in the finite-dimensional parameter space Θ. The "not identically zero" condition holds whenever there exists some θ^* with $\phi_X^{L_3}(\theta^*) \neq 0$ (e.g., Ex. A.13 exhibits nonzero ϕ at generic parameter values when $x_1 \neq 0$). The analyticity condition holds in Ex. A.13, where $\phi_{X_1}^{L_3}$ is a polynomial in the parameters.

A.9.2 Extended L_3 Shapley Values

The measure-zero exception for D_2 can be removed by augmenting the L_3 SV with the variance of the marginal NTE contribution.

Definition A.16 (Extended L_3 Shapley Value)**.** Given SCM $\mathcal{M}$, why-query w, and variable $X \in \mathbf{X}$, the extended L_3 Shapley value is the pair

$$\Phi_X^{L_3}(w) = (\mathbb{E}[\Delta_X],\ \mathrm{Var}[\Delta_X]) \tag{28}$$

where $\Delta_X = f(\pi_{\leq X}) - f(\pi_{<X})$ is the marginal NTE contribution under $\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})$, $\mathbf{u} \sim P(\mathbf{U}|\mathbf{e})$, $\mathbf{u}' \sim P(\mathbf{U})$. We say $\Phi_X^{L_3} = 0$ iff both components vanish, and we order extended L_3 SVs by their first components. □

Theorem A.17 (Extended L_3 SVs satisfy $D_{1:4}$ for all SCMs)**.** *$\Phi_X^{L_3}$ satisfies $D_{1,2,3,4}$ for all SCMs.* □

Proof. D_1, D_3, D_4 inherit from Thm. 3.3: the expectation component already satisfies them, and $\Phi_X^{L_3} = 0$ requires both $\mathbb{E}[\Delta_X] = 0$ and $\mathrm{Var}[\Delta_X] = 0$, which holds iff $\Delta_X = 0$ almost surely. For D_2: when X is a but-for cause via functional dependence, $\Delta_X \neq 0$ with positive probability, since X comes first in π with probability $1/|\mathbf{X}|$ (Lem. A.9). If $\mathbb{E}[\Delta_X] \neq 0$, $\Phi_X^{L_3} \neq 0$. If $\mathbb{E}[\Delta_X] = 0$ (the cancellation case), then $\mathrm{Var}[\Delta_X] > 0$, so $\Phi_X^{L_3} \neq 0$. □

A.10 Startup Funding SCM

The startup funding example (Ex. 2.1, Fig. 3) uses the following SCM:

$$\mathcal{M} = \begin{cases} U_i & \sim \mathrm{Bern}(0.05), \quad \forall i \in \{1,2,3,6\} \\ U_7 & \sim \mathrm{Bern}(0.95) \\ U_j^k & \sim \mathrm{Bern}(0.95) \text{ if } k = 3, \quad \mathrm{Bern}(0.05) \text{ otherwise}, \quad \forall j \in \{4,5\}, k \in \{2,3\} \\ X_i & = U_i, \quad \forall i \in \{1,2,3,6,7\} \\ X_j & = X_1 \wedge ((X_2 \oplus U_3) \vee U_j^2) \wedge U_j^3, \\ & \qquad \forall j \in \{4,5\} \\ Y & = 8X_4 + X_5 + X_6 X_7 \end{cases}$$

Variables X_1, X_2, X_3, X_6, X_7 represent Alice's eloquence, the live demo, the flickering lights, winning the AI grant, and passing the background check, each determined directly by exogenous factors. The funding variables X_4 and X_5 capture whether Alice secured Series A and angel funding; they depend on her eloquence and the demo's effectiveness, where the demo is neutralized by a power fluctuation ($U_3 = 1$) through the XOR interaction $X_2 \oplus U_3$. Alice's total funding of \$10M ($Y = 10$) decomposes as $8 \cdot X_4 + X_5 + X_6 \cdot X_7$, reflecting \$8M from Series A, \$1M from angel funding, and \$1M from the AI grant contingent on both winning the competition and passing the background check.

In Alice's actual situation, $U_3 = 1$ (power fluctuation), so the demo (X_2) had no causal effect: $(X_2 \oplus U_3) = (1 \oplus 1) = 0$. She succeeded due to being well-spoken ($X_1 = 1$), not the live demo. Thus the live demo correctly receives zero attribution ($\phi_{X_2}^{L_3} = 0$) per admissibility (D_1).

A.11 Shapley Axioms

Since L_3 SVs are defined as Shapley values over the NTE value function, they inherit the standard Shapley axioms, which provide theoretical guarantees for attribution decomposition: null player (variables that never affect the outcome receive zero attribution), efficiency (attributions sum to the total effect), symmetry (variables with identical contributions receive identical attributions), and marginality (attribution depends only on marginal contributions, not the absolute value function).

Corollary A.18 (L_3 Shapley axioms)**.** *L_3 SVs satisfy Shapley axioms for value function $f(\mathbf{Z}) = \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{e})$:*

(1) ***Null player****: If $f(\mathbf{S} \cup \{X\}) = f(\mathbf{S})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$, then $\phi_X^{L_3}(w) = 0$*

(2) ***Efficiency****: $\sum_{X \in \mathbf{X}} \phi_X^{L_3}(w) = f(\mathbf{X}) - f(\emptyset)$*

(3) ***Symmetry****: If $f(\mathbf{S} \cup \{X_i\}) = f(\mathbf{S} \cup \{X_j\})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$, then $\phi_{X_i}^{L_3}(w) = \phi_{X_j}^{L_3}(w)$*

(4) ***Marginality****: If $f(\mathbf{S} \cup \{X\}) - f(\mathbf{S}) = f'(\mathbf{S} \cup \{X\}) - f'(\mathbf{S})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$, then $\phi_X^{L_3}(w) = \phi_X'^{L_3}(w)$ where ϕ' is computed using f'* □

The null player axiom ensures marginally irrelevant variables receive zero attribution without requiring explicit minimality constraints as in actual-causation-based explanation frameworks [26]. The proof is given in Sec. D.6.

A.12 When L_1, L_2, and L_3 SVs Coincide

Proposition A.19 (Structural conditions for $L_2 = L_3$ and $L_1 = L_2 = L_3$ SVs)**.** *Assume (i) f_Y depends on $\mathbf{X}$ only (no exogenous noise U_Y), (ii) no causation among explanatory variables ($X_i \not\rightarrow X_j$ for all $i \neq j$), and (iii) no confounding among explanatory variables (the variables in $\mathbf{X}$ are mutually independent). Let $\phi_X^{L_i}$ denote the Shapley value of value function v^{L_i}. Then:*

(a) Under (i)–(ii), $\phi_X^{L_2} = \phi_X^{L_3}$ for all $X \in \mathbf{X}$.

(b) Under (i)–(iii), $\phi_X^{L_1} = \phi_X^{L_2} = \phi_X^{L_3}$ for all $X \in \mathbf{X}$. □

Proof. Define value functions $v^{L_1}(\mathbf{Z}) = \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}]$ (observational), $v^{L_2}(\mathbf{Z}) = \mathbb{E}[Y|\mathrm{do}(\mathbf{Z} = \mathbf{z})]$ (interventional), and $v^{L_3}(\mathbf{Z}) = \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{e})$ (counterfactual, Def. 3.1), evaluated at observed values $\mathbf{z}$. We show $\phi_X^{L_2} = \phi_X^{L_3}$, then $\phi_X^{L_1} = \phi_X^{L_2}$.

Step 1 ($L_2 = L_3$): By Def. 3.1, $v^{L_3}(\mathbf{Z}) = \mathbb{E}_{\mathbf{u},\mathbf{u}'}[Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})]$ where $\mathbf{u} \sim P(\mathbf{U}|\mathbf{e})$ and $\mathbf{u}' \sim P(\mathbf{U})$. Under (ii), setting any $\mathbf{Z} \subseteq \mathbf{X}$ to its factual values $\mathbf{Z}(\mathbf{u})$ does not affect any other variable (no downstream effects within $\mathbf{X}$), so $Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) = Y(\mathbf{u})$ for all coalitions $\mathbf{Z}$. This first term is constant across coalitions, so we can write $v^{L_3}(\mathbf{Z}) = c - \mathbb{E}_{\mathbf{u},\mathbf{u}'}[Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})]$ where $c = \mathbb{E}_{\mathbf{u}}[Y(\mathbf{u})]$. Since Shapley marginal contributions $v(\mathbf{S} \cup \{X\}) - v(\mathbf{S})$ cancel any constant, $\phi_X(v + c) = \phi_X(v)$, so $\phi_X^{L_3} = \phi_X(\tilde{v})$ where $\tilde{v}(\mathbf{Z}) = -\mathbb{E}_{\mathbf{u},\mathbf{u}'}[Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})]$.

Next, we apply complement symmetry. Let $\bar{\mathbf{Z}} = \mathbf{X} \setminus \mathbf{Z}$ and define $\hat{v}(\mathbf{Z}) = -\tilde{v}(\bar{\mathbf{Z}}) = \mathbb{E}_{\mathbf{u},\mathbf{u}'}[Y_{\bar{\mathbf{Z}}(\mathbf{u}')}(\mathbf{u})]$. We show $\phi_X(\hat{v}) = \phi_X(\tilde{v})$. Writing $\phi_X(\tilde{v}) = \sum_{\mathbf{S}: X \notin \mathbf{S}} w_{|\mathbf{S}|}[\tilde{v}(\mathbf{S} \cup \{X\}) - \tilde{v}(\mathbf{S})]$ and substituting $\tilde{v}(\mathbf{S}) = -\hat{v}(\bar{\mathbf{S}})$:

$$\begin{aligned}\phi_X(\tilde{v}) &= \sum_{\mathbf{S}: X \notin \mathbf{S}} w_{|\mathbf{S}|}[-\hat{v}(\overline{\mathbf{S} \cup \{X\}}) + \hat{v}(\bar{\mathbf{S}})] \\ &= \sum_{\mathbf{S}: X \notin \mathbf{S}} w_{|\mathbf{S}|}[\hat{v}(\bar{\mathbf{S}}) - \hat{v}(\bar{\mathbf{S}} \setminus \{X\})]\end{aligned}$$

Relabeling $\mathbf{T} = \bar{\mathbf{S}} \setminus \{X\}$ (so $|\mathbf{T}| = n - |\mathbf{S}| - 1$) and using weight symmetry $w_k = w_{n-k-1}$, we get $\phi_X(\tilde{v}) = \phi_X(\hat{v})$. Thus $\phi_X^{L_3} = \phi_X(\hat{v})$.

Finally, we show $\hat{v}(\mathbf{Z}) = v^{L_2}(\mathbf{Z})$. Under (ii), there is no causation among $\mathbf{X}$, so when $\mathbf{u}' \sim P(\mathbf{U})$, the values $\bar{\mathbf{Z}}(\mathbf{u}')$ are distributed as $P(\bar{\mathbf{Z}})$. Under (i), $Y = f_Y(\mathbf{X})$ has no exogenous noise, so $\hat{v}(\mathbf{Z}) = \mathbb{E}_{\bar{\mathbf{Z}} \sim P(\bar{\mathbf{Z}})}[f_Y(\mathbf{z}, \bar{\mathbf{Z}})]$. The intervention $\mathrm{do}(\mathbf{Z} = \mathbf{z})$ sets $\mathbf{Z}$ to $\mathbf{z}$ while $\bar{\mathbf{Z}}$ follows its natural distribution; under (ii), $P(\bar{\mathbf{Z}}|\mathrm{do}(\mathbf{Z})) = P(\bar{\mathbf{Z}})$, so $v^{L_2}(\mathbf{Z}) = \mathbb{E}_{\bar{\mathbf{Z}}}[f_Y(\mathbf{z}, \bar{\mathbf{Z}})] = \hat{v}(\mathbf{Z})$.

Step 2 ($L_1 = L_2$): Under (ii)–(iii), $P(\bar{\mathbf{Z}}|\mathbf{Z} = \mathbf{z}) = P(\bar{\mathbf{Z}})$ (no confounding or causation among $\mathbf{X}$). Thus $v^{L_1}(\mathbf{Z}) = \mathbb{E}[Y|\mathbf{Z} = \mathbf{z}] = \mathbb{E}_{\bar{\mathbf{Z}}}[f_Y(\mathbf{z}, \bar{\mathbf{Z}})] = v^{L_2}(\mathbf{Z})$, and their Shapley values coincide. □

A.13 Causal Explained Variance

The explanatory variables $\mathbf{X}$ are chosen by the analyst, raising the question: what if $\mathbf{X}$ excludes an important cause? We want a diagnostic that tells us whether to expand $\mathbf{X}$.

One approach is to use the coefficient of determination R^2 from regression, which measures explained variance. However, R^2 can be misleading under confounding, as the following examples show.

Example A.20 (Confounding inflates R^2)**.** Let $U_T, U_X, U_Y \sim N(0,1)$ be independent, with $T = U_T$, $X = T + U_X$, and $Y = T + U_Y$. □

Example A.21 (Confounding masks R^2)**.** Let $U, U_X, U_Y \sim N(0,1)$ be independent, with $X = U + U_X$ and $Y = X - 2U + U_Y$. □

In Example A.20, X does not cause Y, but $\text{Cov}(X,Y) = \text{Var}(T) = 1$ and $\text{Var}(X) = \text{Var}(Y) = 2$, so $R^2 = \text{Corr}(X,Y)^2 = 0.25$. Thus $R^2 > 0$ suggests X explains Y, but X does not cause Y. In Example A.21, X causes Y with coefficient 1, but $\mathbb{E}[U|X = x] = x/2$, so $\mathbb{E}[Y|X = x] = x - 2\cdot(x/2) = 0$, giving $R^2 = 0$. Thus $R^2 = 0$ suggests X does not explain Y, but X does cause Y. We need a causal version of R^2.

Given query $w = \text{Why}(y|\mathbf{e};\mathbf{X})$ and $\mathbf{u}' \sim P(\mathbf{U})$, consider the surprise

$$\Delta(\mathbf{u}' \to \mathbf{u}) = Y(\mathbf{u}) - Y(\mathbf{u}'), \tag{29}$$

the difference between the observed and baseline outcomes. We extend the concept of explained variance R^2 to causal explained variance,

$$R_c^2 = 1 - \frac{\text{Var}[\text{NTE}(\mathbf{X}, Y|\mathbf{u}' \to \mathbf{u}) - \Delta(\mathbf{u}' \to \mathbf{u})]}{\text{Var}[\Delta(\mathbf{u}' \to \mathbf{u})]} \tag{30}$$

$$= 1 - \frac{\text{Var}[Y_{\mathbf{X}(\mathbf{u}')}(\mathbf{u}) - Y(\mathbf{u}')]}{\text{Var}[Y(\mathbf{u}) - Y(\mathbf{u}')]} \tag{31}$$

where $\mathbf{u}' \sim P(\mathbf{U})$ is the baseline unit, as in Def. 3.2. In words, this quantity captures the proportion of surprise that is causally explained by $\mathbf{X}$. Equivalently, R_c^2 is the R^2 of the NTE as a predictor of the surprise Δ. When $R_c^2 = 1$, the explanatory variables account for all variance in the surprise. When $R_c^2 \ll 1$, important causes may be missing from $\mathbf{X}$.

In Example A.20, since $X \not\to Y$, we have $Y_{X(\mathbf{u}')}(\mathbf{u}) = Y(\mathbf{u})$ for all $\mathbf{u}'$, so the NTE is 0 and $R_c^2 = 0$. In Example A.21, $X \to Y$ with coefficient 1, so intervening on X changes Y. The NTE captures part of the surprise, giving $R_c^2 > 0$.

Theorem A.22 (R_c^2 Rejects Non-Causes)**.** *If $X \not\to Y$ but $\mathbb{E}[Y|X]$ is non-constant due to confounding, then $R_c^2 = 0$ while the R^2 of the optimal Bayes regressor $\hat{Y} = \mathbb{E}[Y|X]$ is positive.* □

Proof. Since X does not cause Y, intervening on X has no effect: $Y_x(\mathbf{u}) = Y(\mathbf{u})$ for all x. Thus $Y_{\mathbf{X}(\mathbf{u}')}(\mathbf{u}) - Y(\mathbf{u}') = Y(\mathbf{u}) - Y(\mathbf{u}')$, so the NTE is 0 and $R_c^2 = 0$. For R^2, non-constant $\mathbb{E}[Y|X]$ means $\text{Var}[\mathbb{E}[Y|X]] > 0$ and $R^2 > 0$. □

Example A.20 is a special case of Thm. A.22. The next result shows that when confounding is absent and the evidence fixes $\mathbf{X}$, R_c^2 reduces to R^2.

Theorem A.23 (R_c^2 Generalizes R^2)**.** *Suppose the SCM underlying query w has additive structure $Y = g(\mathbf{X}) + U_Y$, where U_Y is zero-mean exogenous noise and $\mathbf{X} \perp\!\!\!\perp U_Y$, and the evidence $\mathbf{e}$ fixes $\mathbf{X}$. Then $R_c^2 = R^2$ of the optimal Bayes regressor $\hat{Y} = \mathbb{E}[Y|\mathbf{X}]$.* □

Proof. Note that knowing $\mathbf{u}'$ implies knowing $\mathbf{X}(\mathbf{u}')$. Since $\mathbf{e}$ fixes $\mathbf{X}$ and Y, it fixes $U_Y = Y - g(\mathbf{X})$, so each variance below is over $\mathbf{u}'$ alone. Then,

$$\begin{aligned}
R_c^2 &= 1 - \frac{\text{Var}[Y_{\mathbf{X}(\mathbf{u}')}(\mathbf{u}) - Y(\mathbf{u}')]}{\text{Var}[Y(\mathbf{u}) - Y(\mathbf{u}')]} \\
&= 1 - \frac{\text{Var}[g(\mathbf{X}') + U_Y - (g(\mathbf{X}') + U_Y')]}{\text{Var}[g(\mathbf{X}) + U_Y - (g(\mathbf{X}') + U_Y')]} \\
&= 1 - \frac{\text{Var}[U_Y - U_Y']}{\text{Var}[g(\mathbf{X}) + U_Y - g(\mathbf{X}') - U_Y']} \\
&= 1 - \frac{\text{Var}[\mathbb{E}[Y|\mathbf{X}] - Y]}{\text{Var}[Y]} \\
&= R^2
\end{aligned} \tag{32}$$

□

Thus R_c^2 provides the diagnostic sought above: low R_c^2 signals that $\mathbf{X}$ may be missing important causes, regardless of confounding.

Table 4: Comprehensive attribution method comparison on startup funding example (Ex. 2.1). All prior methods violate at least one proxy desideratum, implying violations of oracle desiderata. Only L_3 SVs satisfy all desiderata (D_1-D_4). Bolded values satisfy the expected behavior for that variable (see Def. 2.4), read before rounding: in the PS(x, y) and PNS(x, y) rows, every entry except ϕ_3 is positive and below 0.005. An entry of n/a means that the conditioning event of PS has probability zero. Dashes (—) indicate that we make no statement about whether the method satisfies the desideratum; only L_3 SVs have formal proofs. LIME values are means over 100 runs with maximum 95% CI error bar ±0.02 for $\phi_{1:6}$ and ±0.00 for ϕ_7. L_3 SV values show ground truth computed from known SCM. L_3 SV bounds $[\hat{\phi}^{L_3,-}, \hat{\phi}^{L_3,+}]$ computed by linear programming as described in App. C.

Method	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	D_1	D_2	D_3	D_4
Expected	$D_2 :\neq 0$	$D_1 := 0$	$D_1 := 0$	$D_2 :\neq 0$ $D_3 :> \phi_5$	$D_2 :\neq 0$ $D_3 :< \phi_4$	$D_2 :\neq 0$ $D_4 :> \phi_7$	$D_2 :\neq 0$ $D_4 :< \phi_6$				
L_1 Methods (Observational)											
LIME	0.00	**0.00**	**0.00**	**7.58**	**0.98**	**0.80**	0.00	–	×	–	–
ASV-prox	0.00	**0.00**	0.33	**5.20**	**3.41**	**0.93**	**0.03**	×	×	–	–
L_1 SV (SHAP)	**1.02**	−0.32	−0.32	**5.85**	**2.71**	**0.93**	**0.03**	×	–	–	–
ASV-distal	**3.63**	0.64	−1.63	**5.54**	**0.76**	**0.93**	**0.03**	×	–	–	–
L_2 Methods (Interventional)											
L_2 SV	**1.67**	1.27	**0.00**	**5.33**	**0.67**	**0.93**	**0.03**	×	–	–	–
L_3 Methods (Counterfactual)											
PN(x,y)	**1.00**	1.00	**0.00**	1.00	1.00	1.00	1.00	×	–	×	×
PN(x,y|**x**, y)	**1.00**	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	–	×	×
PS(x,y)	**0.00**	0.00	**0.00**	0.00	0.00	**0.00**	**0.00**	×	–	×	–
PS(x,y|**x**, y)	n/a	n/a	n/a	1.00	1.00	1.00	1.00	×	×	×	×
PNS(x,y)	**0.00**	0.00	**0.00**	0.00	0.00	**0.00**	**0.00**	×	–	×	–
PNS(x,y|**x**, y)	**1.00**	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	–	×	×
CF-Shapley	0.00	**0.00**	**0.00**	**8.00**	**1.00**	0.50	0.50	–	×	–	×
Beckers-CF	0.00	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	×	×	×
Beckers-Suff	0.00	**0.00**	**0.00**	1.00	1.00	1.00	1.00	–	×	×	×
Halpern-Pearl	0.00	**0.00**	**0.00**	0.00	0.00	0.00	0.00	–	×	×	×
L_3 SV (ours)	**4.28**	**0.00**	**0.00**	**4.15**	**0.52**	**0.93**	**0.03**	✓	✓	✓	✓
L_3 SV (min)	2.92	0.00	0.00	2.75	0.34	0.93	0.03				
L_3 SV (max)	4.28	2.92	0.00	4.15	0.52	0.93	0.03				

Table 5: Counterexample classes demonstrating desiderata violations. Each cell lists the simplest example classes that demonstrate a method's violation of the given desideratum. Detailed constructions in subsections below.

Method	D_1 (Admiss.)	D_2 (Power)	D_3 (Scaling)	D_4 (Normality)
SHAP (L_1 SVs)	Spurious, Neg Chain	–	Scaled	–
LIME*	–	Pos Chain, Conjunctive	–	–
L_2 SVs	Neg Chain	–	–	–
PN	Neg Chain	Disjunctive	Scaled	Conjunctive
CF-Shapley	–	Pos Chain	–	Conjunctive
Asymmetric Shapley	–	Pos Chain	–	–
Integrated Grads	–	Pos Chain	–	–
L_3 SVs (ours)	✓	✓	✓	✓

*L_1/L_2 regularization in LIME can reduce attributions to zero for variables that should receive non-zero attribution.

B Method Violations and Counterexamples

We first verify that Ex. 2.1 satisfies the lifting conditions from Thm. A.7, ensuring proxy violations transfer to oracle violations. We then discuss the extended results table (Table 4), organizing methods by type: Shapley-based methods, other ML explainability methods, probabilities of causation, and actual causation based explanations. We illustrate the computation of L_3 SVs for Alice's case. Finally, counterexamples are organized by conceptual mechanism: *confounding issues* (Spurious, Chains), *multivariate causation* (Disjunctive, Conjunctive), and *effect magnitude* (Scaled). The startup funding

example (Ex. 2.1, Table 4) combines multiple violation patterns, demonstrating systematic failures across methods. Each method's violations stem from fundamental structural limitations rather than implementation details.

Remark B.1 (Lifting conditions for Ex. 2.1)**.** We verify that Ex. 2.1 satisfies the lifting conditions in Thm. A.7 for all violations in Table 4.

For D_1 (Admissibility), violations lift unconditionally by conservativity. Methods assigning non-zero attribution to X_2 and X_3 violate $D_1(c)$ since $c = 0$ for these variables. By conservativity, $c = 0$ guarantees $c' = 0$, so these methods also violate $D_1(c')$.

For D_2 (Power), we verify that each of X_1 and $X_{4:7}$ is a but-for cause of $Y = 10$ in some unit consistent with the evidence. Variables $X_{4:7}$ are but-for causes in all consistent units: from $Y = 8X_4 + X_5 + X_6X_7$, we have $Y_{X_4=0} = 2$, $Y_{X_5=0} = 9$, $Y_{X_6=0} = 9$, and $Y_{X_7=0} = 9$, all differing from $Y = 10$. For X_1, consider the one unit consistent with the evidence, which has positive probability. In this unit, X_1 is but-for: setting $X_1 = 0$ yields $X_4 = X_5 = 0$, hence $Y = 0 + 0 + 1 = 1 \neq 10$. Since each variable is a but-for cause in some unit with positive probability, $c \neq 0$ implies $c' \neq 0$: but-for causes are actual causes under all major definitions [27, 28]. Methods assigning $\phi = 0$ to these variables violate both $D_2(c)$ and $D_2(c')$.

For D_4 (Normality), consider the comparison between X_6 (winning the grant) and X_7 (background check). To apply D_4 formally, we construct a comparison SCM $\mathcal{M}'$ by relabeling variables (swapping X_6 and X_7) while swapping their baseline distributions accordingly. Both variables are but-for causes: $Y_{X_6=0} = 9 \neq 10$ and $Y_{X_7=0} = 9 \neq 10$, so $c = c' \neq 0$ for both. Since $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$, violations transfer by Thm. A.7. Methods violating $D_4(c)$ by failing to assign higher attribution to X_6, which has a more normal baseline with $P(X_6 = 0) = 0.95$ compared to $P(X_7 = 0) = 0.05$, also violate $D_4(c')$.

For D_3 (Effect Scaling), consider the comparison between X_4 (Series A funding, \$8M) and X_5 (Angel funding, \$1M). We construct a comparison SCM $\mathcal{M}'$ by relabeling variables (swapping X_4 and X_5). Both variables are but-for causes: $Y_{X_4=0} = 2 \neq 10$ and $Y_{X_5=0} = 9 \neq 10$, so $c = c' \neq 0$ for both. Since $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$, violations transfer by Thm. A.7. Methods violating $D_3(c)$ by failing to assign higher attribution to X_4, which has a larger effect magnitude of \$8M compared to \$1M, also violate $D_3(c')$.

Therefore, all violations in Table 4 satisfy the lifting conditions and transfer from proxy to oracle desiderata. □

B.1 Simple Causal Structures for D2 Power

Beyond the startup funding example, we present additional simple causal structures where LIME and other methods violate power (D_2). These are unambiguous cases where functional dependence coincides with actual causation ($c = c'$) due to simple, deterministic mechanisms.

Chain Structure. Consider the causal chain $X \to Z \to Y$ with SCM:

$$\mathcal{M}_{\text{chain}} = \begin{cases} X & = U_X \\ Z & = X \\ Y & = Z \end{cases} \tag{33}$$

where $U_X \sim \text{Bernoulli}(0.5)$. Observe $\mathbf{v} = \{X = 1, Z = 1, Y = 1\}$.

Variable X causes Y: changing X from 1 to 0 changes Y from 1 to 0 via Z. This holds under functional dependence ($c = 1$) and under any reasonable actual causation definition ($c' = 1$), since the mechanism is deterministic and non-preempted. Yet LIME assigns $\phi_X = 0$ because X and Y are conditionally independent given Z, violating power (D_2).

This violates $D_2(c)$ directly, and since $c = c'$ for this simple deterministic chain, it also violates $D_2(c')$.

Fork Structure. Consider the common cause (fork) $X \leftarrow Z \to Y$ with SCM:

$$\mathcal{M}_{\text{fork}} = \begin{cases} Z & = U_Z \\ X & = Z \\ Y & = Z \end{cases} \tag{34}$$

Observe $\mathbf{v} = \{Z = 1, X = 1, Y = 1\}$.

Here, X does not cause Y (they are both effects of Z). Functional dependence $c(X, Y, z) = 0$ for all witnesses z, since interventions on X do not affect Y. Methods correctly assign $\phi_X = 0$ in this case, satisfying admissibility (D_1).

The fork demonstrates that our framework correctly distinguishes chains (where non-adjacent variables cause the outcome) from common causes (where they do not).

B.2 Extended Results Table and Discussion of Methods

Table 4 presents a comprehensive comparison of 16 explanation methods on the startup funding example (Ex. 2.1), evaluating each method's satisfaction of desiderata D_1 through D_4. The table shows variable attributions $\phi_{1:7}$ for all seven variables in Alice's case, along with checkmarks (✓) or violations (×) for each desideratum. Bolded values indicate expected behavior per Def. 2.4. For D_1 (admissibility), non-causes (ϕ_2, ϕ_3) should receive zero attribution. For D_2 (power), causes ($\phi_1, \phi_{4:7}$) should receive nonzero attribution. For D_4 (normality), the cause with more normal baseline should dominate ($\phi_6 > \phi_7$). For D_3 (effect scaling), the larger effect should dominate ($\phi_4 > \phi_5$). For PN, PS, PNS, Beckers and Halpern-Pearl, we maintain the continuous outcome, as for Shapley-based and ML methods, with $y = 10$.

The key result: only L_3 SVs satisfy all four desiderata simultaneously. All prior methods violate at least one desideratum, with systematic patterns emerging by method family. L_1 methods (SHAP, LIME, ASV-prox, ASV-distal) violate D_1 and/or D_2 due to confounding and lack of causal structure. L_2 methods violate D_1 by conflating population-level interventional effects with unit-specific functional dependence. Prior L_3 methods (PN, PS, PNS, CF-Shapley, Beckers, Halpern-Pearl) violate D_2 and/or D_3 due to structural limitations in how they aggregate counterfactual information.

Explanation-based methods including Beckers-CF, Beckers-Suff, and Halpern-Pearl use binary attributions $\phi_i = \mathbf{1}[X_i \text{ is part of an explanation}]$, where the explanation is chosen to be maximally powerful. Beckers' two explanation methods (counterfactual explanations and sufficient explanations) [16] both require direct sufficiency, meaning $X_1 = 1$ must be sufficient to induce its descendants to their observed values for all exogenous contexts $\mathbf{u}$. Beckers' methods assign $\phi_1 = 0$ to being well-spoken because there exist exogenous contexts under which $X_1 = 1$ is not directly sufficient to induce its descendants to their observed values, violating this requirement. Halpern-Pearl's explanation method [26] assigns all variables $\phi_{1:7} = 0$ because condition EX4 requires that explanatory variables cannot be known. Since all variables are observed in Alice's case, no variable qualifies as an explanation.

B.2.1 Shapley-Based Methods

All Shapley-based methods [32] share the same aggregation formula:

$$\phi_X = \mathbb{E}_{\pi \sim \text{Unif}(\Pi_{\mathbf{X}})} \left[f(\pi_{\leq X}) - f(\pi_{<X}) \right] \tag{35}$$

where π is a random ordering of variables, $\pi_{\leq X}$ denotes variables up to and including X, and $\pi_{<X}$ denotes variables before X. The key distinction lies in the value function f, which determines what each method measures.

Value Functions Across Pearl's Hierarchy. The Shapley-based methods in Table 4 differ in their value functions, corresponding to Pearl's three-layer causal hierarchy [1, 10]:

L_1 **SVs (SHAP)** [6] operate at Layer 1 (association), using observational conditioning:

$$f^{L_1}(\mathbf{Z}) = \mathbb{E}[Y|\mathbf{z}] \tag{36}$$

Asymmetric Shapley Values (ASV-proximal and ASV-distal) [20] modify the L_1 observational value function to respect variable ordering in the causal graph. Both methods compute marginal contributions asymmetrically based on whether a variable is an ancestor or descendant of the outcome. ASV-proximal attributes effects by conditioning on proximal ancestors (those closer to Y in the causal ordering), while ASV-distal conditions on distal ancestors (those further from Y). For the startup example, this asymmetry produces different attributions for variables at different graph distances from Y (see ϕ_1, ϕ_2 differences in Table 4). However, both variants still use observational conditioning

$\mathbb{E}[Y|\mathbf{z}]$ as their base value function, inheriting the admissibility violations (D_1) from L_1 methods: they assign non-zero attribution to non-causal variables when confounding is present.

L_2 SVs (do-SHAP) [22, 21, 23] operate at Layer 2 (intervention), using interventional distributions:

$$f^{L_2}(\mathbf{Z}) = \mathbb{E}[Y|do(\mathbf{z})] \tag{37}$$

CF-Shapley [25] operates at Layer 3 (counterfactuals), using a chosen reference unit $\mathbf{u}'$, typically a previous observation expected to be comparable, as the baseline for all counterfactual comparisons. The method requires Markovian SCMs (no hidden confounders) with additive errors (structural equations of the form $X_i = f_i(\text{pa}(X_i)) + U_i$), ensuring that observing $\mathbf{v}$ with knowledge of $P(\mathbf{V})$ uniquely identifies the true unit $\mathbf{u}$:

$$f^{\text{CF}}(\mathbf{Z}) = Y_{\overline{\mathbf{Z}}(\mathbf{u}')}(\mathbf{u}) \tag{38}$$

where $\overline{\mathbf{Z}} = \mathbf{X} \setminus \mathbf{Z}$ are the complement variables held at their baseline values from the chosen $\mathbf{u}'$. Ex. 2.1 (Alice's startup) violates additivity: the pitch equations (Eq. (2)) combine inputs and noise through logical operations such as $X_2 \oplus U_3$, so the exogenous noise does not enter additively as $X_i = f_i(\text{pa}(X_i)) + U_i$. For Ex. 2.1, we take as reference Alice before any funding, with $X_1 = 1$ and $X_{4:7} = 0$. Beyond the additivity restriction, CF-Shapley's use of a single fixed baseline (rather than averaging over $P(\mathbf{U})$) leads to baseline selection artifacts that violate normality (D_4) and power (D_2) in certain cases (see Positive Chain and Conjunctive examples below).

L_3 SVs (ours) operate at Layer 3 (counterfactuals), using unit-specific counterfactual reasoning with natural baselines $\mathbf{u}' \sim P(\mathbf{U})$. Notably, our method does not require Markovianity or additivity:

$$f^{L_3}(\mathbf{Z}) = \mathbb{E}_{\substack{\mathbf{u}' \sim P(\mathbf{U}) \\ \mathbf{u} \sim P(\mathbf{U}|\mathbf{v})}} \left[\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})\right] \tag{39}$$

The progression from L_1 to L_3 eliminates systematic violations: L_1 violates admissibility (D_1) because observational conditioning conflates correlation with causation (spurious, chains examples), L_2 violates admissibility because population-level interventions differ from unit-specific functional dependence (bow, chains examples), and CF-Shapley violates normality (D_4) because fixed baselines ignore probability weighting (conjunctive example). Only L_3 SVs with probability-weighted baselines satisfy all desiderata.

B.2.2 Other ML Explainability Methods

Local Interpretable Model-agnostic Explanations (LIME) [7] fits a local linear model around the observed instance by sampling perturbations and weighting by proximity. For Alice's startup example, LIME samples training instances from the SCM, then fits a local linear approximation $\hat{Y} \approx \sum_i \beta_i X_i$ near the observed point. The coefficients β_i serve as attributions. LIME violates power (D_2) by assigning zero to multiple causes. Being well-spoken ($\phi_1 = 0.00$) receives zero attribution because X_4 and X_5 (pitch quality and funding sources) are superior predictors of Y with lower variance, and $X_1 \perp Y|X_4, X_5$ in the local approximation. Background check ($\phi_7 = 0.00$ with zero variance across 100 runs) receives zero due to L_2-norm regularization shrinking its coefficient, leading the local linear model to assign the entire $X_6 \cdot X_7$ interaction to X_6 (see Table 4).

Gradient-based methods (Integrated Gradients [18], DeepLIFT [52], GradCAM [53]) compute attributions via gradients of the model output with respect to inputs. These methods require differentiable models and continuous inputs, making them inapplicable to the discrete binary variables in our examples. For settings with continuous features and neural network models, these methods may violate admissibility (D_1) similar to LIME when the learned model conflates correlation with causation.

B.2.3 Probabilities of Causation

Pearl's probabilities of causation [13, 54] measure causal necessity and sufficiency for individual units. For a binary variable X and an outcome Y observed at $x = 1$ and y, with baseline $x' = 0$, and evidence $\mathbf{e}_x$ about variables other than X:

$$\text{PN}(x, y|\mathbf{e}_x) = P(Y_{x'} \neq y|\mathbf{e}_x, X = x, Y = y) \quad \text{(Probability of Necessity)} \tag{40}$$

$$\text{PS}(x, y|\mathbf{e}_x) = P(Y_x = y|\mathbf{e}_x, X = x', Y = y') \quad \text{(Probability of Sufficiency)} \tag{41}$$

$$\text{PNS}(x, y|\mathbf{e}_x) = P(Y_x = y, Y_{x'} \neq y|\mathbf{e}_x, X = x, Y = y) \quad \text{(Prob. of Necessity and Sufficiency)} \tag{42}$$

PN measures whether $X = x$ was necessary for $Y = y$, PS measures whether $X = x$ would be sufficient to produce $Y = y$, and PNS requires both. In PS, $y' \neq y$ denotes the absence of the outcome, so for non-binary Y the event is $Y \neq y$, and $Y_x = y$ requires the exact value y.

The evidence $\mathbf{e}_x$ can be empty (unconditional) or include full evidence $\mathbf{e}_x = \{\mathbf{x}, y\}$ where $\mathbf{x} = (x_1, \ldots, x_n)$ are values for all variables except X.

Unconditional versions ($\text{PN}(x, y|\emptyset)$) measure population-level necessity/sufficiency, while fully conditioned versions ($\text{PN}(x, y|\mathbf{x}, y)$) measure unit-specific necessity/sufficiency. Conditioned PN assigns one to all but-for causes, whatever their effect sizes and baseline probabilities (scaling violation D_3 and normality violation D_4), while unconditioned PN also assigns non-zero to non-causes (admissibility violation D_1), as shown in Table 4.

Connection to NTE. For binary variables observed at $X = 1, Y = 1$, the NTE with univariate baseline relates to conditioned PN as:

$$\text{NTE}(X, Y|\mathbf{v}) = \mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X)}|\mathbf{v}] = 1 - (P(x) \cdot 1 + P(x') \cdot P(Y_{x'} = 1|\mathbf{v})) = P(x') \cdot \text{PN}(x, y|\mathbf{x}, y) \tag{43}$$

Thus, NTE equals scaled PN in binary settings. L_3 SVs generalize this by (1) decomposing effects across all variables via Shapley aggregation rather than measuring univariate necessity, (2) weighting baselines by $P(\mathbf{U})$ rather than treating all counterfactuals equally, and (3) capturing effect magnitudes for non-binary outcomes.

B.2.4 Actual Causation Based Explanations

Halpern-Pearl's (HP) explanation method [26] and Beckers' counterfactual and sufficient explanation methods [16] technically determine whether any subset of variables is an explanation for an outcome. To maximize the power of these methods in our comparison, we modify their output: we assign a binary judgment to each variable according to whether it is included in any (possibly minimal) explanation identified by the method. HP employs sufficient causation with the EX4 constraint (explanatory variables cannot be known), while Beckers requires direct sufficiency for inclusion (variables must be sufficient to produce their effects for all exogenous contexts).

These methods violate power (D_2) due to structural constraints. HP assigns zero to all observed variables because EX4 disqualifies known variables: for Alice's fully-observed case, $\phi_{1:7} = 0$ (Table 4). Beckers assigns zero to variables lacking universal sufficiency: being well-spoken (X_1) gets $\phi_1 = 0$ because there exist contexts where $X_1 = 1$ is insufficient to induce descendants to observed values.

Beyond power violations, binary output structure prevents satisfaction of normality (D_4) and scaling (D_3). For normality, binary judgments cannot weight by baseline probability: in a disjunctive example with $Y = X_1 \vee X_2$ where $P(X_1 = 0) = 0.99$ and $P(X_2 = 0) = 0.01$, both methods output binary 1 for both variables despite the 99-fold baseline probability difference. For scaling, binary judgments cannot represent effect magnitude: in Alice's funding example with $Y = 8X_4 + X_5$, both methods output binary 1 for the \$8M Series A (X_4) and \$1M Angel (X_5) funding despite the eightfold magnitude difference.

The fundamental limitation is binary output: these methods answer the question "is X a cause?" with a yes or no response but cannot distinguish strong from weak causes, normal from abnormal baselines, or large from small effects. Quantitative attribution methods like L_3 SVs are necessary to satisfy all desiderata.

B.3 Computation of L_3 Shapley Values

We illustrate the computation of $\phi_{X_1}^{L_3}(w)$ (being well-spoken) for Alice's case by showing how coalitions contribute to the final attribution. Recall from Def. 3.2 that $\phi_{X_1}^{L_3}(w) = \mathbb{E}_{\pi, \mathbf{u}', \mathbf{u}}[f(\pi_{\leq X_1}) - f(\pi_{<X_1})]$ where $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})$ (from Def. 3.1). The permutation-based Shapley formula is equivalent to enumerating all coalitions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X_1\}$ with Shapley weights $w(|\mathbf{Z}|) = \frac{|\mathbf{Z}|!(n-|\mathbf{Z}|-1)!}{n!}$ where $n = |\mathbf{X}|$.

Table 6 shows NTE values for all coalitions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X_1\}$ (all subsets excluding X_1) for a fixed $\mathbf{u}' \to \mathbf{u}$ pair with $\mathbf{u}' = (0, \ldots, 0)$ (all-zeros baseline) and $\mathbf{u}$ chosen consistent with Alice's observed case

Table 6: L_3 Shapley value computation for X_1. Shows all 64 coalitions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X_1\}$ and marginal contributions $f(\mathbf{Z} \cup \{X_1\}) - f(\mathbf{Z})$ for fixed $\mathbf{u}' \to \mathbf{u}$.

$\mathbf{Z}$	w	$Y_{\mathbf{Z}\cup\{X_1\}(\mathbf{u}')}$	$Y_{\mathbf{Z}(\mathbf{u}')}$	Y	NTE($\mathbf{Z} \cup \{X_1\}$)	NTE($\mathbf{Z}$)	Marginal
$\emptyset$	0.1429	1.0	10.0	10.0	9.0	0.0	9.0
$\{X_2\}$	0.0238	1.0	10.0	10.0	9.0	0.0	9.0
$\{X_3\}$	0.0238	1.0	10.0	10.0	9.0	0.0	9.0
$\{X_4\}$	0.0238	1.0	2.0	10.0	9.0	8.0	1.0
$\{X_5\}$	0.0238	1.0	9.0	10.0	9.0	1.0	8.0
$\{X_6\}$	0.0238	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_7\}$	0.0238	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_3\}$	0.0095	1.0	10.0	10.0	9.0	0.0	9.0
$\{X_2, X_4\}$	0.0095	1.0	2.0	10.0	9.0	8.0	1.0
$\{X_2, X_5\}$	0.0095	1.0	9.0	10.0	9.0	1.0	8.0
$\{X_2, X_6\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_7\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_3, X_4\}$	0.0095	1.0	2.0	10.0	9.0	8.0	1.0
$\{X_3, X_5\}$	0.0095	1.0	9.0	10.0	9.0	1.0	8.0
$\{X_3, X_6\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_3, X_7\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_4, X_5\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_4, X_6\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_4, X_7\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_5, X_6\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_5, X_7\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_6, X_7\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_3, X_4\}$	0.0071	1.0	2.0	10.0	9.0	8.0	1.0
$\{X_2, X_3, X_5\}$	0.0071	1.0	9.0	10.0	9.0	1.0	8.0
$\{X_2, X_3, X_6\}$	0.0071	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_3, X_7\}$	0.0071	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_4, X_5\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_2, X_4, X_6\}$	0.0071	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_4, X_7\}$	0.0071	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_5, X_6\}$	0.0071	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_5, X_7\}$	0.0071	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_6, X_7\}$	0.0071	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_3, X_4, X_5\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3, X_4, X_6\}$	0.0071	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_3, X_4, X_7\}$	0.0071	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_3, X_5, X_6\}$	0.0071	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_3, X_5, X_7\}$	0.0071	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_3, X_6, X_7\}$	0.0071	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_4, X_5, X_6\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_4, X_5, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_4, X_6, X_7\}$	0.0071	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_5, X_6, X_7\}$	0.0071	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_3, X_4, X_5\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_2, X_3, X_4, X_6\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_3, X_4, X_7\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_3, X_5, X_6\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_3, X_5, X_7\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_3, X_6, X_7\}$	0.0095	0.0	9.0	10.0	10.0	1.0	9.0
$\{X_2, X_4, X_5, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_4, X_5, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_4, X_6, X_7\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_5, X_6, X_7\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_3, X_4, X_5, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_6, X_7\}$	0.0095	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_3, X_5, X_6, X_7\}$	0.0095	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_4, X_5, X_6, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_3, X_4, X_5, X_6\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_3, X_4, X_5, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_3, X_4, X_6, X_7\}$	0.0238	0.0	1.0	10.0	10.0	9.0	1.0
$\{X_2, X_3, X_5, X_6, X_7\}$	0.0238	0.0	8.0	10.0	10.0	2.0	8.0
$\{X_2, X_4, X_5, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_2, X_3, X_4, X_5, X_6, X_7\}$	0.1429	0.0	0.0	10.0	10.0	10.0	0.0
						$\phi_{X_1}^{L_3}(w\|\mathbf{u}, \mathbf{u}')$	**4.50**

$\mathbf{v}(\mathbf{u}) = (\mathbf{x} = \mathbf{1}, Y = 10)$. For each coalition $\mathbf{Z}$, we compute the NTE value $f(\mathbf{Z}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})$, which represents the natural total effect of intervening on coalition $\mathbf{Z}$. The marginal contribution of X_1 for coalition $\mathbf{Z}$ equals $f(\mathbf{Z} \cup \{X_1\}) - f(\mathbf{Z})$, obtained by comparing NTE values between coalitions with and without X_1. Coalitions containing X_6 or X_7 and none of X_1, X_4, X_5 produce NTE = 1 (intervening on grant variables changes Y from 10 to 9), while coalitions within $\{X_2, X_3\}$ produce NTE = 0. The Shapley value is computed by weighting each marginal contribution by the Shapley weight $w(|\mathbf{Z}|)$ and summing over all coalitions, yielding $\phi_{X_1}^{L_3}(w|\mathbf{u}, \mathbf{u}') = 4.50$ for this fixed $\mathbf{u}' \to \mathbf{u}$ pair.

Table 6 illustrates the computation for X_1 (being well-spoken), a cause with positive attribution, while Table 7 shows the computation for X_2 (the live demo), a non-cause with zero attribution. For X_1,

Table 7: L_3 Shapley value computation for X_2. Shows all 64 coalitions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X_2\}$ and marginal contributions $f(\mathbf{Z} \cup \{X_2\}) - f(\mathbf{Z})$ for fixed $\mathbf{u}' \to \mathbf{u}$.

$\mathbf{Z}$	w	$Y_{\mathbf{Z}\cup\{X_2\}(\mathbf{u}')}$	$Y_{\mathbf{Z}(\mathbf{u}')}$	Y	NTE($\mathbf{Z} \cup \{X_2\}$)	NTE($\mathbf{Z}$)	Marginal
$\emptyset$	0.1429	10.0	10.0	10.0	0.0	0.0	0.0
$\{X_1\}$	0.0238	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3\}$	0.0238	10.0	10.0	10.0	0.0	0.0	0.0
$\{X_4\}$	0.0238	2.0	2.0	10.0	8.0	8.0	0.0
$\{X_5\}$	0.0238	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_6\}$	0.0238	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_7\}$	0.0238	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_1, X_3\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_4\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_5\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4\}$	0.0095	2.0	2.0	10.0	8.0	8.0	0.0
$\{X_3, X_5\}$	0.0095	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_3, X_6\}$	0.0095	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_3, X_7\}$	0.0095	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_4, X_5\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_4, X_6\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_4, X_7\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_5, X_6\}$	0.0095	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_5, X_7\}$	0.0095	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_6, X_7\}$	0.0095	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_1, X_3, X_4\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_3, X_5\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_3, X_6\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_5\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_4, X_6\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_5, X_6\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_5, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_6, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3, X_4, X_6\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3, X_4, X_7\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3, X_5, X_6\}$	0.0071	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_3, X_5, X_7\}$	0.0071	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_3, X_6, X_7\}$	0.0071	9.0	9.0	10.0	1.0	1.0	0.0
$\{X_4, X_5, X_6\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_4, X_5, X_7\}$	0.0071	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_4, X_6, X_7\}$	0.0071	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_5, X_6, X_7\}$	0.0071	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_1, X_3, X_4, X_5\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_1, X_3, X_4, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_4, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_5, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_5, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_6, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_5, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_5, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_6, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_5, X_6, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5, X_6\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_6, X_7\}$	0.0095	1.0	1.0	10.0	9.0	9.0	0.0
$\{X_3, X_5, X_6, X_7\}$	0.0095	8.0	8.0	10.0	2.0	2.0	0.0
$\{X_4, X_5, X_6, X_7\}$	0.0095	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_4, X_5, X_6\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_4, X_5, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_4, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_5, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_4, X_5, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_3, X_4, X_5, X_6, X_7\}$	0.0238	0.0	0.0	10.0	10.0	10.0	0.0
$\{X_1, X_3, X_4, X_5, X_6, X_7\}$	0.1429	0.0	0.0	10.0	10.0	10.0	0.0
						$\phi_{X_2}^{L_3}(w\|\mathbf{u}, \mathbf{u}')$	**0.00**

coalitions contribute varying marginal values, yielding $\phi_{X_1}^{L_3}(w|\mathbf{u}, \mathbf{u}') = 4.50$. For X_2, all marginal contributions equal 0. Observing $X_4 = X_5 = 1$ with $(X_2 \oplus U_3) = 0$ implies $U_4^2 = U_5^2 = 1$, since these terms compensated for the ineffective live demo. The pitch mechanism reduces to $X_j = X_1 \wedge 1 \wedge U_j^3$, making the outcome independent of X_2. This correctly yields $\phi_{X_2}^{L_3}(w|\mathbf{u}, \mathbf{u}') = 0.00$ per D_1.

To obtain the final attribution $\phi_{X_1}^{L_3}(w)$, we average over all possible $(\mathbf{u}, \mathbf{u}')$ pairs as specified in Def. 3.2: $\phi_{X_1}^{L_3}(w) = \mathbb{E}_{\mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U}|\mathbf{v})}[\phi_{X_1}^{L_3}(w|\mathbf{u}, \mathbf{u}')]$. The baseline distribution $P(\mathbf{U})$ weights each $\mathbf{u}'$ by its probability, and the posterior $P(\mathbf{U}|\mathbf{v})$ weights each $\mathbf{u}$ by its compatibility with the observed evidence $\mathbf{v} = (\mathbf{x} = \mathbf{1}, Y = 10)$. For Alice's startup example with 9 exogenous variables, there are

$2^9 = 512$ possible $\mathbf{u}$ settings, of which 1 is consistent with the observed $\mathbf{v}$. The final attribution $\phi_{X_1}^{L_3}(w) = 4.28$ (Table 4) results from this double expectation.

This generalizes CF-Shapley [25]: when $\mathbf{v}$ uniquely determines $\mathbf{u}$, $P(\mathbf{U}|\mathbf{v})$ concentrates on a single $\mathbf{u}$, recovering CF-Shapley's computation at a fixed $\mathbf{u}'$. Our approach handles cases where multiple $\mathbf{u}$ settings are consistent with observed $\mathbf{v}$, weighting each by its posterior probability.

B.4 Examples

B.4.1 Spurious

We demonstrate that SHAP (L_1 SVs) violates D_1 (admissibility) by attributing effects to non-causal variables when confounding is present.

Example B.2 (Spurious)**.** Consider two variables X_1 and X_2 confounded by a common cause U, where we explain outcome $Y = X_2$ using feature X_1. The simplified SCM is:

$$\mathcal{F} := \begin{cases} X_1 & := U \\ X_2 & := U \\ Y & := X_2 \end{cases}, \quad U \sim \text{Bern}(0.5) \tag{44}$$

The causal diagram has no edge $X_1 \to Y$, only the backdoor path $X_1 \leftarrow U \to X_2 = Y$. Observe the unit: $X_1 = 1, Y = 1$. □

For all baselines z, we have $c(X_1, Y, z) = 0$ because there is no causal path from X_1 to Y in the structural equations. Thus, D_1 requires $\phi_{X_1} = 0$.

L_1 SVs use observational conditioning $\mathbb{E}[Y|X_1 = x]$, which conflates correlation with causation. Since $Y = X_2 = U$, we have $P(Y = 1|X_1 = 1) = 1$ and $P(Y = 1) = 0.5$, yielding strong observational correlation: $\mathbb{E}[Y|X_1 = 1] - \mathbb{E}[Y] = 0.5$. The Shapley value $\phi_{X_1}^{L_1} = 0.5$ assigns non-zero attribution to X_1 despite it not causing Y, violating D_1.

In contrast, L_2 and L_3 SVs both correctly assign $\phi_{X_1}^{L_2} = \phi_{X_1}^{L_3} = 0$ because interventions and counterfactuals respect the causal structure (no $X_1 \to Y$ edge). For L_2: $\mathbb{E}[Y|do(X_1 = 1)] = \mathbb{E}[Y|do(X_1 = 0)] = 0.5$ (intervention breaks the backdoor path). For L_3: $\mathbb{E}[Y|X_1 = 1, Y = 1] - \mathbb{E}[Y_{P(X_1)}|X_1 = 1, Y = 1] = 1 - 1 = 0$ (counterfactually changing X_1 does not affect Y in this unit).

B.4.2 Negative Chain

We demonstrate that L_1 and L_2 SVs violate D_1 by attributing effects to variables that do not functionally determine the outcome in the observed unit.

Example B.3 (Negative Chain)**.** Consider a simplified scenario inspired by Ex. 2.1: being well-spoken (X_1), successful pitches (X_2), and whether Alice got funded (Y). The simplified SCM is:

$$\mathcal{F} := \begin{cases} X_1 & := U_1 \\ X_2 & := X_1 \vee U_2 \\ Y & := X_2 \end{cases}, \quad U_1, U_2 \sim \text{Bern}(0.5) \tag{45}$$

Pitch success depends on being well-spoken or having natural charisma (U_2), and funding depends on pitch success. We observe Alice's unit: $X_1 = 1$ (well-spoken), $X_2 = 1$ (successful pitches), $Y = 1$ (got funded), with $U_2 = 1$ (success came from natural charisma, not being well-spoken). □

For this unit, X_1 is functionally independent of Y: changing X_1 from 1 to 0 would not change Y (since $Y = X_2 = 0 \vee 1 = 1$ when $U_2 = 1$). Thus, $c(X_1, Y, z) = 0$ and D_1 requires $\phi_{X_1} = 0$.

Both L_1 and L_2 methods fail because they aggregate over populations or interventions rather than evaluating unit-specific functional dependence.

L_1 SV computation. L_1 SVs use observational conditioning. Since $Y = X_2 = X_1 \vee U_2$, we have:

$$\mathbb{E}[Y|X_1 = 1] = P(1 \vee U_2 = 1) = 1 \tag{46}$$

$$\mathbb{E}[Y|X_1 = 0] = P(0 \vee U_2 = 1) = P(U_2 = 1) = 0.5 \tag{47}$$

L_1 observes population-level correlation and assigns $\phi_{X_1}^{L_1} = 1 - 0.5 = 0.5 \neq 0$, violating D_1. L_1 arrives at non-zero because it measures population-level correlation rather than recognizing unit-specific functional independence.

L_2 SV computation. L_2 SVs use interventional expectations. Since $Y = X_2 = X_1 \vee U_2$ (there exists a causal path $X_1 \to X_2 \to Y$), interventions on X_1 have population-average effects:

$$\mathbb{E}[Y|do(X_1 = 1)] = P(1 \vee U_2 = 1) = 1 \tag{48}$$

$$\mathbb{E}[Y|do(X_1 = 0)] = P(0 \vee U_2 = 1) = 0.5 \tag{49}$$

L_2 assigns $\phi_{X_1}^{L_2} = 1 - 0.5 = 0.5 \neq 0$, violating D_1. L_2 gets non-zero because there is a population-average causal effect, not recognizing that X_1 is functionally independent in this specific unit.

In contrast, L_3 SVs correctly assign $\phi_{X_1}^{L_3} = 0$ for the right reason: counterfactual reasoning conditioned on the observed unit ($X_1 = 1, X_2 = 1, Y = 1, U_2 = 1$) reveals that changing X_1 would not have changed Y in this unit.

B.4.3 Positive Chain

We demonstrate that LIME and CF-Shapley violate D_2 (power) by failing to detect causation in chain structures where the mediating variable is a stronger predictor than the root cause.

Example B.4 (Positive Chain)**.** Consider a simplified scenario inspired by Ex. 2.1: being well-spoken (X_1) and successful pitches (X_2). The simplified SCM is:

$$\mathcal{F} := \begin{cases} X_1 & := U_1 \\ X_2 & := X_1 \vee U_2 \ , \\ Y & := X_2 \end{cases} \quad U_1, U_2 \sim \text{Bern}(0.5) \tag{50}$$

Pitch success depends on being well-spoken or having natural charisma (U_2). We observe Alice's unit: $X_1 = 1$ (well-spoken), $X_2 = 1$ (successful pitches), $Y = 1$ (funded), with $U_2 = 0$ (success came from being well-spoken). □

In this unit, X_1 functionally determines Y through X_2: changing X_1 from 1 to 0 would change X_2 to 0 (since $U_2 = 0$) and thus Y to 0. Thus, D_2 requires $\phi_{X_1} > 0$ (functional dependence exists).

LIME uses local linear approximations by perturbing variables around the observed point and fitting a linear model. When LIME fits its weighted linear surrogate model over jointly-perturbed inputs, X_2 is a perfect predictor of Y ($Y = X_2$) and absorbs the explained variance, so X_1 receives zero weight. The fitted linear model assigns near-zero weight to X_1 and full weight to X_2, yielding $\phi_{X_1}^{\text{LIME}} \approx 0$. LIME fails D_2 because local linear approximations cannot capture indirect effects when mediating variables are stronger predictors than root causes.

CF-Shapley (using a fixed baseline $\mathbf{u}'$ instead of probability-weighted baselines) can also fail if the baseline is chosen incorrectly. For the observed unit $\mathbf{u} = (U_1 = 1, U_2 = 0)$ with $X_1 = 1, X_2 = 1, Y = 1$:

- If baseline $\mathbf{u}' = (U_1 = 1, U_2 = 0)$ (same as actual unit): $X_2(\mathbf{u}') = 1 \vee 0 = 1$, so $Y(\mathbf{u}) - Y(\mathbf{u}') = 1 - 1 = 0 \Rightarrow \phi_{X_1}^{\text{CF}} = 0$
- If baseline $\mathbf{u}' = (U_1 = 0, U_2 = 0)$ (different U_1): $X_2(\mathbf{u}') = 0 \vee 0 = 0$, so $Y(\mathbf{u}) - Y(\mathbf{u}') = 1 - 0 = 1 \Rightarrow \phi_{X_1}^{\text{CF}} > 0$

When the fixed baseline accidentally matches the actual unit's exogenous values, CF-Shapley fails D_2 by masking true functional dependence.

In contrast, L_2 and L_3 SVs correctly assign positive attribution: $\phi_{X_1}^{L_2} > 0$ and $\phi_{X_1}^{L_3} > 0$ because interventions and probability-weighted counterfactual baselines respect the causal chain structure and avoid baseline selection artifacts.

B.4.4 Disjunctive

We demonstrate that PN (Probability of Necessity) and LIME violate D_2 (power) by failing to detect causation when multiple sufficient causes overdetermine the outcome, using Alice's startup funding sources.

Example B.5 (Funding overdetermination)**.** From Alice's startup funding (Ex. 2.1), consider whether Alice got funded (binary outcome):

$$\mathcal{F} := \begin{cases} X_4 & := U_4 \quad \text{(Series A funding, \$8M if 1)} \\ X_5 & := U_5 \quad \text{(Angel funding, \$1M if 1)} \\ Y & := \mathbf{1}[X_4 = 1 \vee X_5 = 1] \quad \text{(got funded)} \end{cases} \tag{51}$$

with $U_4, U_5 \sim \text{Bern}(0.95)$ (simplified from full startup SCM). Observe Alice's case: $X_4 = 1$ (Series A secured), $X_5 = 1$ (Angel secured), $Y = 1$ (funded). □

Both X_4 and X_5 are causes of Y. With multivariate baseline $z = (x_4' = 0, z' = \{X_5 = 0\})$, we have $c(X_4, Y, z) = 1$ (changing X_4 from 0 to 1 changes Y from 0 to 1 when $X_5 = 0$). Similarly for X_5. Thus, D_2 requires $\phi_{X_4} \neq 0$ and $\phi_{X_5} \neq 0$.

However, the conditioned $\text{PN}(X_4, Y \mid \mathbf{x}, y)$, which fixes the full observed unit (including $X_5 = 1$), assigns:

$$\begin{aligned} \text{PN}(X_4, Y \mid \mathbf{x}, y) &= P(Y_{X_4=0} \neq y | X_4 = 1, Y = 1) && (52) \\ &= P(Y_{X_4=0} = 0 | X_4 = 1, Y = 1) && (53) \\ &= P(X_5 = 0 | X_4 = 1, Y = 1) = 0 \quad \text{(since } X_5 = 1 \text{ in observed case)} && (54) \end{aligned}$$

Thus $\text{PN}(X_4, Y \mid \mathbf{x}, y) = 0$ despite X_4 being a cause of Y. PN uses univariate counterfactuals that only change X_4 while keeping X_5 at its observed value. When $X_5 = 1$, setting $X_4 = 0$ does not change Y (because Alice still gets funded via Angel round: $0 \vee 1 = 1$). PN fails to detect that X_4 would have been necessary if X_5 had been different. This is overdetermination: both funding sources are sufficient, so removing either one alone does not prevent the binary outcome "got funded."

In contrast, L_3 SVs correctly assign $\phi_{X_4}^{L_3} \neq 0$ and $\phi_{X_5}^{L_3} \neq 0$ because the Shapley aggregation considers all baselines, including those where $X_5 = 0$ (making X_4 necessary) and where $X_4 = 0$ (making X_5 necessary). When averaging over baselines $P(X_4, X_5)$, both variables receive appropriate attribution, as shown in Table 4.

B.4.5 Conjunctive

We demonstrate that CF-Shapley and PN violate D_4 (normality) by failing to weight attributions by baseline probability, using Alice's conjunctive grant requirement.

Example B.6 (AI grant with legal requirement)**.** From Alice's startup funding (Ex. 2.1), consider the AI grant that requires both winning the competition and passing the background check:

$$\mathcal{F} := \begin{cases} X_6 & := U_6 \quad \text{(won AI grant competition)} \\ X_7 & := U_7 \quad \text{(background check)} \\ Y & := X_6 \wedge X_7 \quad \text{(\$1M grant received)} \end{cases} \tag{55}$$

with $P(U_6 = 1) = 0.05$ and $P(U_7 = 1) = 0.95$ (from full startup SCM). Observe Alice's case: $X_6 = 1$ (won competition), $X_7 = 1$ (passed check), $Y = 1$ (grant awarded). □

Both X_6 and X_7 are causes of Y with identical counterfactual structure: $c(X_6, Y) = c(X_7, Y) = 1$ (changing either from 1 to 0 changes Y from 1 to 0). The structural equations are symmetric: both variables are equally necessary for the grant.

However, X_6 has a more normal baseline than X_7 because $P(X_6 = 0) = 0.95$ (95% of applicants do not win) while $P(X_7 = 0) = 0.05$ (5% fail the background check). The baseline $X_6 = 0$ is 19 times more likely than $X_7 = 0$, so we expect $\phi_{X_6} > \phi_{X_7}$ per D_4 (normality).

CF-Shapley uses a single fixed baseline unit $\mathbf{u}'$ for all variables. For the all-zeros baseline ($X_6 = 0, X_7 = 0$) and actual unit ($X_6 = 1, X_7 = 1$), the symmetry of $Y = X_6 X_7$ in this configuration gives $\phi_{X_6}^{\text{CF}} = \phi_{X_7}^{\text{CF}} = 0.5$ by direct Shapley computation: each variable contributes $\frac{1}{2}$ on average across coalitions. The 0.5/0.5 attribution holds for any baseline that is symmetric under $X_6 \leftrightarrow X_7$ (e.g., all-zeros, all-ones); for asymmetric baselines, CF-Shapley still violates D_4 but the precise numerical attribution depends on baseline choice. This violates D_4: the cause with more normal baseline (X_6, $P(X_6 = 0) = 0.95$) should receive higher attribution than the one with less normal baseline (X_7, $P(X_7 = 0) = 0.05$), but CF-Shapley's single fixed baseline cannot encode this 19-fold probability ratio. In contrast, L_3 SVs use probability-weighted baselines $P(\mathbf{U})$. In the full startup example (Table 4) where $Y = 8X_4 + X_5 + X_6 X_7$, L_3 SVs assign $\phi_6 = 0.93$ and $\phi_7 = 0.03$, with $\phi_6 > \phi_7$ as D_4 requires.

B.4.6 Scaled

We demonstrate that PN violates D_3 (scaling) by outputting binary judgments that cannot distinguish proportional differences in effect magnitudes, using Alice's two funding sources with different amounts.

Example B.7 (Series A vs Angel funding)**.** From Alice's startup funding (Ex. 2.1), consider the linear contributions from Series A and Angel funding:

$$\mathcal{F} := \begin{cases} X_4 & := U_4 \quad \text{(Series A funding)} \\ X_5 & := U_5 \quad \text{(Angel funding)} \\ Y & := 8X_4 + X_5 \quad \text{(total funding, millions)} \end{cases} \tag{56}$$

with $U_4, U_5 \sim \text{Bern}(0.95)$ (simplified from full startup SCM). Observe Alice's case: $X_4 = 1$ (\$8M), $X_5 = 1$ (\$1M), $Y = 9$ (total \$9M). □

Both X_4 and X_5 contribute linearly to Y, but with different magnitudes: $c(X_4, Y, z) = 8$ vs $c(X_5, Y, z) = 1$. Thus, D_3 requires $\phi_{X_4} > \phi_{X_5}$ (larger effects receive higher attribution).

For deterministic linear contributions, PN measures necessity:

$$PN(X_4, Y = 9) = P(Y_{X_4=0} \neq 9 | X_4 = 1, Y = 9) = P(1 \neq 9) = 1 \tag{57}$$

$$PN(X_5, Y = 9) = P(Y_{X_5=0} \neq 9 | X_5 = 1, Y = 9) = P(8 \neq 9) = 1 \tag{58}$$

PN outputs binary value 1 for both funding sources despite the eightfold difference in effect magnitude: $PN(X_4) = PN(X_5) = 1$. PN is a binary sufficiency/necessity judgment that only indicates whether X_i is a cause, not how strong the causal effect is. Binary causation methods cannot represent proportional differences in causal strength.

In contrast, in the full startup example (Table 4), L_3 SVs assign $\phi_4 = 4.15$ and $\phi_5 = 0.52$, with $\phi_4/\phi_5 = 8$ (the coefficient ratio). L_3 correctly scales attributions with effect magnitude, assigning much higher attribution to the \$8M Series A round than to the \$1M Angel round, as required by D_3.

C Computational Methods and Extended Experiments

Code is available at https://github.com/CausalAILab/improving-causal-explanations. The startup, Color MNIST and CelebA-HQ bounds are linear programs that we solve on CPU. The Color MNIST classifier training and the CelebA-HQ oracle values ran on one NVIDIA H100 80GB GPU; the Color MNIST pipeline takes under 1 minute and the CelebA-HQ oracle computation under 1 hour.

C.1 Computational Methodology

This subsection describes how we compute the bounds of Alg. 1 exactly in all three experiments. The method applies whenever every variable has a finite domain and the law of every c-component but one is identified from $\mathcal{T}$. Sec. C.3.5 states the Color MNIST program in a reduced form over counterfactual responses, and Sec. C.4.6 states the CelebA-HQ program, whose image variable ranges over a finite image bank.

Response-function parameterization. Following Balke and Pearl [34], we write each endogenous variable V as $V = r_V(\mathbf{pa}(V), T_V)$, where the type T_V indexes the $|\text{dom}(V)|^{|\text{dom}(\mathbf{pa}(V))|}$ deterministic functions from parent values to $\text{dom}(V)$. The types of one c-component share a joint law, the types of different c-components are independent, and $\theta \in \Theta$ collects the type laws of all c-components. Any exogenous law $P(\mathbf{U})$ over $\mathcal{G}$ induces a type law by mapping each $\mathbf{u}$ to the functions it selects; the SCM $\mathcal{M}_\theta$ with that type law produces the same counterfactual outcomes as the original SCM for every unit and every intervention, so every $\mathcal{M}^* \in \mathcal{S}(\mathcal{T}, \mathcal{G})$ has the same L_3 distribution as some $\mathcal{M}_\theta$. Θ is a product of simplices, hence compact, and the constraint and objective maps of the next paragraph are linear in θ, hence continuous, so this parameterization satisfies Assump. 4.1.

Linear constraints and objective. Let q denote the law of the one c-component that $\mathcal{T}$ does not identify; the laws of the other c-components are fixed by $\mathcal{T}$. Every distribution $P^{\mathcal{M}_\theta}(\mathbf{v_s})$ is a sum over

joint types of a product of the c-component laws, and all factors but q are constants, so the constraint $\mathcal{M}_\theta \in \mathcal{S}(\mathcal{T}, \mathcal{G})$ is a system of equations linear in q. The evidence event e of the query w is an event over observed variables, so $P(e)$ is identified from $\mathcal{T}$. Each coalition value in Def. 3.2 is a conditional expectation of a counterfactual outcome given e, which we write as $N_{\mathbf{Z}}(q)/P(e)$, where the numerator $N_{\mathbf{Z}}(q)$ sums over types consistent with e the counterfactual outcomes $Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})$ weighted by q and by the fixed laws, and is linear in q. $\phi_X^{L_3}(\mathcal{M}_\theta, w)$ is a fixed signed combination of these coalition values with the Shapley weights of Def. 3.2, so it is linear in q as well. Each endpoint of Alg. 1 is therefore a linear program over the simplex of q.

Solver and verification. We solve the startup and CelebA-HQ programs with the HiGHS solver through `scipy.optimize.linprog`; the Color MNIST program separates across rows and has the closed form of Sec. C.3.5. For every program we record the solver status and the maximal absolute residual of the equality constraints at the returned optimum. Every startup and CelebA-HQ program returned the status optimal; the largest residual is 8.9×10^{-16} over the startup programs and 0 over the CelebA-HQ programs. The optimum is attained by a feasible type law q^*, and $\mathcal{M}_{q^*} \in \mathcal{S}(\mathcal{T}, \mathcal{G})$, so each endpoint is the value of a model consistent with the data and the bounds are sharp. The true SCM's own type law satisfies the constraints and evaluates to the oracle value, so each oracle value lies inside its interval.

C.2 Startup Experiment

Table 2 reports, for each X_i in Ex. 2.1, the oracle L_3 SV computed from the true SCM and the exact interval from the linear program of Sec. C.1, with $\mathcal{T}$ the population law $P(X_{1:7})$ of the true SCM. The mechanism of Y is known, so $P(e)$ follows from $P(X_{1:7})$. The roots X_1, X_2, X_6, X_7 are their own c-components with laws fixed by $\mathcal{T}$, and X_3 shares the exogenous variable U_3 with the mechanisms of X_4 and X_5, so $\{X_3, X_4, X_5\}$ is the one c-component whose law is free. A type of this c-component is a triple (t_3, t_4, t_5): t_3 is the value of X_3, and t_4 and t_5 map each value of (X_1, X_2) to a value of X_4 and X_5. There are $2 \cdot 16 \cdot 16 = 512$ types, and the constraints are one normalization equation and one matching equation $P^{\mathcal{M}_\theta}(x_3, x_4, x_5 \mid x_1, x_2) = P(x_3, x_4, x_5 \mid x_1, x_2)$ per value of $(x_1, x_2, x_3, x_4, x_5)$, which is $1 + 4 \cdot 8 = 33$ equations in total.

X_6 and X_7 are point identified: they are root variables whose laws $\mathcal{T}$ fixes, the mechanism of Y is known, and the term of Y they enter, X_6X_7, does not depend on the free c-component, so the marginal contribution of X_6 or X_7 to every coalition is the same for every feasible q. X_3 has no directed path to Y, so adding X_3 to any coalition leaves every counterfactual outcome unchanged and its interval is $[0, 0]$ for every feasible q. The intervals of X_1, X_2, X_4 and X_5 depend on the joint law of (t_3, t_4, t_5). $\mathcal{T}$ constrains that law only through $P(x_3, x_4, x_5 \mid x_1, x_2)$, which $L_{1:2}$ data do not identify (Thm. 2.5). Such an interval can contain zero when the oracle value is nonzero, and excludes zero only when every feasible q agrees on the sign.

C.3 Color MNIST Experiment

This subsection specifies the structural causal model, classifiers, queries, attribution computations, identification bounds and uncertainty quantification and reports the results behind Fig. 4. All numbers come from one frozen run with seed 42.

C.3.1 Structural Causal Model

The model has five observed variables: the source $Z \in \{0, 1\}$, the hue X, the true digit Y, the image I and the prediction $\hat{Y}$. Hue and digit take values in $\{0, \dots, 9\}$ with the digit law $p = (0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.10, 0.14, 0.20, 0.28)$, so digit d has probability p_d and larger digits are more common. With probability $\alpha = 0.9$ an image comes from source 1, where the hue determines the digit; otherwise it comes from source 0, where the digit is drawn independently of the hue.

Exogenous variables. The source indicator U_Z is Bernoulli with success probability α. The hue noise U_X and the digit noise U_Y are independent draws from p. The image noise $U_I = (U_R, U_G, K_0, \dots, K_9)$ consists of a resampling coin U_R that is Bernoulli with success probability $\epsilon = 0.2$, a resampled label U_G drawn from p, and for each digit j a glyph index K_j drawn uniformly from the MNIST images of label j. All exogenous variables are mutually independent.

Structural equations.

$$Z = U_Z \tag{59}$$
$$X = U_X \tag{60}$$
$$Y = \begin{cases} X & U_Z = 1 \\ U_Y & U_Z = 0 \end{cases} \tag{61}$$
$$G = \begin{cases} Y & U_R = 0 \\ U_G & U_R = 1 \end{cases} \tag{62}$$
$$I = \text{render}(\text{glyph}_G[K_G], X) \tag{63}$$
$$\hat{Y} = \hat{f}(I) \tag{64}$$

The glyph label G equals the digit with probability $1 - \epsilon = 0.8$ and an independent draw from p otherwise, and $\text{glyph}_G[K_G]$ is the grayscale MNIST image selected by index K_G among the images of label G. The source Z records U_Z but does not enter the digit equation, so Z and Y share the exogenous variable U_Z and the graph carries a bidirected edge $Z \leftrightarrow Y$. Intervening on Z changes only the recording.

Rendering. The renderer maps a grayscale glyph to an RGB image with HSV hue $X/10$, saturation one and value equal to the glyph intensity. Saturation one means that every hue is a pure color: the RGB coefficient vector of a hue has one channel equal to one, one channel equal to zero and one intermediate channel, and each pixel of I is the glyph intensity multiplied by that coefficient vector. The pointwise maximum over the three channels therefore recovers the grayscale glyph exactly. Pixel intensities are rescaled to $[-1, 1]$ before entering a classifier.

Graph and observational law. The observed graph is $Z \leftrightarrow Y$, $X \to Y$, $X \to I$, $Y \to I$ and $I \to \hat{Y}$. The observational law of the recorded variables is

$$P(Z = 0, X = h, Y = j) = (1 - \alpha)\, p_h\, p_j, \qquad P(Z = 1, X = h, Y = j) = \alpha\, p_h\, \mathbb{1}[j = h], \tag{65}$$

with the marginal $P(X = h, Y = j) = (1 - \alpha)p_h p_j + \alpha p_h \mathbb{1}[h = j]$. In source 0 the digit is independent of the hue with law $P(Y = j \mid Z = 0) = p_j$.

C.3.2 Classifiers

Both classifiers are LeNet-5 networks: two convolutional layers with 6 and 16 filters of size 5×5, each followed by a hyperbolic tangent and 2×2 max pooling, then fully connected layers of width 120, 84 and 10 with hyperbolic tangent activations. The standard classifier $\hat{f}_{\text{std}}$ receives the three RGB channels. The robust classifier $\hat{f}_{\text{rob}}$ receives one channel, the pointwise maximum over the RGB channels, which equals the grayscale glyph, so its prediction cannot depend on the hue. Both train with Adam at learning rate 10^{-3}, batch size 512 and cross-entropy loss for 20 epochs on 60,000 images sampled from the model with glyphs from the MNIST training split, under one training seed, 42. Held-out accuracy is measured on 10,000 independently sampled images with glyphs from the MNIST test split: 0.905 with 95% Wilson interval $[0.899, 0.911]$ for the standard classifier and 0.808 with 95% Wilson interval $[0.800, 0.815]$ for the robust classifier. We trained the classifiers once and reused them for every query.

C.3.3 Queries and Factual Cohort

Every attribution uses the 10,000 images of the MNIST test split as the glyph bank of the model. A query fixes a digit d and a factual image i with evidence $Z = 0$, $X = d$, $Y = d$ and $I = i$, and explains the event $\hat{Y} = d$. The factual source is 0, so the unit's digit is hue independent and the hue is a non-cause of the digit for that unit. A test image is eligible for digit d when its label is d, its pixel array appears under no other label in the test split, and both classifiers predict d when the image is rendered at hue d. The uniqueness condition makes the glyph label $G = d$ a deterministic function of the evidence, and the prediction condition makes the factual event hold for both classifiers. The eligible counts for digits 0 through 9 are 956, 1113, 968, 949, 949, 857, 947, 1006, 968 and 999. For each digit we draw 100 factual images uniformly with replacement from the eligible set, 1,000 queries in total. The ten illustration images in Fig. 4 are the first eligible test-split index for each digit, fixed before any attribution was inspected and shared by both classifiers; the indices are 3, 2, 1, 30, 4, 8, 11, 0, 61 and 7.

C.3.4 Attribution Computation

All values are exact expectations over the finite glyph bank; no Monte Carlo sampling enters a point value. For a classifier $\hat{f}$, let $B[h, g, k]$ be the fraction of test-split glyphs of label g that $\hat{f}$ predicts as k when rendered at hue h. For the query digit d, the population response $R[h, j] = P(\hat{Y} = d \mid X = h, Y = j) = (1 - \epsilon)B[h, j, d] + \epsilon \sum_g p_g B[h, g, d]$ averages over the glyph label and glyph index and does not depend on Z because rendering uses only the hue and the glyph.

L_1 and L_2 games. The L_1 game conditions the event on the coalition: $v_1(S) = P(\hat{Y} = d \mid X_S = d)$, computed from $P(X, Y)$ and R, so $v_1(\varnothing) = \sum_{h,j} P(h, j)R[h, j]$, $v_1(\{X\}) = \sum_j P(Y = j \mid X = d)R[d, j]$, $v_1(\{Y\}) = \sum_h P(X = h \mid Y = d)R[h, d]$ and $v_1(\{X, Y\}) = R[d, d]$. The L_2 game intervenes on the coalition in the population: $v_2(S) = P(\hat{Y} = d \mid \mathrm{do}(X_S = d))$. Setting $X = d$ sets $Y = d$ in source 1 and leaves Y with law p in source 0, so $v_2(\{X\}) = (1 - \alpha) \sum_j p_j R[d, j] + \alpha R[d, d]$; setting $Y = d$ leaves X with law p, so $v_2(\{Y\}) = \sum_h p_h R[h, d]$; and $v_2(\varnothing) = v_1(\varnothing)$, $v_2(\{X, Y\}) = R[d, d]$. Both games integrate over the source and take the same value for every mechanism of Z with the same $P(X, Y)$.

Image posterior. The L_3 SV for image i needs the counterfactual response $Q[h, j] = P(\hat{Y}_{X=h,Y=j} = d \mid e)$ under the image evidence e. Given $Y = d$ and a glyph of label d, the posterior probability that the glyph followed the label is $w_d = (1 - \epsilon)/((1 - \epsilon) + \epsilon p_d)$. In that branch, intervening to $Y = j \neq d$ draws a fresh uniform glyph of label j, so the response is $B[h, j, d]$; in the resampling branch the observed glyph stays in place for every digit intervention. With $o[h] = \mathbb{1}[\hat{f}(\mathrm{render}(i, h)) = d]$ denoting whether the factual glyph is still predicted as d at hue h,

$$Q[h, d] = o[h], \qquad Q[h, j] = w_d\, B[h, j, d] + (1 - w_d)\, o[h] \quad \text{for } j \neq d. \tag{66}$$

Eligibility gives $Q[d, d] = o[d] = 1$.

L_3 game. The value function of Def. 3.2 is the natural total effect $v_3(S) = \mathrm{NTE}(S, \hat{Y} \mid e)$: the factual probability of the event minus its probability after replacing the coalition by the values it takes in an independent baseline unit $u' \sim P(U)$. The baseline hue has law p, the baseline digit has law p, and the baseline pair has law $P(X, Y)$. Because $Z = 0$ reveals that the digit was drawn independently of the hue, replacing the hue alone leaves the unit's digit at d. Therefore

$$v_3(\varnothing) = 0, \qquad v_3(\{X\}) = Q[d, d] - \sum_h p_h Q[h, d], \tag{67}$$

$$v_3(\{Y\}) = Q[d, d] - \sum_j p_j Q[d, j], \qquad v_3(\{X, Y\}) = Q[d, d] - \sum_{h,j} P(X = h, Y = j)\, Q[h, j]. \tag{68}$$

With two players the Shapley values are $\phi_X = \frac{1}{2}[v(\{X\}) - v(\varnothing) + v(\{X, Y\}) - v(\{Y\})]$ and $\phi_Y = \frac{1}{2}[v(\{Y\}) - v(\varnothing) + v(\{X, Y\}) - v(\{X\})]$ for each game. For the robust classifier $Q[h, j]$ does not depend on h, so $v_3(\{X\}) = 0$ and $v_3(\{X, Y\}) = v_3(\{Y\})$, and $\phi_X^{L_3} = 0$ exactly; the code returns this zero structurally rather than as a difference of rounded sums. For the standard classifier $\phi_X^{L_3} > 0$ in all 1,000 queries.

C.3.5 Identification Bounds

The L_3 bounds of Alg. 1 use the graph, the observational law $P(Z, X, Y)$, the known renderer and the known classifier, and range over the unknown digit mechanism. For a query with digit d, the unknown is the unit's response to hue interventions in source 0, summarized by $q[h, j] = P(Y_{X=h} = j \mid Y_{X=d} = d, Z = 0)$. Only $v_3(\{X\})$ depends on q: replacing the hue by h sends the digit to $Y_{X=h}$, so

$$\phi_X^{L_3}(q) = \tfrac{1}{2}\Big[Q[d, d] - \sum_h p_h \sum_j q[h, j]\, Q[h, j] + \sum_j p_j Q[d, j] - \sum_{h,j} P(X = h, Y = j)\, Q[h, j]\Big], \tag{69}$$

and $\phi_Y^{L_3}(q) = v_3(\{X, Y\}) - \phi_X^{L_3}(q)$ by efficiency. The image posterior Q is known because the renderer and classifier are known and the image noise is independent of the digit mechanism.

Constraints. Let $s = p$ be the digit law in source 0. The feasible set is

$$q[d, \cdot] = \delta_d, \qquad q[h, j] \geq 0, \qquad \sum_j q[h, j] = 1, \qquad q[h, j] \leq s_j / s_d \quad \text{for all } h \neq d,\ j. \tag{70}$$

The first constraint is consistency with the factual digit. The cap follows from $P(Y_{X=h} = j, Y_{X=d} = d \mid Z = 0) \le P(Y_{X=h} = j \mid Z = 0) = s_j$ after dividing by $P(Y_{X=d} = d \mid Z = 0) = s_d$. These constraints are necessary and jointly sufficient for q to be the hue response of some model compatible with the data. For sufficiency, place mass s_d on the event $Y_{X=d} = d$ with the rows of q as conditional responses; each remaining marginal mass $s_j - s_d q[h, j]$ is nonnegative by the cap and sums to $1 - s_d$ over j, so drawing the responses conditionally independently from these remainders on the complementary event yields a joint law of $(Y_{X=0}, \ldots, Y_{X=9})$ with every marginal equal to s. The objective is linear in q and separates across rows, so each extreme is a fractional knapsack: fill row h greedily in decreasing order of $Q[h, j]$ for the largest $\sum_j q[h, j]Q[h, j]$ and in increasing order for the smallest. Because every feasible q is realized by a compatible model and the objective is linear on a convex set, the interval of $\phi_X^{L_3}$ values is exactly the set of values attained by compatible models: the bounds are sharp.

Identification of the inputs. Every interventional law that the bound uses equals an observational conditional on this graph. The hue X is a root with no bidirected edge, so for every hue h, $P(Y, Z \mid \mathrm{do}(X = h)) = P(Y, Z \mid X = h)$ and in particular $P(Y_{X=h} = j \mid Z = 0) = P(Y = j \mid X = h, Z = 0) = s_j$. The image I has no bidirected edge and its parents are X and Y, so $P(I \mid \mathrm{do}(X = h, Y = j)) = P(I \mid X = h, Y = j)$, which the known renderer supplies. The bound never intervenes on Z, so the bidirected edge $Z \leftrightarrow Y$ imposes no further constraint. The bound of a query optimizes the posterior response of that query's own image.

C.3.6 Uncertainty

Whiskers in Fig. 4. The whiskers on the L_1 and L_2 bars are plus and minus one standard error from a stratified bootstrap of the glyph bank with $n_{\mathrm{boot}} = 300$ replicates. Each replicate resamples the test-split images within every label with replacement, keeping each label's count, rebuilds B and R for both classifiers, and recomputes the L_1 and L_2 Shapley values for every digit; the standard error is the sample standard deviation of the replicates. The whiskers on the L_3 bars are the L_3 bounds of the same image and are not confidence intervals.

Table in Fig. 4. Each cell reports the fraction of the 1,000 queries that attribute to hue: for L_1, L_2 and L_3, a value with $|\phi_X|$ above the numerical tolerance $64\varepsilon_{\mathrm{mach}} \approx 1.4 \times 10^{-14}$; for L_3 bounds, an interval that excludes zero beyond the same tolerance. For digit d with k_d positive queries among $n_d = 100$, the exact two-sided binomial interval is computed at error $0.05/10$. The cell estimate is the mean of k_d/n_d over digits, and the cell interval averages the ten per-digit endpoints, so by the Bonferroni inequality it covers the digit-balanced fraction with probability at least 0.95. The paired difference between classifiers uses separate exact binomial intervals for positive and negative discordant queries within each digit at error $0.05/20$, again averaged over digits. These are per-cell guarantees; the run scans ten cell fractions and five paired differences, one of them for a method the figure does not display. The intervals describe factual image sampling conditional on the trained weights, the glyph bank and the known mechanisms; they exclude training-seed, bank and identification uncertainty.

C.3.7 Results

Table. For the standard classifier, L_1, L_2, L_3 SVs and L_3 bounds each attribute to hue in every query (Fig. 4); the 95% interval of each fraction is $[0.942, 1.000]$. For the robust classifier, L_1 and L_2 attribute to hue in every query, with the same 95% interval, while L_3 SVs and L_3 bounds attribute to hue in no query, with 95% interval $[0.000, 0.058]$. The paired difference between the standard and robust fractions is 1.000 with 95% interval $[0.871, 1.000]$ for L_3 SVs and for L_3 bounds, and 0.000 with 95% interval $[-0.065, 0.065]$ for L_1 and for L_2. D_1 requires zero hue attribution for the robust classifier, whose prediction cannot depend on hue: L_3 SVs and L_3 bounds satisfy it in every query and L_1 and L_2 violate it in every query. L_1 and L_2 attribute to hue for the robust classifier because hue and digit are dependent in the population, where hue causes the digit with probability α. D_2 requires nonzero hue attribution for the standard classifier: all four methods satisfy it in every query. An L_3 bounds interval that includes zero leaves zero compatible with the data; it does not identify a causal null.

Illustration images. For the ten illustration images, the standard classifier's $\phi_X^{L_3}$ for digits 0 through 9 is 0.883, 0.718, 0.868, 0.877, 0.764, 0.908, 0.884, 0.643, 0.792 and 0.679. Its L_3 bounds for hue have lower endpoints below the exact value by at most 3×10^{-3} for every digit; the upper endpoints

exceed the exact value by at most 10^{-3} for digits 0, 5, 6, 8 and 9 and by 0.223, 0.069, 0.049, 0.069 and 0.135 for digits 1, 2, 3, 4 and 7. The robust classifier's $\phi_X^{L_3}$ is exactly zero for every illustration image, and its $\phi_Y^{L_3}$ for digits 0 through 9 is 0.974, 0.962, 0.950, 0.938, 0.926, 0.900, 0.877, 0.829, 0.756 and 0.668. This decrease follows the digit law: for the robust classifier $\phi_Y^{L_3} = 1 - \sum_j p_j Q[d, j] \le 1 - p_d$, and p_d grows with d. The robust classifier's L_3 bounds for hue have lower endpoint zero and upper endpoints 0.488, 0.481, 0.475, 0.469, 0.463, 0.451, 0.439, 0.415, 0.380 and 0.336 for digits 0 through 9, so each includes zero; by efficiency its digit bounds are $[\phi_Y^{L_3} - u_d, \phi_Y^{L_3}]$ for hue upper endpoint u_d, with lower endpoints 0.487, 0.481, 0.475, 0.469, 0.463, 0.449, 0.438, 0.413, 0.376 and 0.332, so each excludes zero.

C.4 CelebA-HQ Experiment

This subsection specifies the structural causal models, classifier training, queries, oracle computation and identification bounds and reports the results behind Fig. 5. The oracle values come from one run of 2026-01-29; the bounds come from exact linear programs over the same classifier checkpoints and the full image bank.

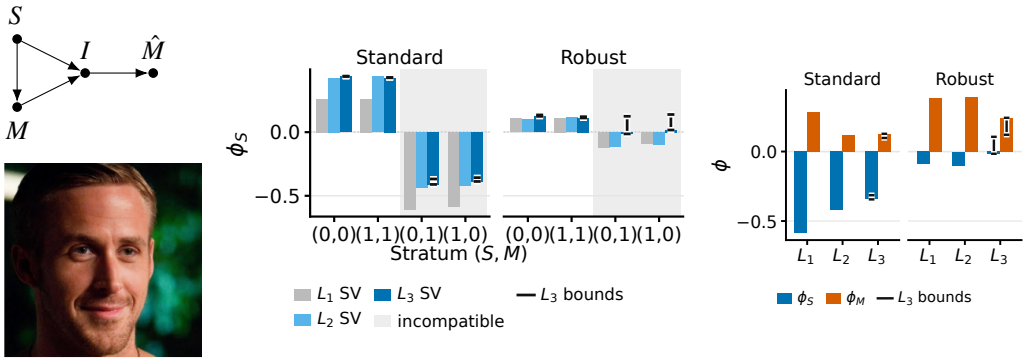


Figure 5: **Causal explanations on CelebA-HQ.** *Left:* Causal diagram and a sample image from the incompatible (S=1, M=0) stratum, smiling with the mouth closed. *Middle:* $L_{1:3}$ SVs ϕ_S per (S, M) stratum for both classifiers; the L_3 bar is the oracle mean over 100 Monte Carlo replicates with its 95% confidence interval, and the whiskers are the identification bounds $[\phi^-, \phi^+]$ of Alg. 1 from $P(S, M)$, the graph and the known classifier. *Right:* Attributions to smiling S and mouth M for the sample image; whiskers denote identification bounds for L_3 SVs. In the incompatible strata $L_{1:2}$ SVs attribute the robust classifier's prediction to S, while L_3 SVs are close to zero, with the residual explained by leakage through the $S \to I \to \hat{M}$ path, which interventional training does not remove. The standard classifier's bounds exclude zero in every stratum (D_2); the robust classifier's bounds contain zero at $(0, 1)$ and exclude it at $(1, 0)$. Table 8 lists the values.

Clf	(S, M)	Oracle $\phi_S^{L_3}$	$[\phi^-, \phi^+]$	$0 \in$?
Std	$(0, 0)$	$+0.438 \pm 0.0020$	$[+0.437, +0.441]$	No
Std	$(1, 1)$	$+0.427 \pm 0.0021$	$[+0.428, +0.430]$	No
Std	$(0, 1)$	-0.411 ± 0.0019	$[-0.411, -0.368]$	No
Std	$(1, 0)$	-0.387 ± 0.0020	$[-0.387, -0.357]$	No
Rob	$(0, 0)$	$+0.124 \pm 0.0006$	$[+0.124, +0.135]$	No
Rob	$(1, 1)$	$+0.111 \pm 0.0007$	$[+0.112, +0.121]$	No
Rob	$(0, 1)$	-0.014 ± 0.0005	$[-0.014, +0.125]$	Yes
Rob	$(1, 0)$	$+0.017 \pm 0.0005$	$[+0.017, +0.139]$	No

Table 8: **CelebA-HQ oracle L_3 SVs and identification bounds** ϕ_S per classifier and (S, M) stratum: the oracle mean over 100 Monte Carlo replicates with its 95% confidence half-width, the exact bounds $[\phi^-, \phi^+]$, and whether they contain zero.

C.4.1 Structural Causal Models

Smiling S influences the mouth-open attribute M, both select the face image I from a bank of CelebA-HQ images, and a classifier outputs $\hat{M} = \hat{f}(I) \in [0, 1]$, its probability that $M = 1$. The image bank is the CelebA-HQ training split, with 24,120 images in four attribute cells (Smiling, Mouth_Slightly_Open) $= (s, m)$ of sizes 7,111 for $(0, 0)$, 5,651 for $(0, 1)$, 2,165 for $(1, 0)$ and 9,193 for $(1, 1)$. Two models share the mechanisms of S and M:

$$U_S \sim \text{Bern}(0.5), \qquad U_M \sim \text{Cat}(0.075, 0.85, 0.075), \tag{71}$$

$$S = U_S, \qquad M = \mathbf{1}[U_M = 1] \cdot S + \mathbf{1}[U_M = 2]. \tag{72}$$

The three values of U_M make the mouth always closed with probability 0.075, equal to S with probability 0.85, and always open with probability 0.075.

Training model. The classifier training data come from a model with a mislabeled image lookup: with $U_{I_1} \sim \text{Bern}(0.4)$ and $U_{I_2} \sim \text{Bern}(0.5)$, the lookup state is $M_I = U_{I_1} U_{I_2} + (1 - U_{I_1})M$ and the image is a uniform draw from cell (S, M_I). The image therefore shows a random mouth state instead of M with probability $0.4 \times 0.5 = 0.2$, so visible mouth features predict M with accuracy 0.8, while S predicts M with accuracy $P(M = S) = 0.85 + 2 \cdot 0.075 \cdot 0.5 = 0.925$, so a classifier gains accuracy by using smiling features.

Attribution model. All attributions and bounds use the model without mislabeling. A unit U_I holds one uniformly drawn image from each of the four cells, the four draws independent of each other and of (U_S, U_M), and I is the image of cell (S, M). The exogenous variables U_S, U_M and U_I are mutually independent, so every c-component is a singleton; the graph is $S \to M$, $S \to I$, $M \to I$, $I \to \hat{M}$. The players are S and M.

C.4.2 Classifier Training

Both classifiers use a LeNet architecture: two convolutional layers with 6 and 16 filters of size 5×5, then fully connected layers of width 120, 84 and 1, trained on 256×256 images. Training uses Adam at learning rate 10^{-4}, batch size 512 and bf16 mixed precision for up to 32 epochs on 100,000 samples from the training model with 5,000 validation samples. The *standard classifier* trains on observational samples, in which S and M are correlated. The *robust classifier* trains on samples generated under $\text{do}(M \sim P(M))$, which removes the dependence between S and M and makes smiling features uninformative for M. The checkpoint of each classifier is the epoch with the smallest validation loss on its own training distribution. At that epoch, the standard classifier has accuracy 0.936 on the observational validation set and 0.528 on the $\text{do}(M \sim P(M))$ validation set; the robust classifier has 0.766 and 0.776. These accuracies are logged scalars on the validation sets and carry no confidence interval.

C.4.3 Strata

Every unit falls into one of four (S, M) strata. In the incompatible strata $S \neq M$, the value $U_M = 1$ is impossible, so S did not cause M and D_1 requires $\phi_S^{L_3} = 0$ for a classifier that does not use smiling features. In the compatible strata $S = M$, the unit has $U_M = 1$ with probability $0.85/0.925 \approx 0.919$ and is an always-closed or always-open unit otherwise.

C.4.4 Queries and Value Function

A stratum query has evidence $e = (S = s, M = m)$ and explains the prediction of the observed mouth state. The score of an evaluation with image I is $P(\hat{M} = m) = \hat{f}(I)$ if $m = 1$ and $1 - \hat{f}(I)$ if $m = 0$: the probability of predicting the factual label, which differs from correctness when the counterfactual evaluation changes M. The value function of Def. 3.2 is the natural total effect with $u \sim P(U \mid e)$ and an independent baseline $u' \sim P(U)$ and the same U_I in both terms:

$$v(T) = \mathbb{E}\big[\text{score}(\hat{M}(u)) - \text{score}(\hat{M}_{T \leftarrow T(u')}(u)) \mid e\big], \tag{73}$$

and $\phi_S = \frac{1}{2}\big[v(\{S\}) - v(\varnothing)\big] + \frac{1}{2}\big[v(\{S, M\}) - v(\{M\})\big]$, with ϕ_M symmetric. The L_1 game conditions on the coalition and the L_2 game intervenes on it in the population, as in Sec. C.3.4. An image query adds $I = i$ to the evidence; since the four cell draws of U_I are independent, the image fixes only the (s, m) cell of the unit, and L_1 and L_2 values of an image query equal those of its stratum.

C.4.5 Oracle Computation

For each classifier and stratum we ran 100 Monte Carlo replicates. Each replicate draws 100,000 candidate units from the attribution model, keeps 2,000 of those whose (S, M) equals (s, m), pairs each with a fresh baseline unit, draws the images of both from the bank, and averages the L_1, L_2 and L_3 Shapley values over the 2,000 pairs. Every reported oracle value is the mean over the 100 replicates; its 95% confidence half-width is 1.96 standard errors across replicates and is at most 0.0021 in every cell of Table 9. The confidence intervals describe Monte Carlo error at fixed classifier weights and image bank.

C.4.6 Identification Bounds

The bounds use the graph, the law $P(S, M)$, the image bank and the known classifier, and range over the unknown mechanism of M. By Sec. C.1, the free c-component is $\{M\}$ and its types are the four functions $t : \{0, 1\} \to \{0, 1\}$: the constants $t_0 = 0$ and $t_1 = 1$, the identity $t_{\mathrm{id}}(s) = s$ and the negation $t_{\mathrm{neg}}(s) = 1 - s$, with a law q over the four types. The data constrain q through $P(M = m \mid S = s)$ for $s \in \{0, 1\}$ and the simplex equation, five equality rows of rank 3, leaving one free dimension, the negation mass $q(t_{\mathrm{neg}}) \in [0, 0.075]$. The true model has $q = (0.075, 0.85, 0, 0.075)$ on $(t_0, t_{\mathrm{id}}, t_{\mathrm{neg}}, t_1)$. For the cell means $g_m(s', m')$, the mean of the score for label m over all bank images of cell (s', m'), the coalition values are

$$v(\{S\}) = g_m(s, m) - \frac{1}{P(m \mid s)} \sum_t q(t)\, \mathbb{1}[t(s) = m] \sum_{s'} P(s')\, g_m(s', t(s')), \tag{74}$$

$$v(\{M\}) = g_m(s, m) - \sum_{m'} P(m')\, g_m(s, m'), \tag{75}$$

$$v(\{S, M\}) = g_m(s, m) - \sum_{s', m'} P(s', m')\, g_m(s', m'), \qquad v(\varnothing) = 0, \tag{76}$$

so only $v(\{S\})$ depends on q and ϕ_S and ϕ_M are affine in q. Each endpoint is one linear program over $q \geq 0$ with the five equality rows. For an image query, $g_m(s, m)$ is replaced by the score of the observed image and the other three cell means are unchanged. The 8 stratum queries and 2 image queries, 2 players and both directions make 40 programs; all returned the status optimal with residual 0. The cell means use the full bank, so the bounds involve no sampling, and the value of ϕ_S at the true q is the quantity the oracle estimates by Monte Carlo.

C.4.7 Results

Oracle Shapley values. Table 9 reports the oracle values per stratum. In the compatible strata both classifiers have positive ϕ_S at all three layers. In the incompatible strata the standard classifier has negative ϕ_S of large magnitude at all three layers. The robust classifier keeps a nonzero $|\phi_S|$ there at L_1 and L_2, which do not condition on the unit and so carry the population dependence between S and M, while its L_3 stratum means are at least five times smaller; the largest $|\phi_S|$ over the 100 replicates is 0.0200 at $(0, 1)$ and 0.0221 at $(1, 0)$. A threshold of 0.05 on $|\phi_S|$ separates these two L_3 means from every other stratum mean in Table 9.

Identification bounds. Table 8 reports the exact interval $[\phi^-, \phi^+]$ per stratum beside the oracle L_3 mean, and the middle panel of Fig. 5 plots them. The lower endpoint equals the value of ϕ_S at the true q in every row: the true model puts zero mass on the negation type, a vertex of the feasible segment, and ϕ_S increases with the negation mass in every stratum. The oracle mean lies strictly inside the interval in 3 of the 8 rows and below the lower endpoint by 0.0002 to 0.0009 in the other 5, within its own 95% confidence interval in 4 of them; the robust $(1, 1)$ mean lies 0.0009 below the endpoint, 1.3 confidence half-widths. Since the lower endpoint is the exact value, Monte Carlo error places the oracle mean below it with probability about one half per row, so 5 of 8 is consistent with this. The robust $(1, 1)$ gap of 0.0009 exceeds that row's oracle confidence half-width in Fig. 5, consistent with the same Monte Carlo error. The L_1 and L_2 values computed in closed form from the cell means agree with the oracle means to within 0.0015 in every cell, which checks the cell means and the value function independently of the programs.

The standard classifier's intervals exclude zero in every stratum, so D_2 holds for every compatible model. The robust classifier's intervals exclude zero in the compatible strata, contain zero at $(0, 1)$

Table 9: **CelebA-HQ oracle SVs** ϕ_S for the four (S, M) strata. Each cell is the mean over 100 Monte Carlo replicates of 2,000 units; the 95% confidence half-width, 1.96 standard errors across replicates, is at most 0.0021 in every cell. The incompatible mean averages the two incompatible strata equally and carries no separate interval.

Classifier	(S, M)	L_1 ϕ_S	L_2 ϕ_S	L_3 ϕ_S
Standard	$(0, 0)$	+0.260	+0.425	+0.438
Standard	$(1, 1)$	+0.261	+0.436	+0.427
Standard	$(0, 1)$	−0.611	−0.437	−0.411
Standard	$(1, 0)$	−0.587	−0.423	−0.387
Standard	Incompat. mean	**−0.599**	**−0.430**	**−0.399**
Robust	$(0, 0)$	+0.108	+0.100	+0.124
Robust	$(1, 1)$	+0.111	+0.117	+0.111
Robust	$(0, 1)$	−0.124	−0.117	−0.014
Robust	$(1, 0)$	−0.092	−0.100	+0.017
Robust	Incompat. mean	**−0.108**	**−0.109**	**+0.001**

and have a positive lower endpoint at $(1, 0)$ (Fig. 5). The $(1, 0)$ interval excludes zero only because its lower endpoint is the true value, which is positive; the positive upper endpoints at both incompatible strata show that the data admit models with positive ϕ_S there. The robust classifier's oracle L_3 values in the incompatible strata (Table 9) are leakage through the $S \to I \to \hat{M}$ path: interventional training removes the dependence between S and M in the training data but does not enforce invariance of $\hat{f}$ to smiling features, unlike the pointwise maximum over the RGB channels in Sec. C.3.2, which removes the hue from the input. The robust classifier's oracle ϕ_S therefore lies within 0.0221 of zero in every replicate but is not exactly zero, so D_1 holds only for a classifier invariant to smiling features.

Image query. The left panel of Fig. 5 shows CelebA-HQ training image 13011, whose labels are Smiling = 1 and Mouth_Slightly_Open = 0, so it belongs to the incompatible $(1, 0)$ stratum. For this image, the standard classifier outputs $\hat{f}(i) = 0.713$ and has $\phi_S^{L_1} = -0.588$, $\phi_S^{L_2} = -0.424$, $\phi_S^{L_3} = -0.344$ under the true law and $[\phi^-, \phi^+] = [-0.344, -0.314]$; the robust classifier outputs $\hat{f}(i) = 0.298$ and has $\phi_S^{L_1} = -0.093$, $\phi_S^{L_2} = -0.101$, $\phi_S^{L_3} = -0.016$ under the true law and $[\phi^-, \phi^+] = [-0.016, +0.106]$. The L_1 and L_2 values equal the closed-form $(1, 0)$ stratum values, because they do not condition on the image; they differ from the Monte Carlo stratum means of Table 9 by at most 0.001. The standard classifier's interval excludes zero and the robust classifier's interval contains zero: for this image the robust classifier's value under the true law is negative, the opposite sign of the $(1, 0)$ stratum mean in Table 9, and the bounds leave the sign undecided.

D Proofs

Notation: Throughout this appendix, we use unit-level notation $\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})$ for counterfactual differences at specific units and expectation notation $\text{NTE}(\mathbf{Z}, Y|\mathbf{e})$ for expectations over units consistent with why-query $w = \text{Why}(y|\mathbf{e})$. In unit-baseline pair $(\mathbf{u}, b)$, unit $\mathbf{u}$ induces the observed evidence $\mathbf{e}$ conditioning the why-query, while baseline $b = (x', \mathbf{z}')$ represents counterfactual values. We write $c(\mathbf{u}, b) := c(\mathcal{M}, \mathbf{u}, X, Y, b)$ when $\mathcal{M}, X, Y$ are clear from context, and $c_{\mathcal{M}}(b)$ (matching Def. 2.3) when $\mathbf{u}$ is also implicit. We abbreviate $D_i(\mathcal{M}, c, \phi)$ (Def. A.6) as $D_i(c)$ when $\mathcal{M}, \phi$ are clear from context. For D_2, we work with the premise at $\mathbf{z}'_\varnothing$ in Def. 2.4.

Proof Structure: We prove six main results in order:

1. **Lifting (Sec. D.1):** Conditions for proxy violations to imply oracle violations
2. **NTE-Functional Equivalence (Sec. D.2):** NTEs equal functional dependence marginal contributions
3. **Impossibility (Sec. D.3):** No data-only method satisfies all desiderata
4. **Validity and Tightness (Sec. D.4):** Our algorithm yields valid and tight bounds
5. **Desiderata Satisfaction (Sec. D.5):** L_3 SVs satisfy all four desiderata
6. **Shapley Axioms (Sec. D.6):** L_3 SVs are the unique Shapley decomposition of NTEs

D.1 Proof of Desiderata Lifting Theorem

We discuss extensions of desiderata to definitions of actual causation in Sec. A.1.1.

Theorem A.7 (Desiderata Lifting)**.** *Under Assump. A.5, violations of D_1, D_3, D_4 and satisfaction of D_2 lift from functional dependence c to causation c'. For D_1, violations lift unconditionally by conservativity. For D_2, satisfaction lifts unconditionally; violations additionally lift when there exists a set of unit-baseline pairs $\{(\mathbf{u}, z)\}$ with positive measure (where $\mathbf{u}$ induces observed evidence $\mathbf{e}$) satisfying $c = c' \neq 0$ and $\phi = 0$. For D_3, D_4, violations lift when $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$ for the SCMs in the comparison. Property P lifts from c to c' if $P(\mathcal{M}, c, \phi) \Rightarrow P(\mathcal{M}, c', \phi)$ for all SCMs $\mathcal{M}$.* □

Proof. By Assump. A.5, we have $c' \neq 0 \Rightarrow c \neq 0$ (necessity) and when $c' \neq 0$, there exists mapping $f : \mathcal{Z}' \to \mathcal{Z}$ such that $c'(z') = c(f(z'))$ (effect preservation).

For D_1 (Admissibility), if $\neg D_1(c, \phi)$, then $c_B \equiv \delta_0$ but $\phi \neq \mathbf{0}$. If $c' \neq 0$ on a set of oracle baselines with positive probability, f maps it to a set of baselines with positive probability where $c \neq 0$, contradicting $c_B \equiv \delta_0$; thus $\neg D_1(c', \phi)$.

For D_2 (Power), satisfaction lifts unconditionally: if the premise of $D_2(c')$ holds, then by Assump. A.5 the premise of $D_2(c)$ holds, so $\phi \neq 0$ and $D_2(c')$ is satisfied. For violations, assume the method violates $D_2(c)$ by assigning $\phi = 0$ despite $\exists(\mathbf{u}, z) : c(\mathbf{u}, z) \neq 0$. Violations lift when there exists a positive-measure set of unit-baseline pairs $\{(\mathbf{u}'', z'')\}$ satisfying (i) $c(\mathbf{u}'', z'') = c'(\mathbf{u}'', z'') \neq 0$ and (ii) the method assigns $\phi = 0$. In this case, the method assigns $\phi = 0$ despite $c'(\mathbf{u}'', z'') \neq 0$ for a positive-measure set of pairs, violating $D_2(c')$.

For D_3, D_4 (Effect Scaling, Normality), assume method violates $D_4(c)$ on SCMs $\mathcal{M}, \mathcal{M}'$. By definition of D_4, the premises hold under c: we have $c_{\mathcal{M}} = c_{\mathcal{M}'}$ and $P_+^{\mathcal{M}}(b) \geq P_+^{\mathcal{M}'}(b)$ for all b, but the conclusion fails since $\phi_{\mathcal{M}} \leq \phi_{\mathcal{M}'}$.

Suppose additionally that $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$. From $c_{\mathcal{M}} = c_{\mathcal{M}'}$ combined with these equality conditions, we obtain $c'_{\mathcal{M}} = c'_{\mathcal{M}'}$.

Baseline probabilities $P^{\mathcal{M}}(B)$ and $P^{\mathcal{M}'}(B)$, and the distribution $P(\mathbf{U} \mid \mathbf{e})$, are determined by the SCM structure and are independent of which causation notion we use. When $c_{\mathcal{M}} = c'_{\mathcal{M}}$, the sets $\{b : c(\mathcal{M}, \mathbf{u}, X, Y, b) \neq 0\}$ and $\{b : c'(\mathcal{M}, \mathbf{u}, X, Y, b) \neq 0\}$ coincide.

Similarly, equality $c_{\mathcal{M}} = c'_{\mathcal{M}}$ implies $\text{sign}(c_{\mathcal{M}}) = \text{sign}(c'_{\mathcal{M}})$, so $P_+^{\mathcal{M}}(b)$ under c equals $P_+^{\mathcal{M}}(b)$ under c'. Equality of these signs at each unit, within each SCM, carries the D4(c) premise $P_+^{\mathcal{M}}(B) \not\equiv P_+^{\mathcal{M}'}(B)$ to c'. Therefore, all three premises of $D_4(c')$ hold but the conclusion still fails with $\phi_{\mathcal{M}} \leq \phi_{\mathcal{M}'}$, so the method violates $D_4(c')$.

For D_3, assume the method violates $D_3(c)$: we have $c_{\mathcal{M}}(b) \geq c_{\mathcal{M}'}(b)$ for all b, $P^{\mathcal{M}}(B) \equiv P^{\mathcal{M}'}(B)$, and $P^{\mathcal{M}}(c_B \mid \mathbf{z}'_{\varnothing}) \not\equiv P^{\mathcal{M}'}(c_B \mid \mathbf{z}'_{\varnothing})$ (effect distributions differ), but $\phi_{\mathcal{M}} \leq \phi_{\mathcal{M}'}$. Under $c_{\mathcal{M}} = c'_{\mathcal{M}}$ and $c_{\mathcal{M}'} = c'_{\mathcal{M}'}$, we have $c'_{\mathcal{M}}(b) \geq c'_{\mathcal{M}'}(b)$ for all b. Baseline distributions and $P(\mathbf{U} \mid \mathbf{e})$ are independent of which causation notion we use, so $P^{\mathcal{M}}(B) \equiv P^{\mathcal{M}'}(B)$ still holds and the two models keep the same $P(\mathbf{U} \mid \mathbf{e})$. Since $c_{\mathcal{M}} = c'_{\mathcal{M}}$ pointwise, the effect distribution $P^{\mathcal{M}}(c_B \mid \mathbf{z}'_{\varnothing})$ under c equals $P^{\mathcal{M}}(c'_B \mid \mathbf{z}'_{\varnothing})$ under c', and similarly for $\mathcal{M}'$; thus $P^{\mathcal{M}}(c'_B \mid \mathbf{z}'_{\varnothing}) \not\equiv P^{\mathcal{M}'}(c'_B \mid \mathbf{z}'_{\varnothing})$. Therefore the premises of $D_3(c')$ hold but $\phi_{\mathcal{M}} \leq \phi_{\mathcal{M}'}$, so the method violates $D_3(c')$. □

Having established the conditions under which desiderata violations lift from proxy to oracle, we next prove that NTEs capture the functional dependence proxy, enabling construction of desiderata-satisfying methods.

D.2 Proof of NTE-Functional Dependence Equivalence

Lemma A.9 (NTE-Functional Dependence Equivalence)**.** *Let $\mathcal{M}$ be a recursive SCM with discrete exogenous domain* $\text{Dom}(\mathbf{U})$*, and let $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$. For any baseline $b = (x', \mathbf{z}') \in \text{Dom}(X) \times \text{Dom}(\mathbf{Z})$ with $P_{X,\mathbf{Z}}(x', \mathbf{z}') > 0$, functional dependence c (Def. 2.3) equals the marginal contribution of X to an* NTE *of $\mathbf{Z} \cup \{X\}$ on Y under transition $\mathbf{u}' \to \mathbf{u}$:*

$$\begin{aligned} c(\mathcal{M}, \mathbf{u}, X, Y, b) =& \text{NTE}(\mathbf{Z} \cup \{X\}, Y | \mathbf{u}' \to \mathbf{u}) \\ & - \text{NTE}(\mathbf{Z}, Y | \mathbf{u}' \to \mathbf{u}) \end{aligned} \tag{17}$$

for any $\mathbf{u}'$ *such that* $(X(\mathbf{u}'), \mathbf{Z}(\mathbf{u}')) = (x', \mathbf{z}')$ *(the right-hand side does not depend on the choice). For continuous* $\mathrm{Dom}(\mathbf{U})$*, the equivalence holds* $P_{X,\mathbf{Z}}$*-almost everywhere via disintegration.* □

Proof. For any baseline $z = (x', \mathbf{z}')$ with $P^{\mathcal{M}}(Z = z) > 0$, there must exist $\mathbf{u}' \in \mathrm{Dom}(\mathbf{U})$ inducing witness values $(X(\mathbf{u}'), \mathbf{Z}(\mathbf{u}')) = (x', \mathbf{z}')$. In the discrete case, this requires $P^{\mathcal{M}}(\mathbf{U} = \mathbf{u}') > 0$; in the continuous case, it requires positive density $\rho(\mathbf{u}') > 0$.

From Def. 2.3, the functional dependence is:

$$c(\mathcal{M}, \mathbf{u}, X, Y, z) = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u}) \tag{77}$$

From Def. 3.1, the natural total effect is:

$$\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \tag{78}$$

Thus we have:

$$\begin{aligned} \mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) &= Y_{\mathbf{Z}(\mathbf{u}),X(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}'),X(\mathbf{u}')}(\mathbf{u}) && (79)\\ &= Y_{\mathbf{z},x}(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u}) && (80) \end{aligned}$$

where $x = X(\mathbf{u})$ and $\mathbf{z} = \mathbf{Z}(\mathbf{u})$.

Similarly:

$$\begin{aligned} \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) &= Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) && (81)\\ &= Y_{\mathbf{z}}(\mathbf{u}) - Y_{\mathbf{z}'}(\mathbf{u}) && (82) \end{aligned}$$

Taking the marginal contribution:

$$\begin{aligned} &\mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) && (83)\\ &= [Y_{\mathbf{z},x}(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u})] - [Y_{\mathbf{z}}(\mathbf{u}) - Y_{\mathbf{z}'}(\mathbf{u})] && (84) \end{aligned}$$

By counterfactual consistency [1] (setting variables to their actual values yields the factual outcome), we have $Y(\mathbf{u}) = Y_{\mathbf{z},x}(\mathbf{u}) = Y_{\mathbf{z}}(\mathbf{u})$ since $\mathbf{z}, x$ are the actual values at unit $\mathbf{u}$. Substituting:

$$\begin{aligned} &= Y(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u}) - [Y(\mathbf{u}) - Y_{\mathbf{z}'}(\mathbf{u})] && (85)\\ &= Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u}) = c(\mathcal{M}, \mathbf{u}, X, Y, z) && (86) \end{aligned}$$

□

The NTE-functional equivalence establishes that our value function captures the causation proxy. We now prove that no method using only observational and interventional data can satisfy all desiderata, motivating the need for structural assumptions.

D.3 Proof of Explanatory Impossibility Theorem

Theorem 2.5 (Explanatory Impossibility Theorem)**.** *There exists no EVA* ϕ *that satisfies* $D_{1:4}$ *for all* $\mathcal{M} \in \Omega$ *and depends only on* $L_{1:2}$ *data: i.e., there is no function* g *with* $\phi_X(\mathcal{M}, w) = g(\mathcal{P}^{L_{1:2}}(\mathcal{M}), w, X)$ *for all* $X \in \mathbf{X}$*, where* $\mathbf{V}_{\mathbf{z}}$ *denotes* $\mathbf{V}$ *under* $\mathrm{do}(\mathbf{Z}{=}\mathbf{z})$ *and* $\mathcal{P}^{L_{1:2}}(\mathcal{M}) = \{P^{\mathcal{M}}(\mathbf{V}_{\mathbf{z}}) : \mathbf{Z} \subseteq \mathbf{V}, \mathbf{z} \in \mathrm{Dom}(\mathbf{Z})\}$. □

Proof. We construct two SCMs that are indistinguishable using only L_1/L_2 data (observational and interventional distributions) but have opposite functional dependence structures: one where X functionally determines Y (requiring $\phi_X \neq 0$ by power) and one where X is functionally independent of Y (requiring $\phi_X = 0$ by admissibility). Since the two models induce identical L_1 and L_2 distributions, any method using only this data will assign the same value ϕ_X in both models, but the desiderata require different values.

Intuition. Consider two scenarios with identical distributions: (i) a random bet X on a coin flip that wins or loses with equal probability, and (ii) two independent fair coin flips X, Y. Both produce uniform $P(X, Y)$, yet the bet causes its outcome while the independent flip does not.

Formal construction. Take $\mathbf{V} = \{X, Y\}$, $\mathbf{X} = \{X\}$, and why-query $w = \text{Why}(Y{=}1 \mid X{=}1, Y{=}1; \{X\})$. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ be binary domains. Define two SCMs:

$$\mathcal{M}_1 : \quad X = U_X, \quad Y = \mathbf{1}[X = U_Y] \quad \text{(causal diagram: } X \to Y\text{)} \tag{87}$$

$$\mathcal{M}_2 : \quad X = U_X, \quad Y = U_Y \quad \text{(causal diagram: } X \quad Y \text{ (no edge))} \tag{88}$$

where $U_X, U_Y \overset{i.i.d.}{\sim} \text{Bernoulli}(0.5)$. $\mathcal{M}_1$ has causal edge $X \to Y$ with functional dependence, while $\mathcal{M}_2$ has no edge and no functional dependence.

We verify $\mathcal{P}^{L_{1:2}}(\mathcal{M}_1) = \mathcal{P}^{L_{1:2}}(\mathcal{M}_2)$ by direct computation across all $\mathbf{Z} \subseteq \{X, Y\}$:

- $\mathbf{Z} = \varnothing$ (observational): $P(X{=}x, Y{=}y) = 0.25$ for all $x, y \in \{0, 1\}$ in both models.
- $\mathbf{Z} = \{X\}$: under $\text{do}(X{=}x)$, X is fixed at x, so $P(X, Y \mid \text{do}(X{=}x)) = \delta_{x'=x} \cdot \text{Bern}(0.5)$ on Y in both models ($\mathcal{M}_1$: $Y = \mathbf{1}[x = U_Y]$ marginalizes to Bern(0.5); $\mathcal{M}_2$: $Y = U_Y$ is independent of the do).
- $\mathbf{Z} = \{Y\}$: under $\text{do}(Y{=}y)$, Y is fixed at y and $X = U_X$ is unaffected, giving $\text{Bern}(0.5) \cdot \delta_{y'=y}$ in both.
- $\mathbf{Z} = \{X, Y\}$: point mass at (x, y) in both.

Thus all L_1 and L_2 distributions coincide between the two models.

However, the functional dependence evaluations differ between these models. In $\mathcal{M}_1$ at context $(X = 1, Y = 1)$, the exogenous values are $\mathbf{u} = (U_X = 1, U_Y = 1)$. The counterfactual outcome under intervention $X = 0$ is $Y_0(\mathbf{u}) = \mathbf{1}[0 = 1] = 0 \neq 1 = Y(\mathbf{u})$, so functional dependence satisfies $c(\mathcal{M}_1, \mathbf{u}, X, Y, b) \neq 0$ for baseline $b = (\mathbf{z}'_\varnothing, x'{=}0)$ (with $\mathbf{Z}' = \varnothing$ in this two-variable model). Since P_B assigns positive mass to baseline $x'{=}0$ and $c \neq 0$ at this baseline, the law $P^{\mathcal{M}_1}(c_B \mid \mathbf{z}'_\varnothing)$ is not δ_0, satisfying D_2's premise. In contrast, for $\mathcal{M}_2$ at any context, the counterfactual satisfies $Y_{x'}(\mathbf{u}) = U_Y = Y(\mathbf{u})$, so $c(\mathcal{M}_2, \mathbf{u}, X, Y, z) = 0$ for all witnesses z.

By power (D_2) and admissibility (D_1), the model $\mathcal{M}_1$ requires $\phi_X \neq 0$ while $\mathcal{M}_2$ requires $\phi_X = 0$. Both $\mathcal{M}_1$ and $\mathcal{M}_2$ induce the same L_1 and L_2 distributions. Any method using only L_1 and L_2 data will therefore assign the same value ϕ_X in both models. However, power (D_2) requires $\phi_X \neq 0$ in $\mathcal{M}_1$, while admissibility (D_1) requires $\phi_X = 0$ in $\mathcal{M}_2$. Therefore, no method using only observational and interventional data can jointly satisfy D_1 and D_2. Thus, no such method can satisfy all desiderata $D_{1:4}$ jointly. The same pair shows that no such method satisfies D_3 alone: the two models have the same $\mathbf{U}$ and the same $P(\mathbf{U} \mid \mathbf{e})$ and equal $L_{1:2}$ distributions, and $c_{\mathcal{M}_1}$ is 1 at $x' = 0$ and 0 at $x' = 1$ while $c_{\mathcal{M}_2}$ is 0, so D_3 requires $\phi_{\mathcal{M}_1} > \phi_{\mathcal{M}_2}$. □

The impossibility result establishes inherent limits on point estimation. We therefore turn to bounding, proving that our algorithm returns valid and tight bounds.

D.4 Proof of Validity and Tightness

Theorem 4.2 (Valid and Tight Bounds)**.** *Under Assump. 4.1, Alg. 1 returns* $[\phi^-, \phi^+]$ *such that:* (validity) $\phi_X^{L_3}(\mathcal{M}, w) \in [\phi^-, \phi^+]$ *for every* $\mathcal{M} \in \mathcal{S}(\mathcal{T}, \mathcal{G})$*;* (tightness) *each endpoint is attained at some* $\hat{\mathcal{M}}^\pm \in \hat{\Omega}(\mathcal{T}, \mathcal{G}) \subseteq \mathcal{S}(\mathcal{T}, \mathcal{G})$*, where* $\hat{\mathcal{M}}^-$ *and* $\hat{\mathcal{M}}^+$ *may differ.* □

Proof. We show the argmin and argmax in Alg. 1 are attained, then verify validity and tightness.

Existence. By Assump. 4.1, Θ is compact and $\theta \mapsto \phi_X^{L_3}(w; \mathcal{M}_\theta)$ is continuous, so the argmin and argmax over $\hat{\Omega}(\mathcal{T}, \mathcal{G})$ in lines 3–4 of Alg. 1 are attained. Let $\hat{\mathcal{M}}^\pm \in \hat{\Omega}(\mathcal{T}, \mathcal{G})$ denote a minimizer and maximizer; set $\phi^\pm := \phi_X^{L_3}(w; \hat{\mathcal{M}}^\pm)$.

Validity. Fix any $\mathcal{M}^* \in \mathcal{S}(\mathcal{T}, \mathcal{G})$. By Assump. 4.1, there exists $\theta \in \Theta$ such that $\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T}, \mathcal{G})$ induces the same L_3 distribution as $\mathcal{M}^*$. Since $\phi_X^{L_3}(w; \cdot)$ is a functional of the L_3 distribution (Def. 3.2 and following remark), $\phi_X^{L_3}(w; \mathcal{M}^*) = \phi_X^{L_3}(w; \mathcal{M}_\theta)$. By definition of $\hat{\mathcal{M}}^\pm$, for every $\mathcal{M}_\theta \in \hat{\Omega}(\mathcal{T}, \mathcal{G})$:

$$\phi^- \leq \phi_X^{L_3}(w; \mathcal{M}_\theta) \leq \phi^+. \tag{89}$$

Combining the two gives $\phi_X^{L_3}(w; \mathcal{M}^*) \in [\phi^-, \phi^+]$.

Tightness. The witnesses $\hat{\mathcal{M}}^{\pm}$ lie in $\hat{\Omega}(\mathcal{T}, \mathcal{G}) \subseteq \mathcal{S}(\mathcal{T}, \mathcal{G})$ by definition of $\hat{\Omega}$ (Sec. 4), so each endpoint $\phi^{\pm}$ is attained at some $\mathcal{M}^{\pm} \in \mathcal{S}(\mathcal{T}, \mathcal{G})$. □

D.5 Proof that L_3 SVs Satisfy All Desiderata

We prove that L_3 Shapley Values satisfy all four desiderata under functional dependence c.

D.5.1 Proof of D_1 (Admissibility)

Lemma D.1 (Causal Admissibility)**.** *L_3 SVs satisfy D_1 (admissibility).* □

Proof. The literal D_1 premise (Def. 2.4) states $c_B \equiv \delta_0$: the random variable $c(B)$ under $B \sim P_B$ and $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$ is almost-surely zero. Since the L_3 Shapley sum is an expectation under P_B and $P(\mathbf{U} \mid \mathbf{e})$, points where $c \neq 0$ on a null set contribute zero. We therefore work with $c \equiv 0$ pointwise on the support of P_B and $P(\mathbf{U} \mid \mathbf{e})$.

Suppose $c(\mathcal{M}, \mathbf{u}, X, Y, z) = 0$ for all baselines z in the support of P_B and all $\mathbf{u}$ in the support of $P(\mathbf{U} \mid \mathbf{e})$. By the NTE-functional equivalence (Sec. D.2), every marginal NTE contribution $\mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = 0$ for all $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$ and all units $\mathbf{u}$ in the support of $P(\mathbf{U} \mid \mathbf{e})$ and $\mathbf{u}'$ in the support of $P(\mathbf{U})$. The L_3 Shapley value (Def. 3.2) is an expectation of these marginal contributions, so $\phi_X^{L_3}(w) = 0$. □

D.5.2 Proof of D_2 (Power)

Lemma D.2 (Weak Causal Power)**.** *Fix a causal diagram G, a finite domain* $\mathrm{Dom}(\mathbf{U})$ *and $P(\mathbf{U})$. For all structural parameters outside a set of Lebesgue measure zero (Sec. A.9), L_3 SVs satisfy D_2 at every why-query whose evidence has positive probability.* □

Proof. Exclude the set of measure zero on which some structural equation takes one value at two inputs. Then the value of a variable determines its input, and hence, by induction along G, the values of its ancestors and their exogenous parents.

Fix a why-query with evidence $\mathbf{e}$ of positive probability, containing $Y = y$, and fix $X \in \mathbf{X}$ for which the premise of D_2 holds. There is one input i^* with $f_Y(i^*) = y$, and every unit consistent with $\mathbf{e}$ reaches it. These units therefore agree on the exogenous parents of Y and of its ancestors, so they have equal counterfactual outcomes. Writing $\mathbf{u}^*$ for any of them, the expectation over $P(\mathbf{U} \mid \mathbf{e})$ reduces to $\mathbf{u}^*$, and the equivalence between the NTE and functional dependence (Sec. D.2) yields

$$\phi_X^{L_3} = \sum_{\mathbf{S} \subseteq \mathbf{X} \setminus \{X\}} w_{\mathbf{S}} \sum_{\mathbf{u}'} P(\mathbf{u}') \left[Y_{\mathbf{S}(\mathbf{u}')}(\mathbf{u}^*) - Y_{\mathbf{S}(\mathbf{u}'), X(\mathbf{u}')}(\mathbf{u}^*) \right], \tag{90}$$

with Shapley weights $w_{\mathbf{S}} > 0$. Each outcome here is f_Y at one input, and that input does not depend on the values of f_Y, so $\phi_X^{L_3} = \sum_i \kappa_i f_Y(i)$ is a linear function of the values of f_Y.

We show $\kappa_{i^*} > 0$. Fix $\mathbf{S}$ and $\mathbf{u}'$, and suppose $Y_{\mathbf{S}(\mathbf{u}')}(\mathbf{u}^*)$ does not reach i^*. Then some parent of Y differs from its value in $\mathbf{u}^*$, and tracing back along differing variables yields a directed path to Y from some $W \in \mathbf{S}$ with $W(\mathbf{u}') \neq W(\mathbf{u}^*)$, with no other variable of $\mathbf{S}$ on the path. If X is not on this path, then setting X as well keeps each variable on the path different from its value in $\mathbf{u}^*$, because its input still differs. If X is on this path, then $X(\mathbf{u}') \neq X(\mathbf{u}^*)$, since equality would force equal values of the ancestor W; so setting $X = X(\mathbf{u}')$ leaves a differing path from X to Y. Either way, $Y_{\mathbf{S}(\mathbf{u}'), X(\mathbf{u}')}(\mathbf{u}^*)$ does not reach i^*. So each bracket contributes a nonnegative amount to κ_{i^*}. Since $\equiv$ in Def. 2.4 is over $B \sim P_B$ and $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$, the premise of D_2 yields a value $x' = X(\mathbf{u}^\circ)$ with $P(\mathbf{u}^\circ) > 0$ and $Y_{x'}(\mathbf{u}^*) \neq y$, so $\mathbf{S} = \varnothing$ and $\mathbf{u}' = \mathbf{u}^\circ$ contribute $w_\varnothing P(\mathbf{u}^\circ) > 0$.

Hence $\{\phi_X^{L_3} = 0\}$ is a hyperplane in the values of f_Y, which has Lebesgue measure zero. Because $\mathrm{Dom}(\mathbf{U})$ is finite, there are finitely many choices of Y, $\mathbf{X}$, X, $\mathbf{E}$, and of inputs fixing the evidence, so the union of these hyperplanes has measure zero, also when $P(\mathbf{U})$ varies, by Fubini's theorem. □

D.5.3 Proof of D_4 (Normality)

Lemma D.3 (Causal Normality)**.** *L_3 SVs satisfy causal normality under functional dependence c.* □

Proof. Consider two SCMs satisfying the D_4 premise (Def. 2.4):

- The two SCMs have the same $\mathbf{U}$ and the same $P(\mathbf{U} \mid \mathbf{e})$, and at each unit $\mathbf{u}$ consistent with the evidence, counterfactual effects are identical: $c_1(z) := c(\mathcal{M}_1, \mathbf{u}, X_1, Y, z) = c(\mathcal{M}_2, \mathbf{u}, X_2, Y, z) =: c_2(z)$ for all baselines z.
- Signed dominance holds at each such unit for all baselines: unfolding $P_+^{\mathcal{M}}$, the premise $P_+^{\mathcal{M}_1}(b) \geq P_+^{\mathcal{M}_2}(b)$ is $\text{sign}(c(b))P_1(b) \geq \text{sign}(c(b))P_2(b)$, where $P_i(b) := P^{\mathcal{M}_i}(b)$ is the unconditional baseline measure.
- The signed baseline measures differ over both $B \sim P_B$ and $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$: $P_+^{\mathcal{M}_1}(B) \not\equiv P_+^{\mathcal{M}_2}(B)$.

Step 0: Strict signed-dominance at some b_0. Since the two SCMs have the same $P(\mathbf{U} \mid \mathbf{e})$ and signed dominance holds at each unit, the third condition holds iff, at each unit of a set of units with positive probability under $P(\mathbf{U} \mid \mathbf{e})$, $P_+^{\mathcal{M}_1}(b) \neq P_+^{\mathcal{M}_2}(b)$ at some b. If at almost every unit under $P(\mathbf{U} \mid \mathbf{e})$ signed dominance held with equality at every baseline b ($\text{sign}(c(b))(P_1(b) - P_2(b)) = 0$ for all b), then $P_1(b) = P_2(b)$ at each b with $c(b) \neq 0$, forcing the signed baseline measures to coincide at almost every unit and contradicting the third condition. Hence, at each unit of a set of units with positive probability under $P(\mathbf{U} \mid \mathbf{e})$, strict signed-dominance inequality holds at some b_0 in the support of P_B with $c(b_0) \neq 0$. The signed form handles negative effects: when $c(b) < 0$, world 1 has *lower* baseline probability, which still contributes positively to the difference.

The L_3 Shapley value for variable X is:

$$\phi_X^{L_3}(w) = \mathbb{E}_{\substack{\pi \sim \text{Unif}(\Pi_{\mathbf{X}}) \\ \mathbf{u}' \sim P_{\mathcal{M}}(\mathbf{U}) \\ \mathbf{u} \sim P_{\mathcal{M}}(\mathbf{U}|\mathbf{e})}} \left[\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u})\right] \tag{91}$$

By the NTE-functional equivalence (Sec. D.2), each marginal contribution equals the functional dependence value:

$$\text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = c(\mathcal{M}, \mathbf{u}, X, Y, z) \tag{92}$$

where $z = (X(\mathbf{u}'), \mathbf{Z}(\mathbf{u}'))$ is the baseline induced by $\mathbf{u}'$.

The Shapley value decomposes into a weighted sum over coalitions:

$$\phi_X^{L_3}(w) = \sum_{\mathbf{S} \subseteq \mathbf{X} \setminus \{X\}} w_{\mathbf{S}} \cdot \Delta_{\mathbf{S}} \tag{93}$$

where $w_{\mathbf{S}} = \frac{|\mathbf{S}|!(|\mathbf{X}|-|\mathbf{S}|-1)!}{|\mathbf{X}|!}$ are Shapley weights (identical in both worlds), and $\Delta_{\mathbf{S}}$ is the expected marginal contribution for coalition $\mathbf{S}$:

$$\Delta_{\mathbf{S}} = \mathbb{E}_{\mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U}|\mathbf{e})}[c(\mathcal{M}, \mathbf{u}, X, Y, z)] \tag{94}$$

where $z = (X(\mathbf{u}'), \mathbf{S}(\mathbf{u}'))$ is the baseline induced by $\mathbf{u}'$.

Step 1: Fixed-coalition comparison. Since $\mathbf{u}'$ is sampled from $P(\mathbf{U})$ independently of $\mathbf{u}$ (Def. 3.2), the baseline distribution $P_{\mathbf{S}}(z) := P_{\mathcal{M}}(\mathbf{u}' : (X(\mathbf{u}'), \mathbf{S}(\mathbf{u}')) = z)$ does not depend on $\mathbf{u}$. The marginal contribution expectation decomposes as:

$$\Delta_{\mathbf{S}} = \sum_z P_{\mathbf{S}}(z) \cdot \mathbb{E}_{\mathbf{u} \sim P(\mathbf{U}|\mathbf{e})}[c(\mathcal{M}, \mathbf{u}, X, Y, z)]. \tag{95}$$

Since the two SCMs have the same $P(\mathbf{U} \mid \mathbf{e})$ and identical counterfactual effects, write $c(z) := c_1(z) = c_2(z)$ at each $\mathbf{u}$:

$$\Delta_{\mathbf{S},1} - \Delta_{\mathbf{S},2} = \mathbb{E}_{\mathbf{u} \sim P(\mathbf{U}|\mathbf{e})}\left[\sum_z c(z) \cdot (P_{\mathbf{S},1}(z) - P_{\mathbf{S},2}(z))\right]. \tag{96}$$

Signed dominance gives $\text{sign}(c(z))P_1(z) \geq \text{sign}(c(z))P_2(z)$ for all z at each $\mathbf{u}$, with strict inequality at some b_0 at each unit of a set of units with positive probability under $P(\mathbf{U} \mid \mathbf{e})$ (Step 0). For coalition $\mathbf{S}$,

the marginal $P_{\mathbf{S},i}(z)$ is the marginalization of $P_i(b)$ to baselines with $\mathbf{Z}'(b) = \mathbf{S}$, so signed dominance lifts: $\text{sign}(c(z))P_{\mathbf{S},1}(z) \geq \text{sign}(c(z))P_{\mathbf{S},2}(z)$ for all z.

Case 1: Positive effects. When $c(z) > 0$, the condition simplifies to $P_{\mathbf{S},1}(z) \geq P_{\mathbf{S},2}(z)$, so $c(z)(P_{\mathbf{S},1}(z) - P_{\mathbf{S},2}(z)) \geq 0$.

Case 2: Negative effects. When $c(z) < 0$, the signed condition gives $P_{\mathbf{S},1}(z) \leq P_{\mathbf{S},2}(z)$; multiplying by negative $c(z)$ reverses the inequality, again giving $c(z)(P_{\mathbf{S},1}(z) - P_{\mathbf{S},2}(z)) \geq 0$.

Combining both cases at each $\mathbf{u}$ and taking the expectation over $\mathbf{u}$: $\Delta_{\mathbf{S},1} \geq \Delta_{\mathbf{S},2}$ for any coalition $\mathbf{S}$.

Step 2: Aggregation over coalitions. Let $b_0 = (x_0', \mathbf{z}_0')$ be, at each unit from Step 0, a baseline of positive P_B-measure where signed dominance is strict, and set $\mathbf{S}_0 := \mathbf{Z}_0'$ (the context set indexing b_0); since $\mathbf{X}$ has finitely many coalitions, $\mathbf{S}_0$ takes one value on a set of these units with positive probability under $P(\mathbf{U} \mid \mathbf{e})$, and we fix that value. Then at each unit of that set the strict-inequality term $c(b_0)(P_{\mathbf{S}_0,1}(b_0) - P_{\mathbf{S}_0,2}(b_0)) > 0$ makes the sum over z in Step 1 positive; every other unit contributes a nonnegative amount, so $\Delta_{\mathbf{S}_0,1} - \Delta_{\mathbf{S}_0,2} > 0$. Since all Shapley weights $w_{\mathbf{S}} > 0$ and all other terms are non-negative (from Step 1):

$$\phi^{L_3}_{X,1}(w) - \phi^{L_3}_{X,2}(w) = \sum_{\mathbf{S}} w_{\mathbf{S}}(\Delta_{\mathbf{S},1} - \Delta_{\mathbf{S},2}) \geq w_{\mathbf{S}_0}(\Delta_{\mathbf{S}_0,1} - \Delta_{\mathbf{S}_0,2}) > 0 \tag{97}$$

Therefore, L_3 SVs satisfy causal normality. □

D.5.4 Proof of D_3 (Effect Scaling)

Lemma D.4 (Causal Effect Scaling)**.** *L_3 SVs satisfy effect scaling under functional dependence c.* □

Proof. Consider two SCMs satisfying the D_3 premise (Def. 2.4):

- Pointwise effect ordering: the two SCMs have the same $\mathbf{U}$ and the same $P(\mathbf{U} \mid \mathbf{e})$, and at each unit $\mathbf{u}$ consistent with the evidence, $c_1(z) := c(\mathcal{M}_1, \mathbf{u}, X_1, Y, z) \geq c(\mathcal{M}_2, \mathbf{u}, X_2, Y, z) =: c_2(z)$ for all baselines z.
- Identical baseline distributions: $P_1(z) = P_2(z)$ for all z.
- Effect-distribution inequivalence over both $B \sim P_B$ and $\mathbf{u} \sim P(\mathbf{U} \mid \mathbf{e})$: $P^{\mathcal{M}_1}(c_B \mid \mathbf{z}'_\varnothing) \not\equiv P^{\mathcal{M}_2}(c_B \mid \mathbf{z}'_\varnothing)$.

With identical baseline distributions and the same $P(\mathbf{U} \mid \mathbf{e})$, the third premise forces $c_1 > c_2$ strictly on a set of unit-baseline pairs with positive probability under $P(\mathbf{U} \mid \mathbf{e})$ and P_B: if $c_1 = c_2$ almost everywhere under these distributions, the distributions of c_B given the but-for baseline $\mathbf{z}'_\varnothing$, which P_B draws with positive probability, would coincide (pushforwards under equal measures), contradicting the third premise.

By the NTE-functional equivalence (Sec. D.2), each marginal contribution equals the functional dependence value:

$$\text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = c(\mathcal{M}, \mathbf{u}, X, Y, z) \tag{98}$$

where $z = (X(\mathbf{u}'), \mathbf{Z}(\mathbf{u}'))$ is the baseline induced by $\mathbf{u}'$.

The L_3 Shapley value aggregates these functional dependence values:

$$\phi^{L_3}_X(w) = \mathbb{E}_{\substack{\pi \sim \text{Unif}(\Pi_{\mathbf{X}}) \\ \mathbf{u}' \sim P_{\mathcal{M}}(\mathbf{U}) \\ \mathbf{u} \sim P_{\mathcal{M}}(\mathbf{U}|\mathbf{e})}} [c(\mathcal{M}, \mathbf{u}, X, Y, z(\pi, \mathbf{u}'))] \tag{99}$$

Since the baseline distributions and $P(\mathbf{U} \mid \mathbf{e})$ are identical, the expectations have the same weights. The pointwise ordering carries to the expectation:

$$\phi^{L_3}_{X_1}(w) = \mathbb{E}_{\mathbf{u},\pi,\mathbf{u}'}[c_1(z(\pi, \mathbf{u}'))] > \mathbb{E}_{\mathbf{u},\pi,\mathbf{u}'}[c_2(z(\pi, \mathbf{u}'))] = \phi^{L_3}_{X_2}(w) \tag{100}$$

where the strict inequality follows because at each $\mathbf{u}$, $c_1 \geq c_2$ pointwise and $c_1 > c_2$ on a set of unit-baseline pairs with positive probability under $P(\mathbf{U} \mid \mathbf{e})$ and P_B, so the expectation is strict.

Therefore, L_3 SVs satisfy effect scaling. □

D.5.5 Main Result

Theorem 3.3 (L_3 SVs satisfy explanatory desiderata)**.** *L_3 SVs satisfy $D_{1,3,4}$ for all SCMs, and D_2 for all but a measure-zero set of structural parameters (Sec. A.9).* □

Proof. The result follows directly from Lems. D.1 to D.4. □

D.6 Proof of Shapley Axioms

Corollary A.18 (L_3 Shapley axioms)**.** *L_3 SVs satisfy Shapley axioms for value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{e})$:*

(1) ***Null player****: If $f(\mathbf{S} \cup \{X\}) = f(\mathbf{S})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$, then $\phi_X^{L_3}(w) = 0$*

(2) ***Efficiency****: $\sum_{X \in \mathbf{X}} \phi_X^{L_3}(w) = f(\mathbf{X}) - f(\emptyset)$*

(3) ***Symmetry****: If $f(\mathbf{S} \cup \{X_i\}) = f(\mathbf{S} \cup \{X_j\})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$, then $\phi_{X_i}^{L_3}(w) = \phi_{X_j}^{L_3}(w)$*

(4) ***Marginality****: If $f(\mathbf{S} \cup \{X\}) - f(\mathbf{S}) = f'(\mathbf{S} \cup \{X\}) - f'(\mathbf{S})\ \forall \mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$, then $\phi_X^{L_3}(w) = \phi_X'^{L_3}(w)$ where ϕ' is computed using f'* □

Proof. We verify each axiom directly for the L_3 Shapley value definition.

Null player: This follows directly from the Shapley value definition [32]. If $f(\mathbf{S} \cup \{X\}) = f(\mathbf{S})$ for all $\mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$, then all marginal contributions are zero, so $\phi_X^{L_3}(w) = 0$.

Efficiency: By definition of the Shapley value:

$$\sum_{X \in \mathbf{X}} \phi_X^{L_3}(w) = \sum_{X \in \mathbf{X}} \mathbb{E}_\pi[\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u})] \tag{101}$$

$$= \mathbb{E}_\pi\left[\sum_{X \in \mathbf{X}} (\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u}))\right] \tag{102}$$

$$= \mathbb{E}_\pi[\text{NTE}(\mathbf{X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\emptyset, Y|\mathbf{u}' \to \mathbf{u})] \quad \text{(telescoping sum)} \tag{103}$$

$$= \text{NTE}(\mathbf{X}, Y|\mathbf{e}) - \text{NTE}(\emptyset, Y|\mathbf{e}) \tag{104}$$

where the third equality follows from telescoping the sum over all variables, and the fourth takes the expectation over $\mathbf{u} \sim P(\mathbf{U}|\mathbf{e})$ and $\mathbf{u}' \sim P(\mathbf{U})$ to obtain the evidence-level NTE.

Symmetry: If variables X_i and X_j have identical marginal contributions for all coalitions $\mathbf{S} \subseteq \mathbf{X} \setminus \{X_i, X_j\}$:

$$\text{NTE}(\mathbf{S} \cup \{X_i\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{S}, Y|\mathbf{u}' \to \mathbf{u}) = \text{NTE}(\mathbf{S} \cup \{X_j\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{S}, Y|\mathbf{u}' \to \mathbf{u}) \tag{105}$$

then by the symmetry of the Shapley value formula, $\phi_{X_i}^{L_3}(w) = \phi_{X_j}^{L_3}(w)$.

Marginality: The L_3 Shapley value depends only on the marginal contributions $\text{NTE}(\mathbf{S} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{S}, Y|\mathbf{u}' \to \mathbf{u})$ for all coalitions $\mathbf{S} \subseteq \mathbf{X} \setminus \{X\}$. If two value functions f_1 and f_2 have identical marginal contributions for all coalitions, then $\phi_X^{L_3}(f_1) = \phi_X^{L_3}(f_2)$.

Since L_3 SVs satisfy all four axioms, they are the unique Shapley value for the NTE value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|\mathbf{e})$. □

E Extended Related Work

Our work focuses on local attribution methods that explain specific observed instances (e.g., "Why did Alice receive $10M?"), distinct from global importance methods [55, 56] and counterfactual recourse methods [57–61] that target actionable interventions rather than retrospective attribution.

Feature Attribution Methods. Following the Pearl Causal Hierarchy, prior methods divide into observational (L_1), interventional (L_2), and counterfactual (L_3) methods. SHAP [6] and related approaches [17, 19] compute additive feature attributions from observational distributions; LIME [7] fits local linear surrogates. Asymmetric Shapley values (ASV) [20] incorporate causal graph structure

into the Shapley computation while operating on L_1 distributions. Interventional Shapley values (also called causal Shapley values or do-SHAP) [22, 21, 23] operate on L_2 distributions, capturing population-level causal effects.

Actual Causation. AC frameworks formalize causation via counterfactual queries on structural-equation modifications, returning binary cause/non-cause verdicts rather than numerical attributions. The literature [15, 41] offers competing definitions, but no definition has been proven correct [44], and counterexamples exist for all current definitions [43]. Following Lewis [11]'s separation of explanation from causation, we develop a framework agnostic to the precise definition of causation.

SCM-Based Explanation Frameworks. Explanation frameworks based on actual causation [26, 16] identify causes but do not, by design, assign unique numeric attributions to individual variables. Probabilities of causation [13, 24] compute necessity/sufficiency probabilities for individual events. Closest to our work, CF-Shapley [25] pioneers unit-level counterfactual attributions by combining Shapley values with SCM counterfactuals under Markovian assumptions with additive errors.

Comparison of these methods against our desiderata appears in Table 1 and App. B.

F Frequently Asked Questions

We address questions that commonly arise when reading the main text, ordered by how often they came up in review.

Why is the live demo (X_2) not a cause of Alice's funding? The argument hinges on abduction from the observed evidence. Alice's evidence includes $X_3 = 1$ (lights flickered). Since $X_3 = U_3$ (Eq. (1)), observing $X_3 = 1$ infers $U_3 = 1$ by abduction. Alice also secures both funding sources ($X_4 = X_5 = 1$), which requires $U_j^3 = 1$ (abduction gives $U_j^{2,3} = 1$ for $j \in \{4, 5\}$). Under these abduced values, the demo term in Eq. (2) becomes $(X_2' \oplus U_3) \vee U_j^2 = (X_2' \oplus 1) \vee 1 = 1$ for any counterfactual value X_2', rendering X_j independent of X_2'. Therefore $c(\mathbf{z}', x_2') = 0$ for all baselines: changing the demo would not change the outcome. The key insight is that unit-level causation requires conditioning on all observed evidence, not just (X_2, Y). Methods that condition on partial evidence (e.g., PN without full $\mathbf{x}$) will incorrectly infer $U_3 = 0$ and attribute causation to X_2.

Is L_3 SV the unique method satisfying all four desiderata? No; $D_{1:4}$ do not uniquely determine a method. They are *necessary* conditions that any good EVA should satisfy. L_3 SVs are the first method shown to satisfy all four, but other value functions plugged into Shapley aggregation could also satisfy them if they correctly capture unit-level causal effects with probability-weighted baselines. The desiderata eliminate all existing $L_{1:2}$ methods (by Thm. 2.5) and impose ordering constraints on L_3 methods, but leave room for alternative formulations. We conjecture that L_3 SVs are the canonical choice given the additional Shapley axioms (efficiency, symmetry, marginality; Corollary A.18), but uniqueness under those joint constraints is not proven here.

Does functional dependence falsely identify non-causes? Functional dependence is a *necessary* condition for causation (see Sec. A.1.1), not sufficient: a variable can exhibit functional dependence under some contingency without being an actual cause (the prisoner example, Ex. A.2). Our framework handles this via the conservative-proxy assumption (Assump. A.5): we use but-for causation as the primary sufficient test (D_1 and D_2), with functional dependence serving as the baseline-sensitivity measure for $D_{3:4}$. A non-cause variable passes the but-for test only when $c(\varnothing, x') \neq 0$, which requires a counterfactual change under the null baseline ($\mathbf{Z} = \varnothing$). In Alice's example, X_2 fails the but-for test for all x_2' given the observed evidence, so D_1 correctly assigns $\phi_2 = 0$ even though X_2 has functional dependence under the population-level contingency $U_3 = 0$.

Can L_3 SVs be computed without a known causal diagram? Not exactly, but partial identification makes them computable in practice. Without a causal diagram, L_3 SVs are not point-identified because the counterfactual distribution $P(\mathbf{U}|\mathbf{e})$ is not recoverable from $L_{1:2}$ data alone (Thm. 2.5). Given a causal diagram and $L_{1:2}$ data, Alg. 1 computes tight bounds $[\phi^L, \phi^U]$ on each L_3 SV via the NTE bounding procedure (Thm. 4.2). The causal diagram encodes qualitative structural assumptions (which variables affect which), which are often available from domain knowledge or can be partially

inferred from data. We view the diagram as the appropriate level of prior knowledge for counterfactual explanation: weaker than specifying the full SCM, stronger than pure data.

What do the expressivity and compactness assumptions in Assump. 4.1 mean? On expressivity, this is the L_3-$\mathcal{G}$ expressivity condition of Xia et al. [36, Thm. 2], proved for feedforward NCMs over discrete variables. There is no general test of expressivity for an arbitrary neural architecture, which is why we work in classes where the property is established.

Expressivity describes which attributions the model class can match from a true SCM. Compactness of the model class (which holds for, for instance, models with discrete-valued variables whose exogenous variables can be represented on simplices, such as the CTM) describes whether the endpoints of the extreme attributions can be achieved exactly (i.e., whether the min and max of the quantity are defined). Should it be violated, we can still guarantee that a learned model that minimizes/maximizes the query comes arbitrarily close to the infimum and supremum of the quantity of interest (L_3 Shapley values), subject to consistency with diagram and observed data distributions.

When are L_3 SVs identifiable? In even the simplest cases with only one explanatory variable, one directed edge, and no unobserved confounders, such as in Ex. 1.1, L_3 SVs are not identifiable from data and a causal diagram. This follows from the bounds on the probability of necessity introduced in Tian and Pearl [54], a quantity we connect to the NTE in settings with one binary explanatory variable and one outcome variable (see Sec. A.6). However, under strong functional and structural assumptions, identifiability is possible. For example, as discussed in Sec. 3, if we assume that the outcome is deterministic with respect to the explanatory variables, *and* that there is no causation among explanatory variables, we can prove that L_3 SVs are equivalent to L_2 SVs and therefore identifiable from interventional data. If we additionally assume that there is no unobserved confounding between the explanatory variables, L_3 SVs are also equivalent to L_1 SVs and therefore identifiable from observational data. These conditions may be viewed as sufficiently strong causal assumptions under which even L_1 and L_2 SVs satisfy the desiderata. Therefore, given that the assumption of a causal diagram is substantially weaker than these conditions, we submit that the bounding algorithm we propose should be used as the default approach to inferring L_3 SVs, in the presence of data but the absence of strong causal assumptions. Following Thm. 4.2, this algorithm theoretically allows a researcher to make the strongest inference possible on L_3 SVs from data and minimal causal assumptions.

How do L_3 SVs scale with many variables? Exact Shapley computation requires $O(2^{|\mathbf{X}|})$ NTE evaluations, which is exponential in the number of explanatory variables. This is a known limitation of all Shapley-based methods. In practice, the approximation methods developed for SHAP [6] (permutation sampling, kernel approximation) apply directly to L_3 SVs: the Shapley permutation formula in Def. 3.2 can be approximated by sampling m random permutations for $O(m|\mathbf{X}|)$ NTE evaluations. The number of i.i.d. samples required to obtain a PAC estimate for a single variable depends only on the desired error threshold and confidence, and on the variance of the NTE. Specifically, L_3 SVs (Def. 3.2) are an expectation over a difference of NTEs Δ. If σ^2 bounds the variance of each NTE over $(\mathbf{Z}, \mathbf{u}, \mathbf{u}')$, then Δ has variance at most $4\sigma^2$, and its average over n samples, $\hat{\phi}_X$, has variance at most $4\sigma^2/n$. By Chebyshev's inequality,

$$P(|\hat{\phi}_X - \phi_X| \geq \epsilon) \leq \frac{4\sigma^2}{n\epsilon^2}.$$

Therefore, as long as the NTE has finite variance over its parameters $(\mathbf{Z}, \mathbf{u}, \mathbf{u}')$, $n \geq \frac{4\sigma^2}{\delta\epsilon^2}$ samples suffice to estimate the attribution within error ϵ with probability $1 - \delta$, with no further dependence on the number of explanatory variables $|\mathbf{X}|$. For typical XAI use cases with $|\mathbf{X}| \leq 20$, exact computation is tractable; for larger feature sets, kernel SHAP approximations transfer to the NTE value function without modification.

Does the theory apply to continuous variables? Yes. The framework is formulated for arbitrary variable domains (Defs. 2.2, 2.3, 3.1 and 3.2): $\mathrm{Dom}(X)$ can be any measurable space. Functional dependence, NTEs, and L_3 SVs are defined for continuous X and Y without modification. The experiments use categorical variables for ease of illustration, but Sec. A.6 derives the PN connection for general binary X, Y, and the bounds algorithm (Alg. 1) applies to discrete domains. Extending the bounds to continuous domains via a discretization or kernel approach is left to future work.

How do L_3 SVs differ from CF-Shapley [25]? CF-Shapley also applies Shapley values to a counterfactual value function, but uses a fixed-baseline approach: it evaluates $Y_{x'}$ at a single fixed counterfactual $x' = 0$ (or a fixed reference point). This design violates D_4 (Normality): a cause whose baseline is highly probable and one whose baseline is rare receive the same attribution if their effects are equal. More fundamentally, CF-Shapley assumes additive noise and no unobserved confounders, which restricts it to a subclass of SCMs. L_3 SVs average over *all* baselines weighted by $P(\mathbf{U})$, handle arbitrary SCMs, and satisfy all four desiderata. In the special case with no confounders, additive noise, and a deterministic outcome function, L_3 SVs reduce to an interventional Shapley value (Prop. A.19), which coincides with CF-Shapley under CF-Shapley's own assumptions.

G Broader Impacts

Our causal desiderata offer a principled framework for theoretically evaluating explanation methods, as well as a negative result showing that data-based methods cannot satisfy our desiderata without further causal information. We hope this work will be adopted by the community in the future evaluation of explanation methods and impact the future development of explanation methods. Potential negative impacts of our work derive from limitations discussed in Sec. 6: an incorrect causal diagram can induce misleading attributions in high-stakes deployment settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction enumerate three contributions (desiderata framework, counterfactual Shapley values, bounding algorithm) that correspond directly to Secs. 2 to 5. Scope limitations (known causal diagram required, semi-synthetic experiments) are stated in Sec. 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Sec. 6 contains an explicit Limitations paragraph covering the requirement for a known causal diagram, the functional-dependence assumption, optimization-dependent bound tightness, and the use of semi-synthetic data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems state their assumptions in the theorem body. Proof sketches appear in the main text; full proofs are in App. D. All results are numbered and cross-referenced via `cleveref`.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: App. C provides full SCM specifications, classifier architectures and hyperparameters, sampling procedures, and runtimes for all three experiments (startup, Color MNIST, CelebA-HQ). Both datasets (MNIST, CelebA-HQ) are publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is available upon request. All datasets used (MNIST, CelebA, CelebA-HQ, Color MNIST) are publicly available, and App. C provides full implementation details sufficient for reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. C reports optimizer (Adam), learning rates, batch sizes, epoch counts, LeNet architecture details, and train/validation splits for the classifier-training experiments. Seed counts (10) and optimizer-restart schedule are reported for the CTM bound computation in the startup experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: For the Color MNIST experiment, oracle Shapley value estimates report 95% bootstrap confidence intervals (B=200 resamples; Sec. C.1). Bound results are reported as binary outcomes (bounds include/exclude zero) over fixed sample sets, which are deterministic given the restricted function space; no variance estimate is required. CelebA-HQ oracle Shapley values are means over 100 samples per stratum; 95% bootstrap CIs (B=200 resamples) have width ≤ 0.004 across all cells and are reported in the caption of Table 9.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. C reports that all experiments ran on a single NVIDIA H100 80GB HBM3 GPU on an internal university cluster (host: 8×H100 80GB, 224 CPU cores, 2 TB system RAM) and gives per-experiment runtimes: startup bounds (~5 GPU-hours), Color MNIST pipeline (< 1 minute), CelebA-HQ (~1 hour). Total is under 7 GPU-hours on a single GPU; no multi-GPU or multi-node parallelism is required.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper presents a theoretical framework and algorithm with no human subjects, no data collection, and no harmful applications. We have reviewed the NeurIPS Code of Ethics and find no conflicts.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: App. G discusses positive impacts (the desiderata framework as a principled adjudicator for explanation methods, and the impossibility result constraining future method design); the negative risk (misplaced confidence from an incorrect causal diagram, which the framework cannot detect) derives from the limitations in Sec. 6.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no models, datasets, or other assets with high misuse risk. The algorithm operates on user-supplied causal diagrams and data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models are cited: MNIST [38], CelebA [62] and CelebA-HQ [63] (CC BY-NC 4.0; used for non-commercial research only), and Color MNIST [37]. No proprietary or restricted assets are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Code implementing the bounding algorithm is available upon request, with full documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing or human subjects. All experiments use synthetic or publicly available benchmark data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing assistance and code implementation support; they contributed no novel methodology, theoretical results, or experimental design.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.