Improving Causal Explanations

Anonymous authors
Paper under double-blind review

Abstract

Answering "Why did this outcome occur?" is central to the empirical science and explainable artificial intelligence (XAI). However, XAI lacks a principled framework for evaluating explanation methods. Following Lewis's view of explanations as summaries of causal histories distinct from causation itself, we formalize four desiderata agnostic to the precise definition of causation: causal admissibility (non-causes receive zero attribution), explanatory power (causes receive non-zero attribution), normality (attribution proportional to baseline normality), and effect responsivity (attribution proportional to effect magnitude). We establish conditions for inferring desiderata violations on unknown causation oracles from known causation proxies. We introduce counterfactual Shapley values (L_3 SVs), extending unit-level total effects with principled baseline selection. We prove L_3 SVs uniquely satisfy all desiderata under functional dependence as a causation proxy and provide a sound bounding algorithm. Experiments demonstrate that L_3 SVs are the first method to satisfy all desiderata on the proxy and correctly discriminate causal structures where existing methods fail.

1 Introduction

Scientific explanation has long relied on mathematical models, from Bacon's induction and Galileo's astronomical laws to Newton's laws of motion, epitomizing the Scientific Revolution of the 16th and 17th centuries. The resulting scientific method combines empirical data (e.g., planetary observations, object trajectories) with falsifiable mathematical models (e.g., Kepler's planetary laws, Newton's law of universal gravitation) to explain specific events (e.g., "Why did the apple fall on Newton's head?"). Modern causal inference adopts an analogous approach by using data and assumptions to construct structural causal models (SCMs) that serve as explanations of reality (**?**).

A similar challenge arises in explainable artificial intelligence (XAI), where growing demands for explainability are central to trust, autonomy, recourse, and debugging (**?**). Yet, unlike the natural sciences, where a model specification may serve as an explanation, XAI has yet to achieve a conceptual and technical definition of explanation that can be computed from data (**????**).

We address this gap by developing a causal explanation framework from first principles. We follow the view of **?** that "to explain an event is to provide some information about its causal history." Building on this insight, we argue that **an explanation of an event is a summary of its causes.**

To operationalize this definition, we characterize causal history using SCMs, which induce the Pearl Causal Hierarchy (PCH): observational (L_1), interventional (L_2), and counterfactual (L_3) distributions (**???**). However, the *Causal Hierarchy Theorem* (CHT) formalizes a fundamental challenge in that SCMs are almost never identifiable from data alone, as the observational distribution may fit multiple models that differ on interventional or counterfactual distributions (**?**, Thm. 1). Even if recoverable, general SCMs are often too complex for human interpretation. Given these challenges, we focus on a more tractable approach of event-level explanations that answer "why" questions about specific observed outcomes by attributing the outcome to individual variables, an approach we call Explanatory Variable Attribution (EVA). This variable-centric approach identifies contributing factors without requiring full model recovery, making it computationally feasible and interpretable while remaining compatible with XAI methods (**???**), actual-causation-based explanations (**??**), and probabilities of causation (**???**).

Following **?**'s insight that explanations must summarize causal histories, we face two fundamental questions: (1) What is a cause? and (2) How do we summarize causes well? The first question

remains unsolved, with causation definitions evolving repeatedly over decades without consensus or any way to prove correctness (**?????**), and recent work showing that even for recent definitions of causation, identical causal structures can yield different judgments depending on contextual framing (**?**). As **?** observes: "Whatever causation may be, there are still causal histories, and what I shall say about causal explanation should still apply." We sidestep this debate by assuming there exists an unknown causation oracle c^* and introducing desiderata given this oracle, then proving that violations on a *conservative proxy* can be used to infer violations on the oracle desiderata (**??**). To ground these concepts, we present a running example:

Example 1 (Startup pitch)**.** *Alice, a well-spoken entrepreneur ($X_1 = 1$), rehearses extensively ($X_2 = 1$) but arrives with red eyes ($X_3 = 1$) due to poor sleep. Extensive rehearsal normally helps but worsens performance for sleep-deprived entrepreneurs due to cognitive overload. Despite this, Alice secures Series A and angel funding ($X_4, X_5 = 1$) worth \$ 9M. She separately secures an AI grant ($X_6 = 1$) worth \$ 1M contingent on legal adult status ($X_7 = 1$), for $Y =$ \$10M total. Alice wonders, "Why did I receive funding?"*

She concludes it was because she was well-spoken ($X_1 = 1$), secured the grants ($X_4, X_5, X_6 = 1$), and was a legal adult ($X_7 = 1$), but not because of her rehearsal ($X_2 = 1$) or red eyes ($X_3 = 1$).

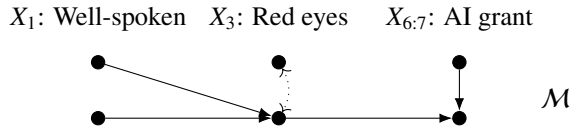
Alice's question is ambiguous (**?**): is she asking why funding versus no funding, why exactly \$10M, or why more than \$5M? We will resolve this by defining precise *why-queries* specifying foil values and explanatory variables.

Alice's intuition motivates four desiderata for explanatory attributions ϕ_i: (1) causal admissibility: non-causes get zero attribution ($\phi_{2:3} = 0$), (2) causal power: actual causes get non-zero attribution ($\phi_{1,4:7} \neq 0$), (3) causal normality: more abnormal causes get higher attribution ($\phi_6 > \phi_7$), and (4) causal effect scaling: larger effects get higher attribution ($\phi_4 > \phi_5$) (**??**).

We now develop the formal framework for explanatory variable attribution, beginning with the foundational questions of causation and summarization. Our contributions follow:

- **Explanatory Desiderata (??)**. We formalize Explanatory Variable Attributions (EVA, **??**) and four explanatory desiderata (**??**) grounded in causal explanation philosophy, assuming an unknown causation oracle exists (**??**). We employ functional dependence (**??**) as a conservative proxy (**??**) and prove the Desiderata Lifting Theorem (**??**) enabling proxy-to-oracle inference. We demonstrate all existing methods violate at least one desideratum (**??**).
- **Explanatory Method (??).** We introduce Natural Total Effects (NTE, **??**), extending probabilistic causation theory, and establish their connection to functional dependence (**??**). We define L_3 Shapley Values aggregating NTEs, proving satisfaction of Shapley axioms (**??**) and our desiderata (**??**).
- **Inferential Machinery (??, ??).** We prove the Explanatory Impossibility Theorem (**??**), which states that explanations satisfying the desiderata cannot be uniquely identified from observational and interventional data alone. We provide a sound algorithm computing information-theoretic bounds on L_3 SVs (**??**, **??**). We demonstrate empirical use cases on semi-synthetic data where our method distinguishes between classifiers when existing approaches cannot.

Preliminaries. Random variables are denoted by capital letters X, with values denoted by lowercase letters x. Sets of random variables $\mathbf{X}$ are bolded, and domains are denoted $\mathcal{D}_{\mathbf{X}}$. A structural causal model (SCM) (**??**) $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ 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 $\mathcal{G}$ (**??**) where $V_i \rightarrow V_j$ 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 $\mathbf{u}$ (**?**, Def. 7.1.8).



$$\mathcal{M} = \begin{cases} U_i & \sim \text{Bern}(0.1 + 0.8 \cdot \mathbf{1}[i = 7]) \\ X_i & = U_i, \quad \forall i \in \{1, 2, 3, 6, 7\} \\ X_j & = X_1 \wedge ((X_2 \oplus U_3) \vee U_j^1) \wedge U_j^2, \\ & \qquad \forall j \in \{4, 5\} \\ Y & = 8X_4 + X_5 + X_6 X_7 \end{cases}$$

Figure 1: "Why did Alice receive funding?" Causal diagram and SCM for the startup pitch example. Variables $X_{4:5}$ (pitch outcomes) and $X_{6:7}$ (AI grant requirements) are clustered for visual clarity. $\oplus$ indicates XOR: excessive rehearsal hurts those who are sleep deprived, who will usually have red eyes. All variables are observed to be 1.

2 Foundations for Explanatory Variable Attribution

We now formalize the explanatory variable attribution framework introduced in Section 1. Returning to Alice's startup pitch example (**??**), we can formalize her scenario with a causal diagram and structural causal model shown in **??**.

Alice's question requires formal specification through why-queries with two key components. Explanatory variables $\mathbf{X} \subseteq \mathbf{V}$ limit the scope to specific candidate causes, focusing the analysis rather than allowing any variable (e.g., "the Big Bang") as a valid explanation (**?**). Foil values specify outcome alternatives for comparison; different foils yield different explanations (e.g., exactly $Y = \$10M$ versus $Y > \$5M$ would make different variables causally relevant). To formalize these intuitions, we define the explanatory variable attribution:

Definition 1 (Explanatory Variable Attribution)**.** *An explanatory variable attribution (EVA) is a function $\phi : \Omega \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}^n$, where Ω is the space of SCMs over observed variables $\mathbf{V}$, $\mathbb{W}$ is the space of why-queries* $\text{Why}(y|\mathbf{x})$ *asking why $Y = y$ given explanatory variable setting $\mathbf{X} = \mathbf{x}$ for $\mathbf{X} \subseteq \mathbf{V}$, and $n = |\mathbf{X}|$ is the attribution dimension.* □

Alice's question "Why did I receive funding?" corresponds to the why-query $\text{Why}(y|x_{1:7})$. In this case, we consider Alice to be asking what factors contributed to her funding amount (continuous). When comparing to methods that only accept binary outcome explananda, we consider her to be asking why $\mathbf{1}[Y > 0]$ (why she received any funding).

We focus on EVAs as explanations for three reasons: (1) a good explanation must compress super-exponentially complex causal histories to a tractable output space (in this case, linear); (2) variable-based attribution approaches dominate existing literature across multiple research communities; and (3) variables correspond naturally to the atomic units of intervention in SCMs. EVA methods span observational approaches like SHAP (**?**) and LIME (**?**), which provide feature importance without causal assumptions but ignore causal structure. Interventional methods including Causal Shapley values (**???**) and Integrated Gradients (**?**) incorporate some causal reasoning through interventions but still diverge from formal causation definitions. The most sophisticated approaches use counterfactual reasoning, where CF-Shapley (**?**) employs counterfactuals but assumes additivity and no hidden confounding while using arbitrary baseline selection. Beyond XAI, the actual causation literature (**????**) and probabilities of causation (**??**) also focus on variable-level causal attributions, though typically for single variables rather than comprehensive explanations. This dominance of EVA approaches across diverse fields reflects their natural alignment with human reasoning about causation, where explanations typically identify which factors mattered for an outcome.

We introduce the *causal measure* as a generic interface to any causation definition, enabling systematic evaluation of explanation quality.

Definition 2 (Causal Measure)**.** *A causal measure c is a mapping $c : \Omega \times \mathcal{D}_{\mathbf{U}} \times \mathbf{V} \times \mathbf{V} \times \mathcal{Z} \to \mathbb{R}$, where Ω is the space of SCMs, $\mathcal{D}_{\mathbf{U}}$ is the domain of the exogenous variables in $\mathcal{M}$, $\mathbf{V}$ is the set of observed variables in $\mathcal{M}$, and $\mathcal{Z}$ is a baseline space equipped with probability measure $P^{\mathcal{M}}$, containing potential proofs of causation.* □

This abstraction enables systematic evaluation without committing to any causation definition, generalizing beyond binary judgments to capture effect magnitudes. Following **?**'s observation that

"Whatever causation may be, [...] what I shall say about causal explanation should still apply," we assume causation exists objectively despite its unknown definition.

Assumption 1 (Causation Oracle)**.** *There exists unknown causation oracle* $c^* \in \mathbb{C}$. □

This assumption allows us to reason about what properties good explanations should have, without defining oracle c^*. Specifically, by sidestepping actual causation debates, we can: (1) formalize desiderata that must hold for good explanations given a causal measure c we take to be the oracle c^* (**??**), and (2) infer violations (and in some cases, satisfactions) of oracle desiderata from known proxies c' under reasonable assumptions (**??**). Our approach will remain relevant as long as establishing correctness for causation definitions remains an open problem.

2.1 Formal Desiderata

Alice's intuitive reasoning about her startup success reflects four fundamental requirements for good explanations. As we'll see in **??**, existing attribution methods systematically violate these principles, producing explanations that conflict with Alice's clear causal understanding. This motivates our formal desiderata for explanatory variable attribution.

First, causal admissibility requires that non-causes like her unhelpful excessive rehearsal (X_2) and red eyes (X_3) should receive zero attribution. Second, causal power demands that actual causes such as being well-spoken (X_1) and securing funding (X_4, X_5, X_6, X_7) should receive non-zero attribution. Third, causal normality dictates that causes with more normal baselines should receive greater attribution in the direction of the effect sign: the specialized AI grant (X_6) deserves more credit than legal adult status (X_7) since not getting the grant is more likely than not being a legal adult. Fourth, causal effect scaling ensures that larger causal effects should receive higher attribution: Series A funding (X_4, \$8M) deserves a larger attribution than angel funding (X_5, \$1M).

These principles, grounded in **?**'s philosophical insights and **?**'s psychological observations, formalize intuitive requirements for good explanations.

Definition 3 (Explanatory Desiderata)**.** *We define four desiderata as mappings* $D_{1:4} : \Omega \times \mathbb{C} \times \Phi \to \{0, 1\}$ *where* Ω *is the space of SCMs,* $\mathbb{C}$ *is the space of causal measures, and* Φ *is the space of EVA methods: causal admissibility, causal power, causal normality, and causal effect scaling.*

Given SCM $\mathcal{M} \in \Omega$*, causation oracle* $c \in \mathbb{C}$*, and EVA method* $\phi \in \Phi$*, we say* $D_i(\mathcal{M}, c, \phi) = 1$ *when:*

$$
\begin{aligned}
D_1: &\quad c = 0 && \implies \phi = \mathbf{0} && \text{(Admissibility)}\\
D_2: &\quad \exists \mathbf{u}, z : c \neq 0 && \implies \phi \neq \mathbf{0} && \text{(Power)}\\
D_3: &\quad c_{\mathcal{M}} = c_{\mathcal{M}'} \ \wedge\ P^{\mathcal{M}}_{\bar{c}}(Z) \not\equiv P^{\mathcal{M}'}_{\bar{c}}(Z) \ \wedge\ P^{\mathcal{M}}_{+}(z) \geq P^{\mathcal{M}'}_{+}(z) && \implies \phi(\mathcal{M}) > \phi(\mathcal{M}') && \text{(Normality)}\\
D_4: &\quad c_{\mathcal{M}} \geq c_{\mathcal{M}'} \ \wedge\ P^{\mathcal{M}}(Z) \equiv P^{\mathcal{M}'}(Z) \ \wedge\ P^{\mathcal{M}}(c_Z) \not\equiv P^{\mathcal{M}'}(c_Z) && \implies \phi(\mathcal{M}) > \phi(\mathcal{M}') && \text{(Scaling)}
\end{aligned}
$$

where universal quantifiers in premises are omitted for clarity. Except for symbols marked with $\exists$*, premises hold for all SCMs* $\mathcal{M}$*, why-queries* w*, units* $\mathbf{u}$*, and baselines* z*. For readability:* c*,* $c_{\mathcal{M}}$ *abbreviate* $c(\mathcal{M}, \mathbf{u}, X, y, z)$*,* ϕ *abbreviates* $\phi_X(\mathcal{M}, w)$*,* $P^{\mathcal{M}}_{\bar{c}}(Z)$ *denotes* $P^{\mathcal{M}}(Z|c \neq 0)$*, and* $P^{\mathcal{M}}_{+}(z)$ *denotes* $\text{sign}(c_{\mathcal{M}})P^{\mathcal{M}}(z)$*. The random variable* Z *ranges over baseline space* $\mathcal{Z}$*, and* c_Z *is the random variable induced by* c *over distribution* $P(Z)$*.*

These desiderata provide objective evaluation criteria that formalize Alice's intuitive reasoning while grounding it in established philosophical principles. D_1 (Admissibility) formalizes Alice's recognition that rehearsal (X_2) and red eyes (X_3) deserve zero attribution, as explanations must provide "negative information about what the causal history does not include" and avoid "false propositions about the causal history" (**?**). D_2 (Power) captures Alice's identification of actual contributors (X_1, $X_{4:7}$), following Lewis's preference for "more correct explanatory information" over "true but weak proposition[s]" (**?**). D_3 (Normality) reflects Alice's insight that legal adult status (X_7) deserves less credit than the AI grant (X_6) due to its more normal baseline, aligning with explanations highlighting "the most remarkable part" of an event's causal history (**?**) and preferring causes that "could easily have been otherwise" (**?**). Unlike graded causation (**?**) which uses only the most normal baseline, D_3 considers all baselines weighted by probability, compatible with empirical observations of context-dependent normality preferences (**?**). D_4 (Scaling) captures that Series A (\$8M) deserves higher attribution than angel funding (\$1M), a need for effect proportionality recognized across

psychology, biostatistics, computer science, econometrics, epidemiology, and XAI (see discussion after **??**).

Having established formal desiderata for evaluating explanatory methods, we now address the fundamental challenge: we don't know the true causation oracle c^*.

2.2 Inference via Conservative Proxies

If we had c^*, we could directly check whether Alice's excessive rehearsal contributed to her funding and verify that methods satisfying D_1 yield $\phi_2 = 0$. Since we do not, our solution is a conservative proxy c' that captures necessary conditions for causation. Methods that fail under this proxy will fail under the true oracle.

Definition 4 (Functional Dependence Causation Proxy)**.** *Define functional dependence as causal measure $c' \in \mathbb{C}$:*

$$c'(\mathcal{M}, \mathbf{u}, X, Y, z) = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}', x'}(\mathbf{u}) \tag{1}$$

where baseline $z = (x', \mathbf{z}') \in \mathcal{Z} = \{(x', \mathbf{z}') : x' \in \mathcal{D}_X, \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{z}' \in \mathcal{D}_{\mathbf{Z}}\}$. □

This measures the difference in outcome when X changes from baseline x' to its actual value in contingent world $\mathbf{z}'$, matching counterfactual Shapley marginal contributions (**?**). For Alice, the AI grant's effect depends on baseline choice: with baseline $z = (x'_6 = 0, \mathbf{z}' = \{X_7 = 0\})$ the grant has no effect, but with $z = (x'_6 = 0, \mathbf{z}' = \{\})$ it contributes \$1M.

Assumption 2 (Conservative Proxy)**.** *Under **??**, there exists a mapping $f : \mathcal{Z}^* \to \mathcal{Z}$ such that $c^*(z^*) \neq 0 \implies c^*(z^*) = c'(f(z^*))$ and $c^* \neq 0 \Rightarrow c' \neq 0$.* □

This combines necessity (no causation theory classifies X as a cause without effect (**????**)) with effect preservation (causal effects are counterfactual differences (**?**)).

Theorem 1 (Desiderata Lifting)**.** *Under **??**: (a) $\neg D_1(c', \phi) \Rightarrow \neg D_1(c^*, \phi)$ (admissibility), (b) $D_2(c', \phi) \Rightarrow D_2(c^*, \phi)$ (power), (c) $\neg D_3(c', \phi) \Rightarrow \neg D_3(c^*, \phi)$ (normality), (d) $\neg D_4(c', \phi) \Rightarrow \neg D_4(c^*, \phi)$ (responsivity).* □

By **??**, violations under proxy c' enable EVA falsification: violations of admissibility, normality, or responsivity under c' guarantee violations under the true oracle c^*, while satisfaction of power under c' guarantees satisfaction under c^*. To ascertain power violations, we require counterexamples where we believe $c' = c^*$. We argue that our startup example provides such cases for variables X_1 (well-spoken) and $X_{4:7}$ (funding sources), where methods yielding $\phi = 0$ despite clear functional dependence violate power.

Evaluating existing methods under our functional dependence proxy reveals systematic violations (**??**) on several popular methods. Our analysis reveals a fundamental gap: no existing method satisfies all four desiderata (detailed counterexamples in **??**). CF-Shapley comes closest but fails power and normality due to arbitrary baseline selection that can eliminate causal effects and ignore baseline probability. In the next section, we show how a principled baseline selection derived directly from the functional dependence proxy yields counterfactual Shapley values that satisfy all desiderata. Like permutation feature importance (**?**), our approach uses the natural (observed) distribution as baseline, ensuring that attributions reflect realistic counterfactual scenarios rather than artificial reference points.

3 Counterfactual Shapley Values

Having established the theoretical framework for evaluating explanatory methods, we now develop a concrete method that satisfies all four desiderata $D_{1:4}$. Our approach builds on two key insights: first, that our functional dependence conservative proxy can be captured through a multivariate extension of the total effect (**??**) with principled baseline selection, which we will call the Natural Total Effect (NTE); second, that Shapley values (**?**) provide a principled decomposition of marginal contributions to multivariate effects.

Table 1: Desiderata violations under proxy c' that lift to oracle c^* via **??**. Left columns show attributions under c', right columns show lifted conclusions about c^*. Methods: SHAP (**?**), LIME (**?**), L_2 SVs (**?**), PN (**?**), CF-SHAP (**?**). Our method provably satisfies all desiderata (**??**). ✓* indicates not proven to violate on this example.

Method	X_1	X_2	X_3	X_4	X_5	X_6	X_7	D_1	D_2	D_3	D_4
Desiderata	D_2 $\neq 0$	D_1 $= 0$	D_1 $= 0$	D_2 $\neq 0$	D_2 $\neq 0$	D_2 $(> \phi_7)$	D_2 $(< \phi_6)$				
SHAP	0.19	−0.25	−0.02	5.23	3.13	0.93	0.03	✗	✓*	✓*	✗
LIME	0.00	0.00	0.00	7.97	1.00	0.80	0.00	✓*	✗	✓*	✓*
L_2 SVs	0.26	−0.40	0.00	7.48	0.94	0.93	0.03	✗	✓*	✓*	✓*
PN	0.81	0.13	0.00	0.29	0.29	0.89	0.30	✗	✓*	✓*	✗
CF-SHAP	4.50	0.00	0.00	4.00	0.50	0.50	0.50	✓*	✓*	✗	✓*
L_3 SVs (**??**)	2.74	0.00	0.00	4.88	0.66	0.38	0.62	✓	✓	✓	✓

3.1 Theoretical Framework

We begin by formalizing the Natural Total Effect, which extends classical total effects to handle multivariate causation while incorporating normality through principled baseline selection.

Definition 5 (Natural Total Effect)**.** *Given SCM $\mathcal{M}$, why-query $w = \text{Why}(y|\mathbf{x})$, and variable subset $\mathbf{Z} \subseteq \mathbf{X}$, the natural total effect (NTE) is defined for actual and baseline units $\mathbf{u}, \mathbf{u}'$ as:*

$$\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \tag{2}$$

This captures the difference in outcome when variables $\mathbf{Z}$ take their actual values versus baseline values, evaluated for unit $\mathbf{u}$. □

For a natural baseline, we select baseline $\mathbf{u}'$ from the natural distribution $P(\mathbf{U})$, capturing outcomes weighted by their baseline normality. In practice, we do not observe $\mathbf{u}$ and therefore we condition on observed values $\mathbf{X} = \mathbf{x}, Y = y$. In practice, this expectation is computed via Monte Carlo sampling given an SCM.

$$\text{NTE}(\mathbf{Z}, Y|w) = \mathbb{E}_{\mathbf{u}\sim P(\mathbf{U}|\mathbf{X}=\mathbf{x},Y=y),\mathbf{u}'\sim P(\mathbf{U})}[\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})] \tag{3}$$

The NTE extends total effects (**??**) through multivariate interventions, natural baseline selection from $P(\mathbf{U})$ incorporating normality weighting, and equivalence to functional dependence information.

Remark 2 (Connection to Functional Dependence)**.** *The functional dependence proxy c' (**??**) is inferrable from the marginal contribution of X to an* NTE*:*

$$c'(\mathcal{M}, \mathbf{u}, X, y, z) = \text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) \tag{4}$$

where $z = (x', \mathbf{z}')$ with $\mathbf{u}'$ inducing witness values $(x', \mathbf{z}')$. □

While NTEs capture our causal proxy, their exponential complexity in $|\mathbf{X}|$ makes them impractical for human comprehension. We address this by summarizing NTEs into a variable contributions using Shapley values (**?**).

Definition 6 (Counterfactual Shapley Values (L_3 SVs))**.** *Given SCM $\mathcal{M}$, why-query $w = \text{Why}(y|\mathbf{x})$ and variable $X \in \mathbf{X}$, the Counterfactual Shapley Value is defined as:*

$$\phi_{L_3,X}(w) = \mathbb{E}_{\substack{\pi\sim\text{Unif}(\Pi_\mathbf{X}) \\ \mathbf{u}'\sim P_\mathcal{M}(\mathbf{U}) \\ \mathbf{u}\sim P_\mathcal{M}(\mathbf{U}|\mathbf{V}=\mathbf{v})}}[\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u})] \tag{5}$$

where π is a uniformly random permutation of variables $\mathbf{X}$, $\Pi_\mathbf{X}$ denotes the set of all permutations of $\mathbf{X}$, $\pi_{<X}$ denotes the variables preceding X in permutation π, and $\pi_{\leq X} = \pi_{<X} \cup \{X\}$. □

L_3 SVs provide a principled decomposition of natural total effects by distributing the causal contribution of all variable subsets $\mathbf{Z} \subseteq \mathbf{X}$ to individual variables $X \in \mathbf{X}$. This approach maintains the theoretical guarantees of Shapley values while incorporating the normality and multivariate causation principles embedded in the NTE.

Algorithm 1 Bounding L_3 Shapley Values

1: Input: Dataset $\mathcal{D}$, causal diagram G, why-query w, variable X
2: Output: Bounds $[\phi^-_{L_3,X}(w), \phi^+_{L_3,X}(w)]$
3: Train ensemble of neural causal models $\{f_k\}_{k=1}^K$ consistent with G on $\mathcal{D}$
4: For each model f_k, compute $\phi_{L_3,X,k}(w)$ using **??**
5: Return $[\min_k \phi_{L_3,X,k}(w), \max_k \phi_{L_3,X,k}(w)]$

Corollary 3 (L_3 SVs satisfy Shapley axioms)**.** *L_3 SVs satisfy the fundamental Shapley axioms of efficiency, symmetry, and marginality when applied to the value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|w)$.* □

The key innovation introduced by this method is natural baseline selection $u' \sim P(\mathbf{U})$, which enables power and normality desiderata satisfaction. In contrast, the most similar method CF-SHAP (**?**) requires selection of a single baselines preventing satisfaction of $D_{2:3}$.

Theorem 4 (L_3 SVs satisfy explanatory desiderata)**.** *Under the functional dependence causation proxy c' (**??**), L_3 SVs satisfy: admissibility (D_1) for all SCMs, power (D_2) for a measure-1 subset of SCMs (see **??** for measure definition), normality (D_3) for all SCMs, and effect responsivity (D_4) for all SCMs.* □

This establishes L_3 SVs as the first method satisfying all fundamental explanation requirements.

3.2 Implementation and Computation

Computing L_3 SVs from data faces a fundamental challenge:

Theorem 5 (Explanatory Impossibility Theorem)**.** *No EVA can compute exact explanatory variable attributions satisfying desiderata $D_{1:4}$ from observational or interventional data alone.* □

This impossibility arises because desiderata satisfaction requires counterfactual relationships unidentifiable from data (**?**). We address this by employing causal diagrams G and neural causal models (**?**) to bound rather than exactly compute L_3 SVs.

Assumption 3 (L_3-G-expressivity)**.** *There exists a neural causal model class $\mathcal{F}_G$ such that for any SCM $\mathcal{M}$ consistent with causal diagram G, some $f \in \mathcal{F}_G$ represents the NTE quantities $\{\text{NTE}(\mathbf{Z}, Y|w) : \mathbf{Z} \subseteq \mathbf{X}\}$ with arbitrary precision.* □

We bound L_3 SVs by training neural models consistent with G and computing the range of values across this class:

$$\hat{\phi}_{L_3,X}(w) \approx \frac{1}{M}\sum_{m=1}^{M}\left[\text{NTE}(\pi^{(m)}_{\leq X}, Y|\mathbf{u}^{(m)}_j \to \mathbf{u}^{(m)}_i) - \text{NTE}(\pi^{(m)}_{<X}, Y|\mathbf{u}^{(m)}_j \to \mathbf{u}^{(m)}_i)\right] \tag{6}$$

where $\{\pi^{(m)}, \mathbf{u}^{(m)}_j, \mathbf{u}^{(m)}_i\}_{m=1}^M$ are Monte Carlo samples.

Theorem 6 (Soundness of bounding algorithm)**.** *Under **??**, the bounds produced by **??**($\mathcal{D}$, $\mathcal{G}$, w, X) contain the true L_3 Shapley Value: $\phi_{L_3,X}(w) \in [\phi^-_{L_3,X}(w), \phi^+_{L_3,X}(w)]$.* □

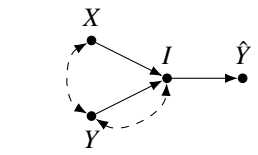
This soundness guarantee provides conservative bounds that reliably identify non-explanatory variables.

4 Case Study: Color MNIST

We demonstrate that counterfactual Shapley values (L_3 SVs) correctly identify non-causal variables (satisfying D_1) while existing methods erroneously assign them non-zero attribution in scenarios with known ground-truth causal structures.

4.1 Color MNIST

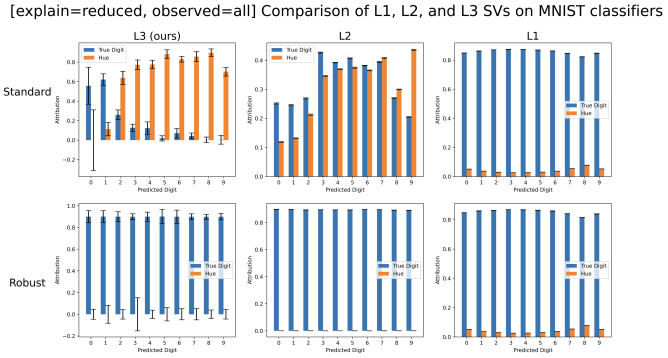
We evaluate our method on a semi-synthetic dataset where we know the ground-truth SCM, allowing us to assess attribution accuracy directly. We generate colored digit images based on the MNIST



(a) Causal diagram.

(b) Real digits.

(c) Shifted digits.



(d) Variable attributions for MNIST.

Figure 2: Color MNIST experiments. (a) Causal diagram. (b) Real digits. (c) Shifted digits. (d) Comparison of L_1, L_2, L_3 SVs on samples from the color MNIST dataset. Error bars denote bounds from interventional data for L_3 SVs; they denote estimation error for L_1, L_2 SVs.

dataset (**?**), where each image has two controlled dimensions: the hue of the digit X and the digit itself Y. The ground truth SCM follows:

$$P(\mathbf{U}) = \begin{cases} u_Y & \sim \text{Unif}(\{0, \ldots, 9\}) \\ u_X & \sim \text{Unif}(0, 1) \\ u_I^i & \sim \text{MNIST}(i) \end{cases}, \qquad \mathcal{F}_{\beta,\hat{f}} = \begin{cases} X & = \frac{u_Y}{9} + 0.5\Phi(u_X) + \beta \mod 1 \\ Y & = u_Y \\ I & = \text{hsv\_to\_rgb}\left(X, \frac{u_Y}{9}, u_I^{Y=y}\right) \\ \hat{Y} & = \hat{f}(I) \end{cases}. \tag{7}$$

Here, β represents a hue shift parameter, f represents an image classifier, MNIST(i) denotes an MNIST image containing the digit i selected uniformly at random, and hsv_to_rgb denotes the conversion of a hue, saturation, and value triplet to a 28×28 RGB image. The causal diagram for this SCM is shown in Fig. **??**.

By design of the MNIST SCM, hue and digit are strongly correlated: lower digits have lower saturation, higher digits have higher saturation, and zeros are entirely white. Thus, digit Y and image I are spuriously confounded. We train two basic convolutional digit classifiers $\hat{Y}$ on this dataset. The *standard* classifier $\hat{Y}^S$ is trained directly on the original data. The *robust* classifier $\hat{Y}^R$ is trained on the same architecture after preprocessing the images to grayscale, removing hue as a predictive feature. We expect for nonzero digits $\phi_X^R \approx 0$ and $\phi_X^S > 0$, while for zero digits $\phi_X^S \approx 0$ since white digits have no hue.

L_1 Shapley values (**?**) fail to distinguish the models (**??**, right), yielding similar attributions for both $\hat{Y}^S$ and $\hat{Y}^R$. L_2 Shapley values (**?**) incorrectly assign nonzero attribution to zero digits for $\hat{Y}^S$ (**??**, middle). Only L_3 counterfactual Shapley values match expectations (**??**, left): zero hue attribution for $\hat{Y}^R$ on all digits, positive attribution for $\hat{Y}^S$ on nonzero digits, and bounds containing zero for zero digits. Therefore, L_3 SVs satisfy causal admissibility where L_1 and L_2 SVs fail: distinguishing models with different behavior and correctly conditioning on observed information to identify irrelevant variables. This experiment demonstrates the practical utility of satisfying admissibility: L_3 SVs correctly identify when variables are non-causal (e.g., hue for grayscale-trained models), enabling practitioners to distinguish between models that use features versus those that ignore them, a distinction L_1 and L_2 SVs fail to make. See **??** for additional results.

5 Conclusions

We developed a principled framework for explanatory variable attribution grounded in Lewis's distinction between causation and causal explanation. We introduce four formal desiderata (admissibility, power, normality, and effect responsivity) as objective evaluation criteria for attribution methods given the true causal oracle. We prove the Desiderata Lifting Theorem, which enables rigorous

evaluation by translating conclusions about conservative causal proxies to claims about desiderata on the causation oracle.

We then apply this framework to introduce counterfactual Shapley values (L_3 SVs), which uniquely satisfy all desiderata among existing methods through their natural baseline selection. We prove the Explanatory Impossibility Theorem, demonstrating that observational and interventional distributions cannot fully determine L_3 SVs, motivating our sound bounding algorithm. Experiments on color MNIST validate that L_3 SVs discriminate genuine causal influence from spurious correlation where existing methods fail.

This framework addresses a fundamental limitation in explainable AI: existing attribution methods lack principled foundations for expressing desiderata on causal explanations. By explicitly modeling the distinction between causation and causal explanation, we provide both theoretical guarantees and practical tools for more reliable explanations.

Key directions for future work include exploring alternative causation proxies, extending to different data modalities, and applying the causation-explanation separation to other explanation modalities beyond variable attribution.

Appendices

A Theoretical Foundations

A.1 Measure on SCMs

We define a measure on SCMs to establish that D_2 (Power) holds for a measure-1 subset of SCMs. For any variable X, SCM $\mathcal{M}$, and evidence $\mathbf{v}$, define the **NTE basis**:

$$\mathcal{B}_X^{\text{NTE}}(\mathcal{M}, \mathbf{v}) = \{\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) : (\mathbf{Z}, \mathbf{u}', \mathbf{u}) \in I_{\mathcal{M}}\} \tag{8}$$

where $I_{\mathcal{M}} = \{(\mathbf{Z}, \mathbf{u}', \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{u}', \mathbf{u} \in \mathcal{D}_{\mathbf{U}}^{\mathcal{M}}, \mathbf{V}_{\mathcal{M}}(\mathbf{u}) = \mathbf{v}\}$. Since this basis is isomorphic to $\mathbb{R}^{|I_{\mathcal{M}}|}$, we define the **standard measure on SCMs** as the product of Lebesgue measures $\mu = \prod_{i \in I_{\mathcal{M}}} \lambda_i$ where each λ_i is the Lebesgue measure on $\mathbb{R}$. This product measure μ is atomless, meaning that any singleton set $\{\mathbf{b}\} \subset \mathbb{R}^{|I_{\mathcal{M}}|}$ has $\mu(\{\mathbf{b}\}) = 0$. Any L_3 Shapley value $\phi_{L_3,X}(w)$ can be viewed as a measurable function $\phi_{L_3,X} : \mathbb{R}^{|I_{\mathcal{M}}|} \to \mathbb{R}$ mapping NTE basis values to attribution values. The key insight is that $\phi_{L_3,X}(w) = 0$ despite functional dependence requires a pathological alignment where all marginal contributions in the Shapley average exactly cancel out, which occurs with probability zero under the atomless measure μ.

A.2 Normality and Human Judgment

? introduces a number of observed phenomena regarding normality and causal judgment. We show that we capture each of them with counterfactual Shapley values.

1. **Abnormal inflation:** under a conjunctive model ($Y = X_1 \wedge X_2$, and $X_1 = X_2 = 1$), if X_1 is less likely, it is viewed as a stronger cause.

2. **Abnormal deflation:** under a disjunctive model ($Y = X_1 \vee X_2$, and $X_1 = X_2 = 1$), if X_1 is more likely, it is viewed as a stronger cause.

3. **Conjunctive supersession:** under a conjunctive model, if X_2 is less likely, then X_1 is viewed as less of a cause than if X_2 was more likely.

4. **Non-supersession under disjunction:** X_2's probability does not affect how X_1 is viewed as a cause relative to X_2.

Abnormal inflation and conjunctive supersession. Abnormal inflation and conjunctive supersession are consistent with the claim that humans prefer more abnormal causes (**?**). We show that L_3 SVs capture both phenomena. Say X_1 and X_2 are independent, $Y = X_1 \wedge X_2$, and we observe $\mathbf{v} = \{X_1 = X_2 = Y = 1\}$. Then, under the standard why query $w = \text{Why}(y|\mathbf{x})$, we have L_3 SV:

$$\begin{aligned}
\phi_{X_1}^{L_3}(w) &= \frac{1}{2}(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{x_1'}|\mathbf{v}]) + \frac{1}{2}(\mathbb{E}[Y_{x_2'}|\mathbf{v}] - \mathbb{E}[Y_{x_1',x_2'}|\mathbf{v}]) && (9)\\
&= \frac{1}{2}(1 - P(X_1 = 1)) + \frac{1}{2}(P(X_2 = 1) - P(X_1 = 1, X_2 = 1)) && (10)\\
&= \frac{1}{2}(1 + P(X_2 = 1))(1 - P(X_1 = 1)) && (11)
\end{aligned}$$

We can see from the expression that when $P(X_1 = 1)$ decreases, we give X_1 a greater attribution (abnormal inflation). When $P(X_2 = 1)$ increases, we also give X_1 a greater attribution (conjunctive supersession).

Disjunctive non-supersession. We argue that our work also captures disjunctive non-supersession. Say X_1, X_2 are independent, $Y = X_1 \vee X_2$, and $\mathbf{v} = \{X_1 = X_2 = Y = 1\}$. Then we have:

$$\phi_{X_1}^{L_3}(w) = \phi_{X_2}^{L_3}(w) = \frac{1}{2}(1 - 1) + \frac{1}{2}(1 - (1 - P(X_1 = 0, X_2 = 0))) = \frac{1}{2}P(X_1 = 0, X_2 = 0) \tag{12}$$

This is consistent with non-supersession under disjunction: X_1, X_2 are equally strong causes.

Abnormal deflation. Finally, we argue that our work also captures abnormal deflation by using an alternative evidence set $\mathbf{e} = \{Y = y\}$.

As shown above, their L_3 SVs under our why query with evidence set $\mathbf{v}$ are identical, and they are equally strong causes. However, to break this tie, we can also consider which of X_1, X_2 is more likely to be a cause of the outcome if we only observe $Y = y$ – that is, which is more likely to be a cause in general, rather than in actuality.

Clearly, the more likely event of $X_1 = 1, X_2 = 1$ is also more likely to be a cause of the outcome. This can be captured using L_3 SVs on why query $w' = \text{Why}(y|\mathbf{e} = \{y\})$:

$$\phi^{L_3}_{X_1}(w') = \frac{1}{2}(\mathbb{E}[Y|y] - \mathbb{E}[Y_{x'_1}|y]) + \frac{1}{2}(\mathbb{E}[Y_{x'_2}|y] - \mathbb{E}[Y_{x'_1,x'_2}|y]) \tag{13}$$

which simplifies to $\phi^{L_3}_{X_1} = \frac{1}{2}(P(X_1 = 0, X_2 = 0) + \epsilon)$, where $\epsilon = \frac{P(X_1=0,X_2=0)}{1-P(X_1=0,X_2=0)}(P(X_1 = 1) - P(X_2 = 1))$.

We can see that when $P(X_1 = 1) > P(X_2 = 1)$, $\epsilon > 0$. Since $\phi^{L_3}_{X_2} = \frac{1}{2}(P(X_1 = 0, X_2 = 0) - \epsilon)$, we know that $\phi^{L_3}_{X_1} > \phi^{L_3}_{X_2}$ when X_1 is more normal than X_2. This is consistent with abnormal deflation, the intuition that it is preferable to communicate the more normal of two sufficient actual causes – not because it is more of a cause in the actual setting, but because it is more likely to be a cause in general and thus more informative.

Notably, the human experiments in **?** test whether humans believe X_1 alone or X_2 alone is a cause in the disjunctive setting. Their results do not distinguish between a preference for X_1 over X_2 due to stronger causation or due to greater utility in communication. We argue that here, humans prefer to communicate the more normal cause because it is more likely to be correct in general, not because it is more likely to be a cause in the actual setting. We further argue that our method resolves these experimental observations that seemingly contradict the human preference for abnormal causes.

B Computational Methods and Extended Experiments

B.1 Methodology

In this section, we prove the Explanatory Impossibility Theorem (**??**): we prove that substantial causal assumptions are needed to infer L_3 SVs from data. Following this motivation, we introduce **??** to bound L_3 SVs from data and assumptions in the form of a causal diagram.

B.1.1 Explanatory Impossibility Theorem

To understand the inherent impossibility of uniquely inferring counterfactual quantities from data, we first define the notion of a *bound* on a counterfactual quantity, following **?**.

Definition 7 (Bound)**.** *Consider SCM class $\Omega' \subseteq \Omega$, counterfactual quantity $f : \Omega \to \mathbb{R}$, and some $a, b \in \mathbb{R}$. Interval $[a, b]$ is a bound on f over SCM class Ω' if for all $\mathcal{M} \in \Omega'$,*

$$a \leq f(\mathcal{M}) \leq b. \tag{14}$$

$[a, b]$ is the tightest bound on f over Ω' if there is no bound $[a', b']$ on f over Ω' such that $a' > a$ or $b' < b$.

We may state that Ω' yields no information about f if the tightest bound $[a, b]$ over Ω' on f is also the tightest bound over Ω on f. In this case, the information that Ω' contains the true SCM does not inform the set of possible values of f. In general, substantive causal information is needed in order to construct valid explanations, as shown below.

Counterfactual Shapley values are quite similar to the probability of necessity **?** in terms of identification and bounding. Indeed, when variables X, Y are binary and observed to be equal to 1 in context $\mathbf{E} = \mathbf{e}$, we have:

$$\begin{aligned} \text{NTE}(X, Y|\mathbf{e}) &= \mathbb{E}[Y|\mathbf{e}] - \mathbb{E}[Y_{P(X)}|\mathbf{e}] && (15)\\ &= 1 - (P(x)P(y|\mathbf{e}) + P(x')P(y_{x'}|\mathbf{e})) && (16)\\ &= P(x')PN(x, y|\mathbf{e}). && (17) \end{aligned}$$

This implies that in certain binary settings, we may use existing bounds on the *PN* **?** to bound counterfactual Shapley values. Particularly, when Markovianity holds, we have:

$$\max\left(0, 1 - \frac{P(y_{x'})}{P(y_x)}\right) \leq PN(x, y) \leq \min\left(1, \frac{P(y'_{x'})}{P(y_x)}\right). \tag{18}$$

Below, we illustrate an application of these bounds.

Example 2 (Two-variable binary Markovian chain)**.** *Consider a binary Markovian SCM with observed variables* $\{X_1, X_2, Y\}$ *with an observational distribution factorizing as:*

$$P(\mathbf{v}) = P(x_1)P(x_2|x_1)\mathbf{1}[y = x_2]. \tag{19}$$

We cannot infer any bounds on ϕ^{L_3} *when the assumption of Markovianity is removed* **?***. With the additional information that the SCM is Markovian, we know that* $\phi_1^{L_1} = \phi_1^{L_2}$*. In addition, we may apply the bounds derived in* **??***, observing that:*

$$\phi_1^{L_1} = \phi_1^{L_2} = P(x'_1)(P(y_{x_1}) - P(y_{x'_1})) \tag{20}$$

$$\leq P(x'_1)\max\left(0, 1 - \frac{P(y_{x'})}{P(y_x)}\right) \tag{21}$$

$$\leq \phi_1^{L_3}, \tag{22}$$

for all choices of $P(y_x), P(y_{x'})$*, with equality holding when* $P(y_x) = 1$ *or* $P(y_x) = P(y_{x'})$*.*

As a simple illustration, consider the following two SCMs where $\mathbf{V} = \{X_1 = 1, X_2 = 1, Y = 1\}$*:*

$$\mathcal{M}_1 = \begin{cases} U_1, U_2 & \sim \text{Bern}(0.5) \\ X_1 & = U_1 \\ X_2 & = X_1 \oplus U_2 \\ Y & = X_2 \end{cases} \tag{23}$$

$$\mathcal{M}_2 = \begin{cases} U_1, U_2 & \sim \text{Bern}(0.5) \\ X_1 & = U_1 \\ X_2 & = U_2 \\ Y & = X_2 \end{cases} \tag{24}$$

We may observe that $P(y_{x_1}) = P(y_{x'_1}) = 0.5$ *in both SCMs, implying that* $\phi_1^{L_1} = \phi_1^{L_2} = 0$ *in both SCMs. However, in* $\mathcal{M}_1$*, changing* X_1 *will always change* Y*, while in* $\mathcal{M}_2$*, changing* X_1 *will never change* Y*. This yields:*

$$\phi_1^{L_3}(\mathcal{M}_1) = \frac{1}{2}(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]) \tag{25}$$

$$= \frac{1}{2}(1 - 0) = \frac{1}{2}$$

$$\phi_1^{L_3}(\mathcal{M}_2) = \frac{1}{2}(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]) \tag{26}$$

$$= \frac{1}{2}(1 - 1) = 0$$

This corresponds to the bounds obtained in **??***, illustrating that* $\phi_1^{L_3}(\mathcal{M}_1) \in [0, \frac{1}{2}]$ *and is not identified by data, even in a setting where* L_1 *and* L_2 *Shapley values are both identified and equal to zero.*

As a concrete example, it would be reasonable in $\mathcal{M}_1$ *to claim that choosing to buy rather than short a stock,* $X_1 = 1$*, is an explanation for positive returns* $Y = 1$*, even if the sign happens to be entirely uncorrelated with the decision of whether to buy or short; contrarily, it would be absurd to claim in* $\mathcal{M}_2$ *that an unrelated coin flip landing on heads* $X_1 = 1$ *is an explanation for positive returns.*

The two types of scenarios captured by $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are indistinguishable given the causal diagram and these particular observational and experimental data distributions, and we argue that it is correct to output a bound in this case rather than claiming an exact value of zero, as* L_1 *and* L_2 *Shapley values do.*

We generalize this intuition in the Explanatory Impossibility Theorem.

In this subsection, we have motivated the problem of bounding explanatory variable attributions in **??**, illustrating that it is not a limitation of our method but rather a result of *epistemic* uncertainty about the underlying data-generating model that cannot be reduced by obtaining more data, and which can only be reduced by making substantive causal assumptions about the underlying data-generating model. Therefore, any sound explanation technique must either require more information than observational data, such as interventional data or structural causal assumptions, in order to output any inference on a variable's contribution to the outcome.

B.1.2 Bounding Counterfactual Shapley values

In this subsection, we introduce **??**, a method to bound counterfactual Shapley values, extending the counterfactual identification algorithm of **?**. The approach constructs two neural causal models $\hat{M}_1, \hat{M}_2$ consistent with causal diagram $\mathcal{G}$; it respectively minimizes and maximizes $\phi_X(w)$ for some $X \in \mathbf{V}$, subject to the constraint that the optimized model is consistent with observed data $\mathbb{Z}(\mathcal{M})$.

Algorithm 2 Bounding counterfactual Shapley values

Require: Query $q : \Omega \to \mathbb{R}$, variable of interest $X \in \mathbf{V}$, L_2 datasets $\mathbb{Z}(\mathcal{M})$, and causal diagram $\mathcal{G}$.
Ensure: Bounds on $\phi_X^{L_3}$.
$\hat{M} \leftarrow \text{NCM}(\mathcal{G}; \theta)$
$\phi_X^{\min} \leftarrow \arg\min_\theta \Omega(\hat{M})$ s.t. $\mathbb{Z}(\hat{M}(\theta)) = \mathbb{Z}(\mathcal{M})$
$\phi_X^{\max} \leftarrow \arg\max_\theta \Omega(\hat{M})$ s.t. $\mathbb{Z}(\hat{M}(\theta)) = \mathbb{Z}(\mathcal{M})$
return $\phi_X^{\min}, \phi_X^{\max}$

B.2 Experimental Setup

In this appendix, we provide additional details on our experimental setup and approach, complementing the experiments described in Sec. **??** of the main text. Our experimental setting can be described as semi-synthetic – we generate our data from a ground truth SCM, while the data is modeled on a real-world dataset (MNIST and CelebA examples). In addition to these examples, in this appendix we also discuss synthetic examples, which illustrate some further failure modes of the methods in the literature.

Our experimental approach consists of two separate steps. In our first step, we are interested in establishing whether the L_3 Shapley values match with the human intuition on explanations. For this step, having access to the ground truth SCM is helpful, since we can compute any quantity (such as L_1, L_2, or L_3 Shapley values) based on the SCM, without the limitations of finite samples or identifiability issues. After establishing that our method is aligned with human intuition (using the ground truth SCM), while other methods are not, we move to the second step of our experimental setup – inferring L_3 Shapley values from a combination of causal assumptions (encoded in the causal diagram) and data. This second step corresponds to real-world settings, in which we almost never have access to the underlying SCM.

The remainder of this appendix is organized according to the above two steps. First, we go over our examples, describing the ground truth SCMs we constructed (Apps. **??**-**??**). After this, we discuss how to use the bounding technique described above in order to infer L_3 Shapley values from assumptions and data (App. **??**).

B.2.1 Color MNIST – Ground Truth

We first described the ground truth SCM for the color MNIST experiment, based on **?**. In this example, we consider four variables, namely: the hue X of the image, the digit Y appearing the image. The values of X, Y influence the 28×28 colored MNIST image I. Additionally, we consider the digit classifier $\hat{Y}$, which is a deterministic function of I. In our constructed SCM, hue X and digit Y are confounded, and digit Y and image I are confounded through the image's saturation (the confounding

is through the latent variable u_Y). The full SCM is given by:

$$P(U) = \begin{cases} u_Y & \sim \text{Unif}(\{0, \dots, 9\}) \\ u_X & \sim \text{Unif}(0, 1) \\ u_I^i & \sim \text{MNIST}(i) \end{cases} \tag{27}$$

$$\mathcal{F}_{\beta,f} = \begin{cases} X & = \left(\frac{u_Y}{9} + 0.5\Phi(u_X) + \beta\right) \mod 1 \\ Y & = u_Y \\ I & = \text{hsv\_to\_rgb}\left(u_I^{Y=y}, \frac{u_Y}{9}, X\right) \\ \hat{Y} & = \hat{f}(I) \end{cases} . \tag{28}$$

Here, β represents a hue shift parameter, f represents an image classifier, MNIST(i) denotes an MNIST image containing the digit i selected uniformly at random, and hsv_to_rgb denotes the conversion of a hue, saturation, and value triplet to a 28×28 RGB image. The causal diagram for this SCM is shown in Fig. **??**.

The is aim to explain two LeNet **?** classifiers trained on the color MNIST dataset: a standard LeNet classifier f, and a "robust" model g which applies a greyscale transform to the data before fitting to it (these are the classifiers constructed by Bob and Alice, respectively). The relevant why-query to explain either model's prediction is Why($\hat{y}|x, y, i$). The detailed interpretation of the different explanation methods is described in the main text (see Sec. **??**).

B.2.2 CelebA – Ground Truth

We next describe the ground truth SCM for the CelebA experiment, based on **?**. We consider four variables: the smiling indicator S, the indicator of whether the person's mouth is open M, the image of the person I (affected by S, M). Additionally, we also consider a classifier $\hat{M}$, predicting whether the person's mouth is open, based on the image I. The full CelebA SCM is given by:

$$\mathcal{F} = \begin{cases} S & = U_S \\ M & = \begin{cases} 0 & U_M = 0 \\ s & U_M = 1 \\ 1 & U_M = 2 \end{cases} \\ I & = U_I^{s,m} \\ \hat{M} & = f_{\hat{M}}(I) \end{cases} \tag{29}$$

$$P(\mathbf{U}) = \begin{cases} U_S & \sim \text{Bern}(0.5) \\ U_M & \sim \text{Categorical}([0.05, 0.9, 0.05]) \\ U_I & \sim \text{CelebA-HQ}(\texttt{Smiling, Mouth\_Slightly\_Open}) \end{cases} . \tag{30}$$

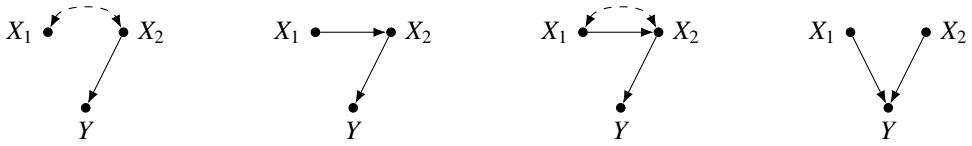
Here, CelebA-HQ(`Smiling, Mouth_Slightly_Open`) denotes a distribution over a list of four CelebA-HQ images, such that $U_I^{s,m}$ denotes an image where `Smiling` $S = s$ and `Mouth_Slightly_Open` $M = m$. In the SCM, as in the real world, smiling S has a positive effect on the mouth being open M. The causal diagram for the setting is shown in Fig. **??**.

The aim is to explain two diffusion-based classifiers, constructed by Bob and Alice. Bob constructed a "standard" classifier $\hat{M}^B$, while Alice constructed a "robust" classifier $\hat{M}^A$, who applied a re-weighing transformation to her dataset before fitting a model. The relevant why-query to explain either model's prediction is Why($\hat{m}|s, m, i$).

Smiling $S = 1$ never causes the mouth to be closed $M = 0$, implying no causal effect on the path $S \to M \to I \to \hat{M}$. By construction, the standard classifier is expected to have an effect along the path $S \to I \to \hat{M}$, while the robust classifier is not. Thus, by causal admissibility (D_1), when $S = 1, M = 0$, an EVA should satisfy $\phi_S^{\text{rob}} \approx 0, \phi_S^{\text{std}} > 0$.

We empirically compute SVs for a single sample under this setting. We show that L_1 and L_2 yield $\phi_S^{\text{rob}}, \phi_S^{\text{std}} > 0$ with statistical confidence, failing to distinguish between classifiers. In contrast, L_3 satisfies the explanatory desiderata: ϕ_S^{rob} is statistically indistinguishable from zero, while $\phi_S^{\text{std}} > 0$.

We extend our single-sample result to 20 samples per configuration ($S = s, M = m$). Applying a two-tailed z-test, L_3 yields $p < 0.001$ for $(0, 1)$ and $(1, 0)$, indicating a significant difference in



(a) Spurious graph (b) Chain graph (c) Bow graph (d) Conjunction graph

Figure 3: Causal diagrams for toy experiments.

attribution sign distributions between the standard and robust classifiers. In these cases, L_1 and L_2 remain unable to distinguish between classifiers.

We therefore conclude that L_3 Shapley values distinguish between standard and robust classifiers consistent with causal admissibility, whereas L_1 and L_2 cannot. The detailed interpretation of the different explanation methods and figures are described in the main text (see Sec. **??**).

B.2.3 Synthetic examples

In this section, we introduce several synthetic examples, which further highlight how our method improves upon prior work (on top of the semi-synthetic examples discussed above and in the main text). In particular, we evaluate L_1, L_2, and L_3 Shapley values on four SCMs, which we refer to as (a) spurious SCM; (b) chain SCM; (c) bow SCM. In all settings, variables are binary, and in the observed event $\mathbf{E} = \mathbf{e}$ they equal 1 ($\mathbf{v} = \{x_1, x_2, y\}$). Also, in each SCM, the unobserved U_i variables are sampled from Bern(0.5). $Y = 1$ is our event explanandum, and $\text{Why}(y|x_1, x_2)$ our query. Throughout, we focus on the attribution assigned to the first variable, X_1. Graphs for each SCM are shown in **??**. We next discuss each SCM in order.

Spurious SCM (Shark Attacks, ????) Daily shark attacks are high today ($X_1 = 1$), and so are ice cream sales ($X_2 = 1$); store profit is also high ($Y = 1$). The graph is shown in **??**. The ground truth SCM is given by:

$$\mathcal{M}_1 = \begin{cases} X_1 & := U_{12} \\ X_2 & := U_{12} \vee U_2 \\ Y & := X_2 \end{cases} \tag{31}$$

Given that the high shark attack incidence $X_1 = 1$ has no effect on Y, it should be assigned a zero attribution. However, in Fig. **??** (first column, blue bar), we observe that L_1 Shapley values give X_1 a non-zero attribution, violating the property of causal admissibility (**??**). Conversely, L_2, L_3 Shapley values satisfy admissibility in this example, giving a zero attribution to the variable X_1.

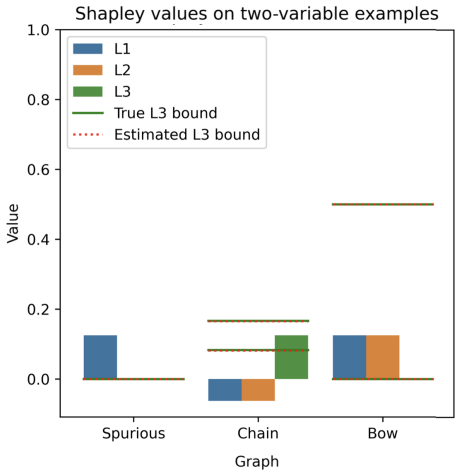


Figure 4: SVs for X_1 in toy experiment SCMs. Green and dotted red lines denote true and estimated bounds. Error bars are negligible.

Chain SCM (??) The causal diagram for the chain SCM is shown in **??**, and the SCM is given by:

$$\mathcal{M}_2 = \begin{cases} X_1 & := U_1 \\ X_2 & := (U_2^1 \wedge X_1) \vee (\neg U_2^1 \wedge (U_2^2 \vee \neg X_1)) \\ Y & := X_2 \end{cases} \tag{32}$$

In our observed event, $Y = 1$, meaning that the variable Y takes its maximum value. Therefore, we expect X_1 to have a non-negative effect on Y; X_1 could not have had a negative effect on Y, since Y attains its maximum. Contrary to this expectation, L_1 and L_2 SVs give negative attributions to the X_1 variable (see Fig. **??** second column, blue and red bars). Therefore, both of these methods provide

counterintuitive explanations. In contrast, the L_3 SVs for X_1 are strictly positive, in line with human intuition. Thus, even in this simple setting, one can see that L_3 SVs produce attributions superior to L_1, L_2 SVs.

Bow SCM (??) The causal diagram for the bow SCM is shown in **??**, and the SCM is given by

$$\mathcal{M}_3 = \begin{cases} X_1 & := U_{12} \\ X_2 & := (U_2^1 \wedge U_2^2) \vee ((U_2^1 \vee U_2^2) \wedge (X_1 \vee U_{12})) \\ Y & := X_2 \end{cases} \tag{33}$$

Given the observed even $X_1 = 1, X_2 = 1, Y = 1$, we can infer that that $U_{12} = 1$ based on the SCM. The fact that $U_{12} = 1$ further implies that X_1 has no effect on X_2 and thus could not have an effect on Y. Intuitively, therefore, we expect the variable X_1 to be given a zero attribution. The SCM is constructed such that X_1 has a positive effect on X_2 in some settings, and no effect in the observed setting $\{x_1, x_2, y\}$. We see that the L_3 SV for X_1 are approximately zero (third column of Fig. **??**). However, both L_1 and L_2 Shapley values violate our expectations and admissibility; once again, X_1 is incorrectly given a non-zero attribution while having no effect on Y.

Summary of synthetic examples. We argue that L_1 SVs differ from L_2, L_3 SVs in the spurious setting because L_1 SVs capture spurious effects, violating admissibility. Discrepancies in the chain and bow settings arise because L_1, L_2 SVs average over all units when considering the effect of X_1 on Y, where the effect is on average negative and positive, respectively. On the contrary, L_3 SVs only consider units consistent with the observations, where the effects are strictly non-negative and zero, respectively. Therefore, in the toy examples above, L_3 SVs are better aligned with human intuition for explanations.

B.2.4 Bounding L_3 Shapley values from Assumptions & Data

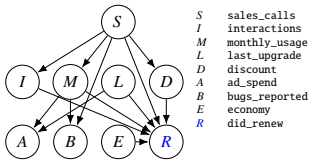
In this section, we discuss the bounding of L_3 SVs from real data and assumptions. We start with the MNIST example. As a sanity check, we first test whether the standard and robust classifiers behave as expected. For this, we investigate the performance of these models on a sample of 60000 generated samples from the color MNIST dataset. In this setting, both models achieve near-perfect accuracy (standard: 1, robust: 0.994). However, if we compare their performance on 60000 samples from the data-generating model with $\beta = 0.5$ (that is, with a distribution shift), we find that the standard model performance drops to close to random chance, while the robust model's performance is unaffected (standard: 0.181, robust: 0.994). Therefore, this empirically validates that the standard and robust classifiers behave as expected by our construction.

We then move onto bounding the L_3 SVs based on the data and the causal diagram. For computing the bounds, we make use of L_2 (interventional) data, and apply the bounding method described in App. **??**. Specifically, we train a conditional diffusion model with 4000 steps **?** to model $P(I|X, Y, U_Y)$ for 150 epochs. At inference time, we reduce this to 25 steps (**?**), still achieving realistic results. The bounds obtained on the SVs are shown in **??** as error bars, and we can see that the computed bounds from assumptions and data include the ground truth values computed from the SCM.

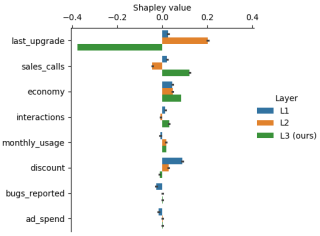
We next move onto estimating bounds on L_3 SVs for the synthetic examples (again using the methodology described in App. **??**), based on the causal diagram and the observational distribution. For the synthetic examples, as an additional verification, we can compute analytical bounds (as expressions based on the observational distribution) by leveraging the bounds on the probability of necessity (PN), introduced in **?**. This is possible given that all variables in the synthetic examples are binary, and L_3 SVs may therefore be computed using the observational distribution and the PN. The true bounds for L_3 SVs (based on the SCM) are shown as green intervals in **??**, while the bounds computed from the causal diagram and the observational distribution are shown as dotted orange intervals. We can see that the true and estimated bounds are identical, and that the computed bounds consistently include the true value of the L_3 SV, empirically corroborating the validity of our bounding approach.

B.3 Subscriber retention

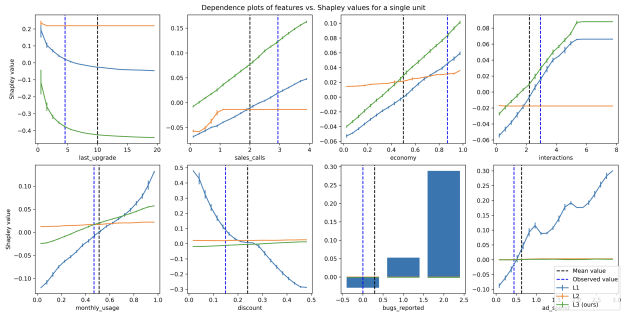
We provide evidence in support of L_3 SVs in a higher-dimensional tabular setting: the subscriber retention data generating process from **?**, describing customer cohorts' rates of subscription renewal



(a) Subscriber retention model causal diagram.



(b) Shapley values for observed sample. Error bars denote SEM.



(c) Dependence plots of L_1, L_2, L_3 SVs with respect to observed sample, holding all variables but one fixed. Error bars denote SEM. Blue dotted lines denote the observed value of the variable. Black dotted lines denote the mean. Outcome `did_renew` takes a value of 0.493 and has a mean of 0.350.

Figure 5: Subscriber retention experiments.

R, as affected by several other observed variables. The causal diagram is shown in **??**, and the underlying SCM is shown in **??**. Our goal is to understand why specific customer cohorts, identified by their attributes $\mathbf{X} = \mathbf{V} \setminus \{R\}$, have certain subscription renewal rates (R); our query is Why($r|\mathbf{x}$). We sample a random data point and first plot its L_1, L_2, and L_3 Shapley values, shown in **??**. Then, we plot its per-feature dependence plot, shown in **??**. Observed variable settings (blue lines) and variable means (black lines) are shown in per-feature dependence plots.

Comparison to L_1 SVs Because there is no causal path from the variables A, B to outcome R (see **??**), admissibility (**??**) requires $\phi_A = \phi_B = 0$. However, L_1 SVs yield non-zero attributions for A, B in **??**, violating admissibility. We argue this is due to the confounders S, M, which are ancestors of each of A, B, R and which induce spurious correlation between A, B and R. On the other hand, L_2 and L_3 SVs correctly assign zero attribution to A, B.

Comparison to L_2 SVs We observe that the retention mechanism R is monotonically increasing and point symmetric about $R(D) = 0.5$. In other words, a positive change Δ in D increases R by the same amount that the change $-\Delta$ decreases it, even if other variables are held fixed. Because our observed $R \approx 0.5$ is close to this midpoint, and D is below-average, we expect $\phi_D < 0$, since R would usually be higher. Indeed, the L_3 SV for D reflects this. By contrast, L_2 SVs instead capture how R behaves on average, around $R \approx 0.35$, where the the sigmoid R plateaus on the negative side. Because the negative slope is much shallower than the positive slope, negative effects arising from decreasing D to its below-average value are insignificant compared to positive effects arising from increasing D to it current value from even lower values. For concision, we discuss SV discrepancies for the L, S, I variables in **??**.

Dependence plot analysis Examining the dependence plots for `ad_spend` A and `bugs_reported` B in **??**, we see that L_1 SVs (blue) vary significantly with respect to A and B, while L_2 and L_3 SVs remain zero. This further supports the claim that L_1 SVs violate admissibility. Examining the dependence plots for D in **??**, we observe that for `discount` D, L_3 SVs (green) have a small positive slope, while L_2 SVs (orange) have plateaued at a small positive constant. This plateau is more clearly visible in the `sales_calls` plot: L_3 SVs (green) vary near-linearly in S, while L_2 SVs (orange) experience a plateau at a small negative number for $S > 1$. Given that we observe $R \approx 0.5$, we expect SVs to be approximately point symmetric about the blue line. This expectation is satisfied for L_3 SVs but not for L_2 SVs; thus, we conclude L_3 SVs better explain the actual cohort's renewal rate than L_2 SVs.

We conclude that L_3 SVs offer more intuitive explanations than L_1, L_2 SVs in this setting, primarily because their explanations capture effects on the observed cohort's renewal rate, while L_1, L_2 target their explanations towards the average cohort's renewal rate.

C Proofs

C.1 Proof of Desiderata Lifting Theorem

Theorem ?? (Desiderata Lifting)**.** *Under* **??***: (a)* $\neg D_1(c', \phi) \Rightarrow \neg D_1(c^*, \phi)$ *(admissibility), (b)* $D_2(c', \phi) \Rightarrow D_2(c^*, \phi)$ *(power), (c)* $\neg D_3(c', \phi) \Rightarrow \neg D_3(c^*, \phi)$ *(normality), (d)* $\neg D_4(c', \phi) \Rightarrow \neg D_4(c^*, \phi)$ *(responsivity).* □

Proof. By **??**, 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).

Admissibility: If $\neg D_1(c', \phi)$, then $\exists z : c'(\cdot, z) = 0$ but $\phi \neq \mathbf{0}$. By necessity ($c' = 0 \Rightarrow c^* = 0$), we have $c^*(\cdot, f^{-1}(z)) = 0$, so $\neg D_1(c^*, \phi)$.

Power: If $D_2(c', \phi)$ and $c^*(\mathbf{u}, z^*) \neq 0$, then by necessity $c'(\mathbf{u}, f(z^*)) \neq 0$, so by assumption $\phi \neq \mathbf{0}$, proving $D_2(c^*, \phi)$.

Normality and Responsivity: By effect preservation via mapping f, any violation of normality or responsivity patterns under c' corresponds to the same violation under c^*. □

C.2 Proof of Explanatory Impossibility Theorem

Theorem ?? (Explanatory Impossibility Theorem)**.** *No EVA can compute exact explanatory variable attributions satisfying desiderata* $D_{1:4}$ *from observational or interventional data alone.* □

Proof. We prove by contradiction that no single method can satisfy admissibility (D_1) across all SCMs consistent with observational and interventional data.

Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ be binary domains. We construct two SCMs that demonstrate the impossibility. First, consider SCM $\mathcal{M}_1$ defined by

$$X = U_X \tag{34}$$
$$Y = X \tag{35}$$

where $U_X \sim$ Bernoulli(0.5). Second, consider SCM $\mathcal{M}_2$ defined by

$$X = U_X \tag{36}$$
$$Y = U_Y \tag{37}$$

where $(U_X, U_Y) \sim P(U_X, U_Y)$ with $P(U_Y = 1|U_X = 1) = 1$, $P(U_Y = 1|U_X = 0) = 0$, and $P(U_X) =$ Bernoulli(0.5).

Both SCMs induce identical observational distributions given by

$$P^{\mathcal{M}_1}(X, Y) = P^{\mathcal{M}_2}(X, Y) = \{(0, 0) : 0.5, (1, 1) : 0.5, (0, 1) : 0, (1, 0) : 0\}. \tag{38}$$

The interventional distributions differ between the two models. For $\mathcal{M}_1$, interventions on X directly affect Y, yielding

$$P^{\mathcal{M}_1}(Y = 1|\text{do}(X = x)) = x. \tag{39}$$

For $\mathcal{M}_2$, interventions on X have no effect on Y, yielding

$$P^{\mathcal{M}_2}(Y = 1|\text{do}(X = x)) = P(Y = 1) = 0.5. \tag{40}$$

Under our conservative proxy c' from Definition **??**, the functional dependence evaluations yield different results. In $\mathcal{M}_1$, we have

$$c'_1(\mathcal{M}_1, \mathbf{u}, X, Y, z) = Y_{X(\mathbf{u})}(\mathbf{u}) - Y_{x'}(\mathbf{u}) \neq 0 \tag{41}$$

for some witness z. In $\mathcal{M}_2$, we have

$$c'_2(\mathcal{M}_2, \mathbf{u}, X, Y, z) = Y_{X(\mathbf{u})}(\mathbf{u}) - Y_{X(\mathbf{u})}(\mathbf{u}) = 0 \tag{42}$$

for all witnesses z.

By the admissibility requirement D_1, these functional dependence evaluations impose contradictory constraints. For $\mathcal{M}_1$, since $c'_1 \neq 0$, admissibility requires $\phi_X \neq 0$. For $\mathcal{M}_2$, since $c'_2 = 0$, admissibility requires $\phi_X = 0$.

Any method ϕ operating on observational and interventional data must output a single attribution value ϕ_X for the observed data. This creates an unavoidable contradiction:

$$\phi_X = 0 \implies \text{satisfies } D_1 \text{ for } \mathcal{M}_2 \text{ but violates } D_1 \text{ for } \mathcal{M}_1 \tag{43}$$

$$\phi_X \neq 0 \implies \text{satisfies } D_1 \text{ for } \mathcal{M}_1 \text{ but violates } D_1 \text{ for } \mathcal{M}_2. \tag{44}$$

The Causal Hierarchy Theorem establishes that counterfactual quantities required for evaluating functional dependence are not identifiable from observational and interventional data alone. Since our desiderata evaluation depends on these Layer 3 quantities, no method can satisfy admissibility consistently across all SCMs compatible with the observed data. This impossibility extends to all desiderata $D_{1:4}$, establishing that exact computation of desiderata-satisfying explanations requires additional structural assumptions about the data-generating process. □

C.3 Proof of Bounding Soundness

Theorem ?? (Soundness of bounding algorithm)**.** *Under* **??***, the bounds produced by* **??***($\mathcal{D}$, $\mathcal{G}$, w, X) contain the true L_3 Shapley Value: $\phi_{L_3,X}(w) \in [\phi^-_{L_3,X}(w), \phi^+_{L_3,X}(w)]$.* □

Proof. Under the L_3-G-expressivity assumption (**??**), there exists a neural causal model $f^* \in \mathcal{F}_G$ that can represent the true NTE quantities with arbitrary precision.

Since $f^* \in \mathcal{F}_G$, the true L_3 Shapley value $\phi_{L_3,X}(w)$ can be computed by f^* with arbitrary precision. Let $\phi^*_{L_3,X,f^*}(w)$ denote this value computed by f^*.

?? computes bounds over all trained models in the ensemble $\{f_k\}_{k=1}^K \subset \mathcal{F}_G$:

$$[\phi^-_{L_3,X}(w), \phi^+_{L_3,X}(w)] = [\min_k \phi_{L_3,X,k}(w), \max_k \phi_{L_3,X,k}(w)] \tag{45}$$

Since the ensemble training procedure can approximate any model in $\mathcal{F}_G$ (including f^*) on the dataset $\mathcal{D}$, and $f^* \in \mathcal{F}_G$ by assumption, there exists a model f_j in the ensemble such that $\phi_{L_3,X,j}(w) \approx \phi^*_{L_3,X,f^*}(w) = \phi_{L_3,X}(w)$.

Therefore:

$$\phi^-_{L_3,X}(w) \leq \phi_{L_3,X,j}(w) \approx \phi_{L_3,X}(w) \leq \phi^+_{L_3,X}(w) \tag{46}$$

Thus, $\phi_{L_3,X}(w) \in [\phi^-_{L_3,X}(w), \phi^+_{L_3,X}(w)]$. □

C.4 Proof that L_3 SVs Satisfy All Desiderata

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

C.4.1 Proof of D_1 (Admissibility)

Lemma 1 (Causal Admissibility)**.** *L_3 SVs satisfy causal admissibility under proxy c'.*

Proof. Following the definition of causal admissibility, assume that for all baselines $z = (x', \mathbf{z}') \in \mathcal{Z}$:

$$c'(\mathcal{M}, \mathbf{u}, X, Y, z) = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}',x'}(\mathbf{u}) = 0 \tag{47}$$

for all $\mathbf{u}$ consistent with the why-query.

By **??**, this implies:

$$\text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = 0 \tag{48}$$

for all subsets $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$ and all units $\mathbf{u}, \mathbf{u}'$.

The L_3 Shapley value is:

$$\phi_{L_3,X}(w) = \mathbb{E}_{\substack{\pi\sim\text{Unif}(\Pi_{\mathbf{X}}) \\ \mathbf{u}'\sim P_{\mathcal{M}}(\mathbf{U}) \\ \mathbf{u}\sim P_{\mathcal{M}}(\mathbf{U}|\mathbf{V}=\mathbf{v})}} \left[\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u})\right] \tag{49}$$

Since every marginal contribution $\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u}) = 0$ for all permutations π and units $\mathbf{u}, \mathbf{u}'$, the expectation equals zero: $\phi_{L_3,X}(w) = 0$.

Therefore, L_3 SVs satisfy causal admissibility. □

C.4.2 Proof of D_2 (Power)

Lemma 2 (Weak Causal Power)**.** *L_3 SVs satisfy weak causal power (measure-1 over SCMs) under proxy c'.*

Proof. If there exists baseline $z^* = (x'^*, \mathbf{z}'^*)$ and unit $\mathbf{u}^*$ such that:

$$c'(\mathcal{M}, \mathbf{u}^*, X, Y, z^*) = Y_{\mathbf{z}'^*}(\mathbf{u}^*) - Y_{\mathbf{z}'^*, x'^*}(\mathbf{u}^*) \neq 0 \tag{50}$$

Then there exists at least one subset $\mathbf{S} = \mathbf{Z}'^* \cap \mathbf{X}$ where the marginal contribution is non-zero:

$$\text{NTE}(\mathbf{S} \cup \{X\}, Y|\mathbf{u}'^* \to \mathbf{u}^*) - \text{NTE}(\mathbf{S}, Y|\mathbf{u}'^* \to \mathbf{u}^*) \neq 0 \tag{51}$$

where $\mathbf{u}'^*$ induces the witness values $(x'^*, \mathbf{z}'^*)$.

The L_3 Shapley value is a linear combination of NTE marginal contributions. For $\phi_X^{L_3}(w) = 0$, the weighted sum must exactly equal zero. Under the measure defined in **??**, this exact cancellation occurs with probability zero.

Therefore, L_3 SVs satisfy weak causal power for a measure-1 subset of SCMs. □

C.4.3 Proof of D_3 (Normality)

Lemma 3 (Causal Normality)**.** *L_3 SVs satisfy causal normality under proxy c'.*

Proof. Consider two SCM-unit pairs $(\mathcal{M}_1, \mathbf{u}_1)$ and $(\mathcal{M}_2, \mathbf{u}_2)$ satisfying the conditions of D_3:

- All counterfactuals on Y are identical for X_1 (in world 1) and X_2 (in world 2)
- Both cause the outcome according to c'
- For all baselines $z = (x', \mathbf{z}')$, the normality condition holds:

$$\text{sign}(\Delta Y) \cdot P_1(z) \leq \text{sign}(\Delta Y) \cdot P_2(z) \tag{52}$$

  where $\Delta Y = Y_{\mathbf{z}'}(\mathbf{u}) - Y_{\mathbf{z}', x'}(\mathbf{u})$, with strict inequality for at least one baseline.

The L_3 Shapley value involves an expectation over baselines that weights contributions by baseline probability through $P(\mathbf{U})$. For positive effects ($\Delta Y > 0$), lower baseline probability means higher attribution. Since the counterfactuals are identical but baseline probabilities differ, we have $\phi_{X_1}^{L_3} > \phi_{X_2}^{L_3}$.

Therefore, L_3 SVs satisfy causal normality. □

C.4.4 Proof of D_4 (Effect Responsivity)

Lemma 4 (Effect Responsivity)**.** *L_3 SVs satisfy effect responsivity under proxy c'.*

Proof. Consider two SCM-unit pairs satisfying the conditions of D_4 with identical baseline distributions but different effect magnitudes. Since functional dependence c' directly measures counterfactual differences and the L_3 Shapley value aggregates these effects, larger effects with identical baseline distributions result in larger attributions.

Therefore, L_3 SVs satisfy effect responsivity. □

C.4.5 Main Result

Theorem ?? (L_3 SVs satisfy explanatory desiderata)**.** *Under the functional dependence causation proxy c′ (**??**), L_3 SVs satisfy: admissibility (D_1) for all SCMs, power (D_2) for a measure-1 subset of SCMs (see **??** for measure definition), normality (D_3) for all SCMs, and effect responsivity (D_4) for all SCMs.* □

Proof. The result follows directly from Lemmas **??**, **??**, **??**, and **??**. □

C.5 Proof of Shapley Axioms

Corollary 7 (L_3 SVs satisfy Shapley axioms)**.** *L_3 Shapley Values satisfy the fundamental Shapley axioms of efficiency, symmetry, and marginality when applied to the value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|w)$.*

Proof. This follows directly from Young's theorem (**?**), which establishes that any value satisfying efficiency, symmetry, and marginality must be the Shapley value.

Since L_3 SVs are defined as:

$$\phi_{L_3,X}(w) = \mathbb{E}_{\substack{\pi\sim\text{Unif}(\Pi_{\mathbf{X}}) \\ \mathbf{u}'\sim P_{\mathcal{M}}(\mathbf{U}) \\ \mathbf{u}\sim P_{\mathcal{M}}(\mathbf{U}|\mathbf{V}=\mathbf{v})}} \left[\text{NTE}(\pi_{\leq X}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\pi_{<X}, Y|\mathbf{u}' \to \mathbf{u})\right] \tag{53}$$

which is equivalent to applying the Shapley value to the NTE value function $f(\mathbf{Z}) = \text{NTE}(\mathbf{Z}, Y|w)$, they inherit the Shapley axioms by construction.

Efficiency: $\sum_{X\in\mathbf{X}} \phi_{L_3,X}(w) = \text{NTE}(\mathbf{X}, Y|w) - \text{NTE}(\emptyset, Y|w)$

Symmetry: If variables X_i and X_j have identical marginal contributions for all coalitions, then $\phi_{L_3,X_i}(w) = \phi_{L_3,X_j}(w)$.

Marginality: The attribution depends only on the marginal contributions of each variable to all possible coalitions. □

D Large Language Model Usage Disclosure

LLMs were used for writing assistance, literature search query formulation, and technical proofreading. LLMs were not used for research ideation or theoretical development. The authors take full responsibility for all content.