TECHNICAL REPORT
R-135
May, 2025

Causal Explanations through Counterfactual Variable Attributions

Kai-Zhan Lee and **Drago Plecko** and **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
kl2792@columbia.edu, dp3144@columbia.edu, eb@cs.columbia.edu

Abstract

Answering the question "Why did this outcome occur?" is fundamental to the empirical sciences and, in particular, the explainable artificial intelligence (XAI) literature. However, there is a need for a conceptually grounded and technically precise definition of this query, along with valid corresponding explanations. We address this gap by developing a formal causal explanation framework from first principles, grounding Lewis's view that an explanation is a summary of an event's causal history. We formally define and justify three novel counterfactual quantities that capture the causal history of an event: the natural total effect, the generalized direct effect, and counterfactual Shapley values. We introduce three formal explanatory desiderata – causal admissibility, explanatory power, and normality – and demonstrate that our method satisfies all desiderata, unlike any prior approach found in the literature. Simulated experiments support our results.

1 Introduction

The challenge of explaining observed phenomena – patterns of events – has long been a central goal of science. From Newtonian mechanics to quantum field theory, scientific explanation typically relies on identifying underlying mechanisms that govern a system's behavior through observations, interventions, and mathematical modeling. A complete system of equations is considered adequate to explain a phenomenon not merely because it enables prediction, but because it captures how different parts of the system interact, evolve over time, and give rise to the observed behavior. Naturally, this learning process is both challenging and ongoing – it may take centuries before we can confidently say we truly understand a given phenomenon.

A similar challenge in explaining the behavior and decision-making of AI systems has recently emerged. As AI plays an increasingly central role in society, there is a growing demand for intelligent systems to explain themselves – to articulate the reasons behind their actions, describe their current understanding of the world, and acknowledge what they do not know. This requirement is essential for ensuring stakeholder trust in AI models, supporting end-user autonomy, providing recourse for contesting model outputs, and improving our ability to debug these systems. An expanding body of research under the rubric of explainable AI (XAI) seeks to meet this challenge [24].

Despite this parallel, there are fundamental obstacles in directly applying the scientific method to AI systems. Scientific advances have traditionally focused on physical systems that do not involve humans – or even deliberately remove them from the loop. In contrast, XAI methods often operate in environments that are inherently social, interactive, and adaptive – yet they are typically blind to the causal mechanisms underlying the systems in which they are embedded. Thus, the methods and assumptions that work well in the natural sciences cannot be applied to AI systems without significant retooling. Indeed, a proper definition of explanation itself has eluded formalization in much of the existing literature of causality and explainable AI [31, 23, 26, 24].

We address the challenge of developing a causal explanatory framework with the intent of capturing the essence of scientific explanation while accounting for the complexity of real-world AI settings – where humans play a central role. Our first key observation is that modern causal language provides a natural foundation for this task: both physical and AI systems can be described using *Structural Causal Models (SCMs)* [30, 31]. Each SCM encodes a generative process – a collection of causal mechanisms – and gives rise to the *Pearl Causal Hierarchy (PCH)*, which organizes different levels of reasoning about the system [1]. The *observational* level captures what is seen in the world, factual events; the *interventional* level models how outcomes would change under hypothetical actions; and the *counterfactual* level allows us to ask what would have happened under different conditions, even if those conditions never occurred in this world.

Despite its power, one observation is that SCMs are almost never identifiable from data. Given only observational data, multiple SCMs may be consistent with the observed distribution, while yielding vastly different answers at the interventional or counterfactual levels. This limitation is formalized in the *Causal Hierarchy Theorem (CHT)* [1, Thm. 1], which shows that data from lower levels of the PCH underdetermine an SCM and thus our ability to learn an explanation. And even if a complete SCM were somehow learnable, the resulting equations are often too intricate to serve as human-friendly explanations.

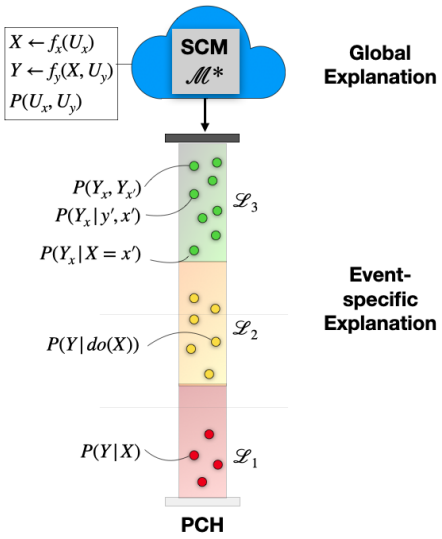


Figure 1: An SCM (global explanation) induces the PCH, which enables reasoning about specific events.

Given this high bar required from an explanation, we shift focus from constructing global explanations (the whole SCM) to explaining *specific events*, under only a partial observability of the SCM and the corresponding PCH. To address this, we propose a formal framework for event-level explanation grounded in causal inference. We study questions of the form: "Why is the outcome $Y = y$, given that $\mathbf{X} = \mathbf{x}$?", assuming event $\mathbf{E} = (\mathbf{X} = \mathbf{x}, Y = y)$ was in fact observed. Here, $Y = y$ is referred to as the *event explanandum*, the event we aim to explain.There are three key challenges in pursuing our goal. First, the why-question must be formalized, which requires resolving ambiguities such as the choice of contrasts or *foil* – a well-acknowledged issue in philosophy and cognitive science [18, 26]. Second, the construction of an explanation requires attributing the variation in the explanandum to specific variables – a process closely tied to notions of variable importance and decomposition [34, 5]. Third, we must ask what makes an explanation *good*. Drawing on insights from philosophy, psychology, cognitive science, and jurisprudence [9, 18, 4], we propose a formal set of *desiderata* – including causal admissibility, explanatory power, and normality – that any explanation should satisfy.

In this paper, we address these challenges and propose a framework for generating and evaluating explanations through a causal lens. More specifically, our contributions are as follows:

- **Causal Foundations.** We introduce the *Natural Total Effect (NTE)* and the *Generalized Direct Effect (GDE)* to extend univariate total effects to multivariate interventions. We show that these quantities capture the causal history of an event (Thms. 1 and 2). These quantities yield our proposed explanation method: *counterfactual Shapley values* (L_3 SVs).
- **Causal-Explanation Desiderata.** We formally define *Explanatory Variable Attribution (EVA)* and introduce three properties it should satisfy: *causal admissibility*, *causal explanatory power*, and *causal normality*. We prove that our proposed method satisfies the desiderata (Thm. 4).

Our framework contributes to the broader goal of building interpretable and trustworthy AI systems grounded in causality. Experiments corroborate our theoretical findings. Formal proofs are provided in App. A. Further discussion, including counterexamples to existing methods, appears in App. B.

Preliminaries. Random variables are denoted by capital letters X, and their values with corresponding lowercase letters x. Sets of random variables $\mathbf{X}$ are bolded. The domain of a random variable X is denoted by $\mathcal{D}_X$. We use the language of structural causal models (SCMs) as our basic semantic

framework [30], following the presentation in [1]. An SCM is a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$, where $\mathbf{V}$, $\mathbf{U}$ are sets of endogenous (observables) and exogenous (latent) variables, respectively, $\mathcal{F}$ is a set of functions f_{V_i}, one for each $V_i \in \mathbf{V}$, where $V_i \leftarrow f_{V_i}(\mathbf{pa}(V_i), \mathbf{U}_{V_i})$ for some $\mathbf{pa}(V_i) \subseteq \mathbf{V}$ and $\mathbf{U}_{V_i} \subseteq \mathbf{U}$. $P(\mathbf{u})$ is a strictly positive probability measure over $\mathbf{U}$. Each SCM $\mathcal{M}$ is associated with a causal diagram $\mathcal{G}$ [30, 1] over the node set $\mathbf{V}$ 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. An instantiation of the exogenous variables $\mathbf{U} = \mathbf{u}$ is often called a *unit*. By $Y_x(\mathbf{u})$ we denote the potential outcome of Y when setting $X = x$ for the unit u, which is the solution for $Y(u)$ to the set of equations obtained by evaluating the unit u in the submodel $\mathcal{M}_x$, in which all equations associated with X are replaced by $X = x$. We assume all observations are generated by a causal world [30][Def. 7.1.8], a tuple $(\mathcal{M}, \mathbf{u}) \in \mathcal{W}$, the space of all worlds, where $\mathcal{M}$ is an SCM, and $\mathbf{u}$ is a particular realization of exogenous $\mathbf{U}$.

2 A Causal Framework for Explanations

To explain specific events, we must first formally define a *why query*. We distinguish why queries from natural language why questions, which may be ambiguous, as they often fail to specify an alternative value of $Y = y'$ against which the observed event $Y = y$ is compared. In the literature, this is known as the choice of *foil* [26], and can substantially affect valid answers to the question [18]. For instance, one's explanation in response to "Why is the GDP growth rate in the United States is 2%?" may differ significantly when comparing to the expected 5% vs. a negative growth rate elsewhere in the world. Therefore, for precision in our why query, we precisely specify the **observed events** $\mathbf{E} = \mathbf{e}$, the **event explanandum** $Y = y \in \mathbf{e}$, and **explanatory variables** $\mathbf{X}$.

Definition 1 (Why Query)**.** *Given SCM $\mathcal{M}$, a why query is a tuple $(Y = y, \mathbf{X}, \mathbf{E} = \mathbf{e}) \in \mathcal{W}$, where $\mathcal{W}$ is the space of why queries. The tuple consists of the SCM $\mathcal{M}$ with observed variables $\mathbf{V}$, observed evidence $\mathbf{E} = \mathbf{e}$, where $\mathbf{E} \subseteq \mathbf{V}$, event explanandum $Y = y$ implied by $\mathbf{e}$, and explanatory variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$. This tuple is denoted by $\mathrm{Why}(y|\mathbf{e}; \mathbf{X})$, or $\mathrm{Why}(y|\mathbf{x})$ when $\mathbf{e} = \mathbf{x} \cup \{y\}$.* □

To answer why queries, we follow the philosophical insights of Lewis [21]: "an explanation of an event explanandum is information about the causal history of that event." In other words, we argue that **an explanation of an event is a summary of the causes of that event.**

Under this framing, our goals are twofold. In Sec. 2.1, we aim to formalize the entire causal history of an event explanandum. Our discussion will demonstrate that the causal history is captured by a collection of multivariate total effects of explanatory variable subsets $\mathbf{Z} \subseteq \mathbf{X}$ on event Y, and that these effects need to be weighted by their baseline probabilities $P(\mathbf{X} = \mathbf{x}')$. Next, in Sec. 2.2, we convert this causal history into a set of variable-specific effects that describe how each single variable $X_i \in \mathbf{Z}$ contributes to a multivariate total effect of $\mathbf{Z}$ on Y, and we construct a variable attribution method which summarizes this set of variable-specific effects.

2.1 Multivariate Total Effects

How do we capture the full *causal history* of an event, reflecting the set of effects of explanatory variables $\mathbf{X}$ on the observed event $Y = y$? In traditional causal inference literature, the *unit-level total effect* of changing $X = x'$ to $X = x$ on Y for a particular unit $\mathbf{U} = \mathbf{u}$,

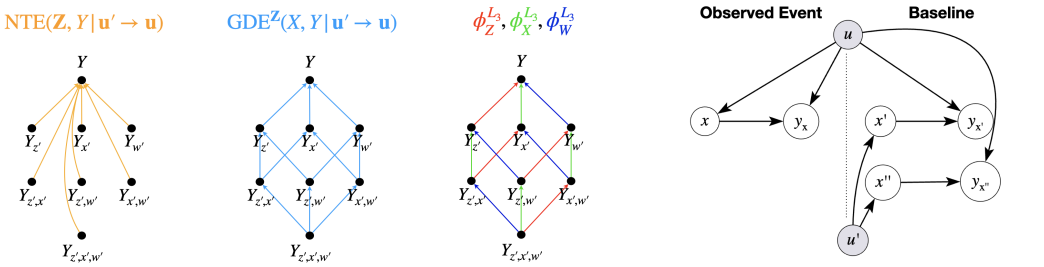
$$\mathrm{TE}_{x',x}(y|\mathbf{u}) = Y_x(\mathbf{u}) - Y_{x'}(\mathbf{u}), \tag{1}$$

describes how changing X affects Y [29, 32]. Indeed, there is consensus in the literature that if there is a non-zero unit level total effect from some $x' \rightarrow X(\mathbf{u}) = x$, then X causes Y [15, 29, 9, 3].[1]

There are three major drawbacks of this classical approach. First, the total effect only considers the variations of Y with respect to a singleton X, even though interventions on multiple variables at once are sometimes necessary to determine an events causal history.[2] Second, when comparing an observed event $Y = y$, with a potential cause $X = x$, against an alternative value x', there is a need for a principled method to select such a baseline value x'. Third, and more challengingly, having

[1] For example, if a lightning strike ($X = 1$) ignites a forest ($Y = 1$), one can argue that the lightning caused the fire because it had a non-zero total effect $Y_{X=1}(\mathbf{u}) - Y_{X=0}(\mathbf{u}) = 1$ (App. B.2).

[2] For example, if it rains and a sprinkler waters the grass, both rain and sprinkler are causes of the grass being wet. However, neither has a total effect on the outcome, since changing either event alone will not affect the outcome. (App. B.3).



(a) Explanatory bases: natural total effects (orange arrows, NTE, Def. 2), generalized direct effects (blue arrows, GDE, Def. 3) and counterfactual Shapley values (tricolor arrows, L_3 SVs, Def. 4) in a setting with three explanatory variables.

(b) The NTE, averaging over effects $Y_{X(\mathbf{u})}(\mathbf{u}) - Y_{X(\mathbf{u}')}(\mathbf{u})$ with natural baseline $\mathbf{u}' \sim P(\mathbf{U})$.

Figure 2: Visual representations of counterfactual quantities developed in this work.

selected a baseline x', the total effect does not reflect the normality or abnormality of the cause $X = x$, even though this information is needed to evaluate its importance relative to other causes [18].[3] To address these challenges in capturing the causal history [21] of an event, we introduce the *natural total effect*.

Definition 2 (Natural Total Effect (NTE))**.** *The unit-level natural total effect of a set of variables* $\mathbf{Z} \subseteq \mathbf{X}$ *on outcome* Y*, transitioning from baseline unit* $\mathbf{U} = \mathbf{u}'$ *to actual unit* $\mathbf{U} = \mathbf{u}$*, is defined as*

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

In words, the NTE *captures the effect of a set of variables* $\mathbf{Z}(\mathbf{u})$ *on* $Y(\mathbf{u})$ *with respect to a baseline* $\mathbf{Z}(\mathbf{u}')$ *determined by the natural value of* $\mathbf{Z}$ *under baseline unit* $\mathbf{U} = \mathbf{u}'$*.* □

The NTE (Fig. 2a, left) differs from previous methods in two key ways, addressing the drawbacks above. First, the NTE is a *multivariate* extension of the total effect, which allows one to capture causes that require intervention on multiple variables to observe a change in the outcome. We prove that this extension is necessary to allow the NTE to capture the causes of an event fully (App. B.5 contains the full formal statement with a proof):

Theorem 1 (Causal Necessity and Sufficiency of the NTE (informal))**.** *Given an SCM* $\mathcal{M}$ *and why query* $\text{Why}(y|\mathbf{x})$*, define the explanatory basis as*

$$\mathcal{B}(\mathcal{M}, w) = \{\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X}; \mathbf{u}', \mathbf{u} \in \mathcal{D}_{\mathbf{U}}; \mathbf{X}(\mathbf{u}) = \mathbf{x}, Y(\mathbf{u}) = y\}, \tag{3}$$

the set of all natural total effects. Then, $\mathcal{B}(\mathcal{M}, w)$ *is sufficient to ascertain the causes of* Y*. In addition, each* NTE $b \in \mathcal{B}(\mathcal{M}, w)$ *is necessary to distinguish some* $\mathcal{M}$ *from another SCM* $\mathcal{M}'$ *that has identical settings of* $\mathcal{B}(\mathcal{M}, w) \setminus \{b\}$*, but for which* $Y = y$ *has a different causal history.* □

Second, because baseline $\mathbf{Z}(\mathbf{u}')$ is determined by unit $\mathbf{u}'$, the NTE can be used to capture the human preference for abnormal causes in an explanation by constructing a *natural baseline* $\mathbf{u}' \sim P(\mathbf{U})$ (Fig. 2b), which selects baseline $X(\mathbf{u}')$ from its natural distribution,

$$\text{NTE}(\mathbf{Z}, Y \mid \mathbf{u}) = \mathbb{E}_{\mathbf{u}' \sim P(\mathbf{U})}[Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})]. \tag{4}$$

Finally, we note that since, in practice, the value of $\mathbf{U}$ is unobserved, unit-level NTEs cannot be estimated, despite their conceptual usefulness. For concreteness, we will instead work with NTEs conditional on observed variables $\mathbf{v}$, or the $\mathbf{v}$-specific NTE,

$$\text{NTE}(\mathbf{Z}, Y \mid \mathbf{v}) = \mathbb{E}_{\mathbf{u} \sim P(\mathbf{U}), \mathbf{u}' \sim P(\mathbf{U})}[Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u})]. \tag{5}$$

Having established that NTE fully captures an event's causal history, we are faced with one final challenge regarding complexity: the cardinality of the set of all $\mathbf{v}$-NTEs is exponential in the number of explanatory variables $|\mathbf{X}|$. We address this issue in the next section by decomposing NTEs into the contributions of their individual variables $X \in \mathbf{Z}$ and summarizing these contributions.

[3] For example, oxygen's presence is technically a cause of a hypothetical forest fire, but the probability of the baseline (the absence of oxygen) is so low that this cause would be considered less important to communicate relative to other causes, such as the fact that lightning struck a tree in the forest (App. B.4).

2.2 Univariate Contributions to Multivariate Total Effects

To quantify the contribution of a single variable X to the effect $\mathrm{NTE}(\{X\} \cup \tilde{\mathbf{Z}}, Y \mid \mathbf{u}' \to \mathbf{u})$ for some $\tilde{\mathbf{Z}} \subseteq \mathbf{V}$, we compare the effect of $\{X\} \cup \tilde{\mathbf{Z}}$ on Y to the effect of $\tilde{\mathbf{Z}}$ alone on Y. We call this quantity the the *generalized direct effect (GDE)*.

Definition 3 (Generalized Direct Effect (GDE))**.** *The unit-level generalized direct effect of a single variable $X \in \mathbf{V}$ on outcome Y, adjusting for $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$ and transitioning from baseline unit $\mathbf{U} = \mathbf{u}'$ to actual unit $\mathbf{U} = \mathbf{u}$, is defined as*

$$\mathrm{GDE}^{\mathbf{z}}(X, Y \mid \mathbf{u}) = \mathrm{NTE}(\{X\} \cup \mathbf{Z}, Y \mid \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y \mid \mathbf{u}). \tag{6}$$

Analogously, the GDE *with natural baseline $\mathbf{u}'$ is defined by averaging over $\mathbf{u}' \sim P(\mathbf{U})$, and the $\mathbf{v}$-specific* GDE *is defined by conditioning on $\mathbf{v}$ instead of $\mathbf{u}$.* □

In words, the GDE (Fig. 2a, middle) captures the contribution of $X(\mathbf{u})$ to the effect of $\mathbf{Z}(\mathbf{u})$ on $Y(\mathbf{u})$, with respect to a baseline $\mathbf{Z}(\mathbf{u}')$ under baseline unit $\mathbf{U} = \mathbf{u}'$. Armed with this notion, we next show how one can decompose a multivariate total effect represented by the NTE into contributions attributable to single variables:

Theorem 2 (NTE decomposition)**.** *For any permutation π over the elements of potential cause $\mathbf{Z} \subseteq \mathbf{X}$ of Y, the following decomposition holds:*

$$\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{v}) = \sum_{X_i \in \mathbf{Z}} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{v}), \tag{7}$$

where $\pi_{<i}$ denotes the set of variables prior to X_i in the permutation π. □

In words, Thm. 2 shows that GDEs disentangle the contributions of each $X_i \in \mathbf{Z}$ to the multivariate effect of $\mathbf{Z}$ on Y, $\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{v})$ (see Fig. 2a, left and middle). Notably, the $\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{v})$ decomposition is shown for an arbitrary permutation of elements in $\mathbf{Z}$. However, different permutations can change the result Y in different ways, and therefore an exhaustive approach is to consider the contribution of X_i to an NTE averaged in all possible permutations. This results in our explanation method, the counterfactual Shapley value, which averages GDEs using Shapley weights [34].

Definition 4 (Counterfactual Shapley value (L_3 SV))**.** *Consider why query $w = \mathrm{Why}(y|\mathbf{x})$ such that $\mathbf{v} = \mathbf{x} \cup \{y\}$. Then, the counterfactual Shapley value for $X \in \mathbf{X}$ is defined as*

$$\phi_X^{L_3}(w) = \mathbb{E}_{\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})} \left[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v}) \right], \tag{8}$$

where $\Pi_{\mathbf{X}}$ denotes the set of orderings on $\mathbf{X}$, and $\pi_{<X}$ denotes the variables prior to X in π. □

Intuitively, counterfactual Shapley values summarize all variable-specific attributions to effects of subsets of $\mathbf{X}$ on Y, as captured by the GDE.

Corollary 3 (ϕ^{L_3} decomposes NTE)**.** *L_3 Shapley values decompose the NTE of all variables $\mathbf{X}$ on Y as follows:*

$$\mathrm{NTE}(\mathbf{X}, Y|\mathbf{v}) = \sum_{X \in \mathbf{X}} \phi_X^{L_3}(\mathrm{Why}(y|\mathbf{x})). \tag{9}$$

□

This result follows from Thm. 2 and shows that our explanation method ϕ^{L_3} decomposes the NTE and accounts for all the variations appearing in it, thus capturing the causal history of the event explanandum while providing a parsimonious single attribution for each variable.

3 Explanatory Variable Attributions: Properties & Desiderata

Having developed a causally consistent explanation method, counterfactual Shapley values (L_3 SVs), we now compare it to existing methods in the literature. To do so, we first define a general data structure for explanations: the explanatory variable attribution (EVA). [4]

[4] For concreteness, we limit our discussion to variable attributions. In App. C, we discuss generalized variable attributions, an overview of prior methods, and properties they satisfy and violate.

Definition 5 (Explanatory Variable Attribution (EVA))**.** *An explanatory attribution is a mapping $\phi : \Omega \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}^n$, where Ω is the set of SCMs, $\mathbf{X}$ is the set of explanatory variables for why query $w \in \mathbb{W}$, and attribution dimensionality $n \in \mathbb{Z}^+$. We denote the space of EVAs outputting vectors in $\mathbb{R}^n$ as Φ_n. When unambiguous denote X as a subscript in $\phi_X(w)$.* □

Next, we formalize properties that represent desirable features of an explanation and allow ~~the~~ comparison to other methods. Finally, we will show that L_3 SVs satisfy all desiderata, unlike any prior work found in the literature. The first property, *causal admissibility*, follows a simple intuition: if a variable could not possibly have a causal effect on the outcome, given our observed knowledge, then it should be assigned a zero attribution $\phi_i(w) = \mathbf{0}$. In other words, an attribution should not incorrectly tell us that a variable is a cause of the outcome.

Property 1 (Causal Admissibility)**.** *Consider why query $w = \text{Why}(y|\mathbf{X}; \mathbf{e}) \in \mathbb{W}$. Let there be an absence of causation from $X_i \in \mathbf{X}$ to Y in consistent worlds if for all units $\mathbf{u} \in \mathcal{D}_\mathbf{U}$, there is no setting $\mathbf{z}' \in \mathcal{D}_{\mathbf{X}\setminus\{X_i\}}$ and $x_i' \in \mathcal{D}_{X_i}$ such that*

$$\mathbf{E}(\mathbf{u}) = \mathbf{e} \wedge Y_{\mathbf{z}', x_i'}(\mathbf{u}) \neq Y_{\mathbf{z}', x_i}(\mathbf{u}). \tag{10}$$

In words, there is an absence of causation from X_i to Y if it is impossible to change Y by changing X_i, under any unit consistent with the observations. Then, EVA $\phi_n \in \Phi$ is causally admissible if for all why queries w and variables $X_i \in \mathbf{X}$, the absence of causation from X_i to Y in consistent worlds implies a zero attribution $\phi_i(w) = \mathbf{0}$. □

However, satisfying admissibility does not imply a complete explanation. For instance, $\phi(w) = \mathbf{0}$ trivially satisfies admissibility, while giving no information about the causes of the event. To require informative explanations, we introduce the property of *causal explanatory power*, which requires a non-zero attribution whenever an admissible attribution might detect causation.

Property 2 (Causal explanatory power)**.** *EVA $\phi \in \Phi_n$ satisfies strong causal explanatory power if for every causally admissible EVA $\phi' \in \Phi_n$, all why queries $w \in \mathbb{W}$, and explanatory variable $X_i \in \mathbf{X}$,*

$$\phi_i'(w) \neq \mathbf{0} \implies \phi_i(w) \neq \mathbf{0} \tag{11}$$

ϕ satisfies weak causal explanatory power if the above criterion holds almost surely over the space of SCMs $\mathcal{M} \in \Omega$. □

In words, an admissible attribution has causal explanatory power if it is one of the most informative: it must inform us of causation by yielding a non-zero attribution whenever any other admissible method yields a non-zero attribution.

Finally, we formalize a property related to *normality*, following the work in cognitive psychology by the Nobel Prize awardee Daniel Kahneman [18]. His primary observation is that humans tend to prefer abnormal causes over normal ones when explaining an event: "the affective response to an event is enhanced if its causes are abnormal [...] A cause must be an event that could easily have been otherwise." Intuitively, the more abnormal a positive effect of X on Y is, the higher the attribution the variable should receive; the more abnormal a negative effect, the lower the attribution should be.[5]

Property 3 (Causal Normality)**.** *EVA $\phi \in \Phi_n$ satisfies causal normality if, given why query $w \in \mathbb{W}$ and valid corresponding SCMs $\mathcal{M}_1, \mathcal{M}_2$ with identical observed variables $\mathbf{V}$, when some $X \in \mathbf{V}$ is a more of an abnormal cause in the positive direction in $\mathcal{M}_1$ than $\mathcal{M}_2$ but affects Y identically in both SCMs, $\phi_X(\mathcal{M}_1, w) > \phi_X(\mathcal{M}_2, w)$.* □

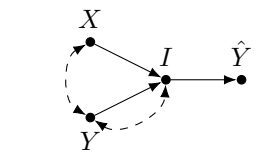
Finally, we observe that L_3 SVs may be viewed as an explanatory variable attribution and prove that they satisfy all three properties.

Theorem 4 (L_3 SVs satisfy desiderata)**.** *L_3 SVs satisfy causal admissibility, weak causal explanatory power, and normality.* □

4 Experiments

In this section, we evaluate counterfactual Shapley values (L_3 SVs) in practice, aiming to determine whether they provide more intuitive feature attributions than existing methods. We support the claim

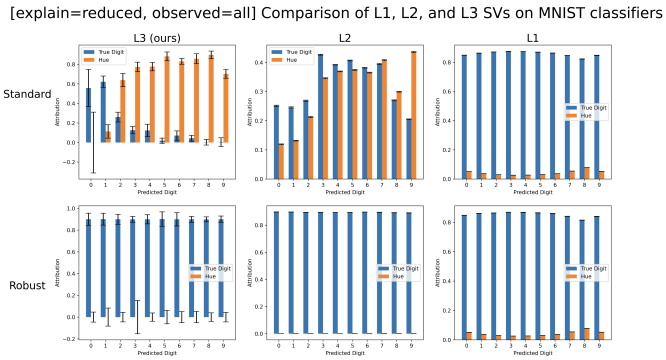
[5] See App. C.4 for a formal definition of greater abnormal causation in the positive direction, counterexamples illustrating violations of normality, and further discussion of normality.



(a) Causal diagram.

(b) Real digits.

(c) Shifted digits.



(d) Variable attributions for MNIST.

Figure 3: 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.

that our method yields more intuitive feature attributions by comparing it to others on two vision datasets: color MNIST [6, 28] (Sec. 4.1) and CelebA-HQ [19] (Sec. 4.2). Experimental details and additional experiments on toy examples and on a synthetic dataset are included in App. E.[6] We compare against the methods most similar to ours in the literature, observational Shapley values [25] and interventional Shapley values [16, 12, 17], which, from here on, will be called L_1 and L_2 Shapley values, respectively, given that they are related to the first and second layers of the Pearl Causal Hierarchy.

4.1 Color MNIST

Consider a dataset of images of colored digits, inspired by [28]. There are two dimensions of variation for each image in the dataset: the hue of the digit X, and the digit itself Y; the causal diagram is shown in Fig. 3a. There exists a strong correlation between the hue and the digit, as shown in samples from the dataset Fig. 3b. Furthermore, the digit and image I are spuriously confounded: lower digits have lower saturation, while higher digits have higher saturation, to the extent that zeros in the dataset are entirely white.

Say Alice and Bob are training basic convolutional digit classifiers $\hat{Y}$ on this dataset. Bob trains his classifier in an entirely standard fashion and obtains $\hat{Y}^S$. On the other hand, Alice is more perceptive, and is concerned about the spurious correlation between the hue of the digit and the outcome. She then converts her images to grayscale in a preprocessing step to make them robust to color shift before training her convolutional classifier $\hat{Y}^R$. The S and R stand for "standard" and "robust," respectively.

Now, Alice and Bob want to compare their classifiers and plan to use explanation techniques available in the literature. A priori, Alice expects that the robust classifier will ignore hue and considers only the digits themselves in its prediction. In addition, she expects Bob's standard classifier to ~~always~~ use both hue and digit in its prediction, except for zero digits, which are white and therefore unaffected by changing hues; she expects only the digit to be used when considering zero digits. For example, on the shifted-hue digits in Fig. 3c, she would expect her classifier to perform well and Bob's to perform poorly. When Alice tries to use L_1 Shapley values [25], a popular method in the literature, she finds to her dismay that the classifiers' attributions are identical, as shown on the right side of Fig. 3d: the top and bottom plots are identical.

Using L_2 Shapley values [17] (Fig. 3d, middle), she is able to distinguish her robust classifier from Bob's standard classifier: the attribution always yields a zero attribution to hue for her robust classifier (bottom middle plot), and a positive attribution for Bob's standard classifier (top middle plot). However, she is still puzzled when she examines attributions to the zero digit for Bob's standard classifier, which should be zero but are not.

[6] We provide code at `https://anonymous.4open.science/r/causal-explanation-framework/`.

Finally, Alice decides to use counterfactual Shapley values (Fig. 3d, left), which are able to correctly assign zero to the hue variable for all robust classifier predictions (bottom left plot), distinguishing her classifier from Bob's (top left plot). In addition, the bounds for the zero digit always contain zero: for Bob's standard classifier, the attribution method correctly assigns no attribution to hue for zero digits (top left plot). Alice decides that of the three methods she used, only L_3 Shapley values behave exactly as expected based on her intuition.

Given the experiments above, we conclude that L_3 SVs are superior to L_1 SVs in that they can distinguish between ML models with different behavior. Furthermore, L_3 SVs are superior to L_2 SVs in that they correctly condition on the observed information, rather than averaging over all units; in this experimental set, this behavior manifests in allowing L_3 SVs to correctly identify variables without an effect on the outcome in every setting.

4.2 CelebA-HQ

We next consider an example based on the CelebA-HQ dataset [19], including two ML engineers called Bob and Alice. They are interested in constructing classifiers for predicting whether a person has their mouth open (labeled M), based on the CelebA-HQ data labeled $\mathcal{D}$. After this, they wish to use explanation techniques to see if the constructed classifiers align with the human intuition. Bob, eager to get results quickly, constructs a classifier using a LeNet-style model [20] to predict the label M using the dataset $\mathcal{D}$, labeled I, thereby constructing a classifier $\hat{M}^B$. Alice, being a more experienced engineer, is concerned about her classifier using spurious correlations. She knows that, often, the mouth being open may be caused by the person smiling (variable S), and this information is also captured in the image I (see causal diagram in Fig. 4). Thus, she wants to avoid her classifier using information on smiling, and for this reason uses a pre-processing technique: she re-weighs the data $\mathcal{D}$, so as to make the variables M and S probabilistically independent. In this new reweighted dataset, $\mathcal{D}^{rw}$, she constructs a LeNet classifier for M based on I, labeled $\hat{M}^A$.

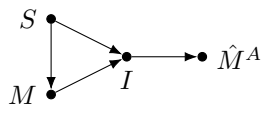
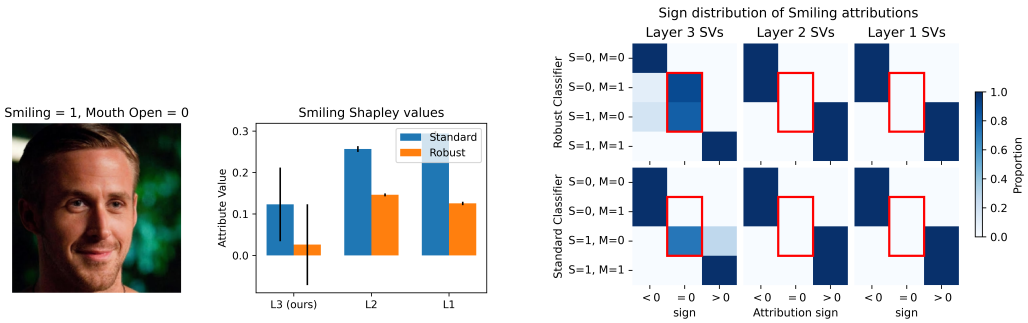


Figure 4: CelebA-HQ causal diagram.

After constructing their classifiers, Bob and Alice evaluate them on a held-out part of data $\mathcal{D}$, and to see what explanation techniques can tell them about $\hat{M}^A, \hat{M}^B$. They are interested in the why query $\text{Why}(\hat{m}|m, s)$, ~~that is~~ the explanation of the classifiers $\hat{M}^A, \hat{M}^B$ based on the variables M, S. They start with a randomly sampled image shown in Fig. 5a, of Ryan Gosling, who is smiling ($S = 1$) but has his mouth closed ($M = 0$), and select L_1, L_2, and L_3 Shapley values as the explanation techniques ~~they will~~ to use. They first look at Alice's classifier $\hat{M}^A$, which was trained in the setting where M, S are uncorrelated. Since $S = 1, M = 0$ for the given image, they realize that S had no effect on M, since smiling can only positively affect the mouth being open, which was not the case since $M = 0$. Therefore, they conclude that for the given image, S must have had no effect on $\hat{M}^A$, neither through the influence $S \to I \to \hat{M}^A$ (due to training setting), nor through $S \to M \to I \to \hat{M}^A$ (since $S = 1, M = 0$ implies no influence $S \to M$). They then inspect the attributions for S, M by L_1 and L_2 Shapley values – and find, to their surprise, that the variable S is given a strictly positive attribution by both methods (see Fig. 5a middle and right orange bars). Given that they feel that Ryan Gosling's closed mouth was not due to his smile, they examine the L_3 Shapley values, finding that the smiling variable S has an attribution indistinguishable from 0, aligned with their intuition.

They then move to open up Bob's classifier $\hat{M}^B$. Bob is somewhat embarrassed by the fact that he did not construct a clever, robust classifier like Alice. They note that in his classifier, an influence $S \to I \to \hat{M}^B$ should exist, therefore expecting a positive attribution even in the setting where $S = 1, M = 0$. They observe such positive attributions to S for L_1, L_2 Shapley values (Fig. 5a, middle and right blue bars). However, based on their experience of being misled when looking at $\hat{M}^A$, they no longer trust these – and note that L_1, L_2 Shapley values cannot qualitatively distinguish $\hat{M}^A, \hat{M}^B$, even though these were trained in a very different way. Luckily, they once again check L_3 Shapley values, and also find a positive attribution to S, aligned with their expectation and that allows them to distinguish the classifiers.

After being convinced that L_3 Shapley values performed a better job at explaining their classifiers than L_1, L_2 Shapley values on the particular instance, they wish to understand if this was a coincidence on



(a) (b)

Figure 5: CelebA-HQ experiments. (a) Left: Why do we predict Ryan Gosling's mouth is closed? Surely not because of or despite his smile. Right: L_3, L_2, and L_1 Shapley values computed for the Smiling variable with respect to the standard and robust classifiers. Error bars denote SEM. (b) Shapley value sign distribution for Smiling on correctly-classified instances. Areas in which the standard and robust classifiers are expected to differ are boxed in red. Standard and robust classifiers' L_3 distributions differ with $p < 0.001$.

a single sample or a broader phenomenon about the methods. For this, they compute Shapley values on 20 samples within each of the four classes $S = s, M = m$ and check if each estimated attribution is positive, negative, or indistinguishable from zero, with a two-tailed z-test with a significance of 0.05. Considering their expectation that L_1 and L_2 SVs are unable to qualitatively distinguish between their classifiers, Alice and Bob first compute L_1 and L_2 Shapley values and observe that, indeed, the sign distributions conditional on each class Fig. 5b, illustrate that, as expected, L_1 and L_2 Shapley values are generally unable to distinguish between standard and robust classifiers in terms of attribution sign. Next, they compare their L_3 Shapley values' behavior with respect to their classifiers. They observe that for images with $S = 0, M = 1$, there is a significant ($p < 0.001$) difference between the sign distributions of the smiling attribution, and they are therefore able to distinguish their classifiers' behavior in this way. Interestingly, for images with $S = 0, M = 1$ and $S = 1, M = 0$, where S is expected to have no effect on the Alice's classifier and a non-zero effect on Bob's classifier, standard and robust classifiers are distinguishable ($p < 0.001$); the attribution signs match with their expectations (see the boxes marked in red in the figure). Ending their investigation, they conclude that L_3 Shapley values' superior performance in comparing their classifiers was true in general on the CelebA-HQ dataset.

Summary. Experiments support the theoretical findings of this work, namely: L_3 Shapley values provide a fine-grained variable attribution method and satisfy the desiderata grounded in the sciences. For further experimental results, see App. E; for discussion on related literature, see App. C.2.

5 Conclusions

We introduced a precise mathematical formalization of the explanatory query and proposed the natural total effect, generalized direct effect, and counterfactual Shapley values to summarize the causal history of an event. We articulated explanatory desiderata grounded in key insights from philosophy, cognitive science, and psychology – such as causal admissibility, explanatory power, and normality – and demonstrated that our method is the first to satisfy them. Synthetic and semi-synthetic experiments corroborate our findings. We expect this to be a first step toward a new generation of causally grounded explanation methods, which are increasingly needed at a time when AI systems are pervasive and only expected to grow. In terms of future research directions, one aspect that warrants further investigation is the scalability of the methods proposed here. There is an intricate relationship between the proposed quantity, its identifiability status, and its evaluability from observational data, which makes this task both computationally and statistically challenging. Still, since causality is an indispensable ingredient of any possible explanation framework, the only viable path is to delve deeper into these challenges, which we believe we have made significant strides.

Acknowledgements

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

References

[1] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. *On Pearl's Hierarchy and the Foundations of Causal Inference*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501743.

[2] S. 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.

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

[4] L. I. I. Cornell Law School. but-for test. https://www.law.cornell.edu/wex/but-for_test, 2023. Accessed: 2023-3-16.

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

[6] L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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

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

[9] J. Y. Halpern. *Actual causality*. MiT Press, 2016.

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

[11] J. Harding, T. Gerstenberg, and T. Icard. A communication-first account of explanation. *arXiv preprint arXiv:2505.03732*, 2025.

[12] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] X. Huang and J. Marques-Silva. On the failings of shapley values for explainability. *International Journal of Approximate Reasoning*, 171:109112, 2024.

[15] D. Hume. A treatise of human nature. Clarendon Press, 1739.

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

[17] 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.

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

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[22] D. Lewis. Causal explanation. In D. Lewis, editor, *Philosophical Papers Vol. Ii*, pages 214–240. Oxford University Press, 1986.

[23] Z. 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.

[24] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. 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.

[25] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[26] T. 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.

[27] C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[28] Y. Pan and E. Bareinboim. Counterfactual image editing. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. 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.

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

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

[31] J. Pearl and D. Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[32] D. Plecko and E. Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

[33] M. T. Ribeiro, S. Singh, and C. 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.

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

[35] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

[36] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[37] L. Ter-Minassian, O. Clivio, K. Diazordaz, R. J. Evans, and C. C. Holmes. PWSHAP: A path-wise explanation model for targeted variables. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34054–34089. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/ter-minassian23a.html.

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

[39] J. Wang, J. Wiens, and S. Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2021.

[40] K. M. Xia, Y. Pan, and E. Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2022.

Appendices

A Proofs

Assumptions. In all theorems, we make the following assumptions regarding well-behaved SCMs: 1) $P(\mathbf{U})$ is positive on its domain $\mathcal{D}_{\mathbf{U}}$ and continuous with respect to each continuous variable $U \in \mathbf{U}$; 2) for every $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there exists $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ such that $\mathbf{V}(\mathbf{u}) = \mathbf{v}$; 3) the arguments to every causal mechanism $f_i(\mathbf{pa}_i, \mathbf{u}_i)$ can be separated into continuous and discrete parents $\mathbf{X}_i^c, \mathbf{X}_i^d$ such that $\mathbf{X}_i^c \cup \mathbf{X}_i^d = \mathbf{Pa}_i \cup \mathbf{U}_i$, and $f_i(\cdot, \mathbf{x}_i^d)$ is continuous with respect to its continuous parents $\mathbf{X}_i^c$ for every setting of its discrete parents $\mathbf{x}_i^d \in \mathcal{D}_{\mathbf{X}_i^d}$.

A.1 NTE is necessary and sufficient to capture causal history

See Apps. B.5 and D for further discussion of the results below.

Definition 6 (Event Explanatory Basis)**.** *The event basis of SCM $\mathcal{M}$ with respect to query* $\mathrm{Why}(y|\mathbf{X}; \mathbf{e}' \to \mathbf{e})$ *may be written as $\mathcal{C}(\mathcal{M}, w) := \langle Y_{\mathbf{X}^*}, \mathbf{X}, P(\mathbf{U}|\mathbf{e}'), P(\mathbf{U}|\mathbf{e}) \rangle$, where $Y_{\mathbf{X}^*}$ is the event counterfactual basis, defined as:*

$$Y_{\mathbf{X}^*}(\mathbf{u}) := \{Y_{\mathbf{z}}(\mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X}, \mathbf{z} \in \mathcal{D}_{\mathbf{Z}}\}. \tag{12}$$

Note that $\mathbf{X}$ above denotes the function $\mathbf{X}(\mathbf{u}) := \{X(\mathbf{u}) : X \in \mathbf{X}\}$. □

Definition 7 (Natural Total Effect (NTE) Basis)**.** *Consider SCM $\mathcal{M}$,* Why *query w, event counterfactual basis Y_*, and explanatory variable subset $\mathbf{Z} \subseteq \mathbf{X}$. The unit-level natural total effect of $\mathbf{Z}$ on Y with respect to baseline $\mathbf{u}'$ and knowledge $\mathbf{u}$ is defined as*

$$\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}). \tag{13}$$

Assume that for any $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there is $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ inducing $\mathbf{V}(\mathbf{u}') = \mathbf{v}$. Then the NTE *basis of $\mathcal{M}$ for why query w is defined as*

$$\mathcal{B}_{\mathrm{NTE}}^{\mathcal{M},w}(\mathbf{u}) := \{\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X}, \mathbf{u}' \in \mathcal{D}_{\mathbf{U}}\}. \tag{14}$$

□

Lemma 1 (NTE basis equivalence)**.** *Consider SCM $\mathcal{M}$ and why query w. Assume that for every setting of observed variables $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there exists unobserved setting $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ such that $\mathbf{V}(\mathbf{u}) = \mathbf{v}$. Then, the* NTE *basis $\mathcal{B}_{\mathrm{NTE}}^{\mathcal{M},w}$ and the value $Y(\mathbf{u})$ uniquely determine and are uniquely determined by the event counterfactual basis $Y_{\mathbf{X}^*}$.* □

Proof. Consider some actual $\mathbf{u}$. We will show that there is a bijective mapping from $Y_{\mathbf{X}^*}(\mathbf{u})$ to $\{Y(\mathbf{u}), \mathcal{B}_{\mathrm{NTE}}^{\mathcal{M},w}(\mathbf{u})\}$. For the forward direction, we examine the definition of the NTE,

$$\begin{aligned} \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) &= Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) && (15) \\ &= Y(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}). && (16) \end{aligned}$$

Since the NTE is defined using counterfactuals on Y, it is identified by the event counterfactual basis. To prove the converse, we select $\mathbf{u}'$ to induce $\mathbf{Z}(\mathbf{u}') = \mathbf{z}'$. Then, we can compute any counterfactual $Y_{\mathbf{z}'}$ as

$$Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) = Y(\mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}). \tag{17}$$

This concludes our proof. □

Lemma 2 (NTE necessity)**.** *Consider SCMs $\mathcal{M}_A, \mathcal{M}_B$, described by the following functions, for $C \in \{A, B\}$ and $n \in \mathbb{Z}^+$, $m \in \mathbb{Z}^{\geq 0}$*

$$\mathcal{F}_C(n,m) = \begin{cases} Z_i & = u_Z^i & \forall i \in [n] \\ W_j & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0] & C = A \wedge u_W^j = 2 \wedge j = 1 \\ 1 & C = B \wedge u_W^j = 2 \wedge j = 1 \\ W_{j-1} & u_W^j = 2 \wedge j > 1 \\ u_{W^j} & u_W^j \in \{0,1\} \end{cases} & \forall j \in [m] \\ Y & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0 \vee W_m = 1] & m > 0 \\ \mathbf{1}[\mathbf{Z} \neq 0] & m = 0 \end{cases} \end{cases} \tag{18}$$

and why query $w = \text{Why}(Y = 1 | \mathbf{Z} = \mathbf{1}_n, \mathbf{W} = \mathbf{1}_m)$. Every element of the NTE *basis $\mathcal{B}_{\text{NTE}}^{\mathcal{M},w}(\mathbf{u})$ for units $\mathbf{u} \in \mathcal{D}_\mathbf{U}$ consistent with observed events is necessary to distinguish between $\mathcal{M}_A(n,m)$ and $\mathcal{M}_B(n,m)$ under at least one setting of n, m.* □

Proof. We consider the actual world $\mathbf{U} = \{\mathbf{U}_\mathbf{Z} = \mathbf{1}_n\}$. First, we note that $Y^A_{\mathbf{Z}=\mathbf{0}_n}(\mathbf{u}) = 0$, while $Y^B_{\mathbf{Z}=\mathbf{0}_n}(\mathbf{u}) = 1$. Next, we note that for interventions that do not set $\mathbf{Z} = \mathbf{0}_n$, $Y = 1$ in both SCMs. Finally, we note that if $\mathbf{Z} = \mathbf{0}_n$ is in the intervention, in addition to at least one $W_j \in \mathbf{W}$ that is the last in the chain to be intervened upon, $Y = 0$ if we set $W_j = 0$, and $Y = 1$ if we set $W_j = 1$. Therefore, the counterfactual $Y_{\mathbf{Z}=\mathbf{0}_n}$ is necessary within the event counterfactual basis to distinguish the two SCMs, as all other counterfactuals are identical between $\mathcal{M}_A, \mathcal{M}_B$. Thus, $\text{NTE}(\mathbf{Z}, Y | \mathbf{u}' \to \mathbf{u})$ for all $\mathbf{Z}(\mathbf{u}') = \mathbf{0}_n$ is necessary to distinguish the two SCMs, given that only this NTE is a function of $Y_{\mathbf{Z}=\mathbf{0}_n}(\mathbf{u})$, and all other counterfactuals are identical between the two SCMs. □

Theorem 5 (NTE necessity and sufficiency)**.** *Consider SCM $\mathcal{M}$ and why query $w = \text{Why}(y|\mathbf{x})$. Every element of the* NTE *basis $\mathcal{B}_{\text{NTE}}^{\mathcal{M},w}(\mathbf{u})$ for actual units $\mathbf{u} \in \mathcal{D}_\mathbf{U}$ consistent with observed events is necessary to distinguish between $\mathcal{M}_A(n,m)$ and $\mathcal{M}_B(n,m)$ under at least one setting of n, m. Furthermore, the* NTE *basis is sufficient to describe the causal history of $Y = y$ contained in the event counterfactual basis $\mathbf{Y}_{\mathbf{X}^*}$.* □

Proof. This result follows directly from Lems. 1 and 2. □

A.2 NTE decomposes into GDEs and L_3 SVs

Theorem 2 (NTE decomposition)**.** *For any permutation π over the elements of potential cause $\mathbf{Z} \subseteq \mathbf{X}$ of Y, the following decomposition holds:*

$$\text{NTE}(\mathbf{Z}, Y | \mathbf{v}) = \sum_{X_i \in \mathbf{Z}} \text{GDE}^{\pi_{<i}}(X_i, Y | \mathbf{v}), \tag{7}$$

where $\pi_{<i}$ denotes the set of variables prior to X_i in the permutation π. □

Proof. We prove this theorem by induction. For $|\mathbf{Z}| = 0$, and any $X \in \mathbf{X}$, we observe that

$$\begin{aligned} \text{GDE}^{\varnothing}(X, Y | \mathbf{u}' \to \mathbf{u}) &= \text{NTE}(X, Y | \mathbf{u}' \to \mathbf{u}) - \text{NTE}(\varnothing, Y | \mathbf{u}' \to \mathbf{u}) && (19) \\ &= \text{NTE}(X, Y | \mathbf{u}' \to \mathbf{u}) - (Y(\mathbf{u}) - Y(\mathbf{u})) && (20) \\ &= \text{NTE}(X, Y | \mathbf{u}' \to \mathbf{u}). && (21) \end{aligned}$$

For the inductive step, we assume that for every $|\mathbf{Z}| < k < |\mathbf{X}|$ and ordering π on $\mathbf{Z}$,

$$\text{NTE}(\mathbf{Z}, Y | \mathbf{u}' \to \mathbf{u}) = \sum_{X_i \in \mathbf{Z}} \text{GDE}^{\pi_{<i}}(X_i, Y | \mathbf{u}' \to \mathbf{u}). \tag{22}$$

Consider subset $\mathbf{Z}' \subseteq \mathbf{X}$ with $|\mathbf{Z}'| = k > 0$. Let X be the last element of $\mathbf{Z}'$ according to ordering π on $\mathbf{Z}'$, and denote $\mathbf{Z} = \mathbf{Z}' \setminus \{X\}$. Note that $|\mathbf{Z}| = k - 1$, satisfying Eq. (22). Then, following Def. 3:

$$\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = \mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) \tag{23}$$

$$\mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) = \mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) + \mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) \tag{24}$$

$$= \sum_{X_i \in \mathbf{Z}} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{u}' \to \mathbf{u}) + \mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) \tag{25}$$

$$= \sum_{X_i \in \mathbf{Z} \cup \{X\}} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{u}' \to \mathbf{u}) \tag{26}$$

$$\mathrm{NTE}(\mathbf{Z}', Y|\mathbf{u}' \to \mathbf{u}) = \sum_{X_i \in \mathbf{Z}'} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{u}' \to \mathbf{u}) \tag{27}$$

We have proven Eq. (22) by induction. We now take the expectation over $\mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U}|\mathbf{v})$ on both sides to obtain

$$\mathbb{E}_{\mathbf{u}',\mathbf{u}}\left[\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u})\right] = \mathbb{E}_{\mathbf{u}',\mathbf{u}}\left[\sum_{X_i \in \mathbf{Z}} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{u}' \to \mathbf{u})\right] \tag{28}$$

$$\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{v}) = \sum_{X_i \in \mathbf{Z}} \mathrm{GDE}^{\pi_{<i}}(X_i, Y|\mathbf{v}) \tag{29}$$

We conclude our proof. □

Corollary 3 (ϕ^{L_3} decomposes NTE)**.** *L_3 Shapley values decompose the NTE of all variables $\mathbf{X}$ on Y as follows:*

$$\mathrm{NTE}(\mathbf{X}, Y|\mathbf{v}) = \sum_{X \in \mathbf{X}} \phi_X^{L_3}(\mathrm{Why}(y|\mathbf{x})). \tag{9}$$

□

Proof. This corollary follows from Thm. 2. We may write:

$$\mathrm{NTE}(\mathbf{X}, Y|\mathbf{v}) = \mathbb{E}_{\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})}\left[\mathrm{NTE}(\mathbf{X}, Y|\mathbf{v})\right] \sum_{X \in \mathbf{X}} \phi_X^{L_3}(w) \tag{30}$$

$$= \mathbb{E}_{\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})}\left[\sum_{X \in \mathbf{X}} \mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v})\right] \tag{31}$$

$$= \sum_{X \in \mathbf{X}} \mathbb{E}_{\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})}\left[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v})\right] \tag{32}$$

$$= \sum_{X \in \mathbf{X}} \phi_X^{L_3}(w). \tag{33}$$

We conclude our proof. □

A.3 L_3 SVs satisfy causal explanation properties

A.3.1 Causal Admissibility

Lemma 3 (Causal Admissibility)**.** *L_3 SVs satisfy causal admissibility.* □

Proof. First, we prove that causal admissibility holds for L_3 SVs. If $Y_{\mathbf{z},x'}(\mathbf{u}) = Y_{\mathbf{z}}(\mathbf{u})$ for all variables $X \in \mathbf{X}$ and interventions $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{z} \in \mathcal{D}_{\mathbf{Z}}$, then it is clear that

$$\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}'),X(\mathbf{u}')}(\mathbf{u}) \tag{34}$$

$$= Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \tag{35}$$

$$= 0. \tag{36}$$

This implies that the corresponding L_3 Shapley values must be zero:

$$\phi_X^{L_3}(w) = \mathbb{E}_{\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}})}\left[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v})\right] = 0. \tag{37}$$

Therefore, causal admissibility holds. □

A.3.2 Weak Causal Explanatory Power

Lemma 4 (Weak Causal Explanatory Power)**.** *L_3 SVs satisfy weak causal explanatory power.* □

Proof. Following the definition of weak causal explanatory power, consider causally admissible EVA $\phi' \in \Phi_n$, why query $w \in \mathbb{W}$, and explanatory variable $X_i \in \mathbf{X}$.

If $\phi'(w) \neq 0$, the contrapositive of causal admissibility implies there exists some $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ such that for some $\mathbf{z}' \in \mathcal{D}_{\mathbf{X}\setminus\{X_i\}}$ and $x_i' \in \mathcal{D}_{X_i}$,

$$\mathbf{X}(\mathbf{u}) = \mathbf{x} \wedge Y(\mathbf{u}) = y \wedge Y_{\mathbf{z}', x_i'}(\mathbf{u}) \neq Y_{\mathbf{z}', x_i}(\mathbf{u}). \tag{38}$$

Let $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ induce $\mathbf{Z}(\mathbf{u}') = \mathbf{z}'$ and $X_i(\mathbf{u}') = x_i'$. Then, by the above equation,

$$\mathrm{GDE}^{\mathbf{Z}}(X_i, Y | \mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u}'), X(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) \neq 0. \tag{39}$$

We have $\phi_i^{L_3}(w) = 0$ if and only if

$$\mathbb{E}_{\pi, \tilde{\mathbf{u}}', \tilde{\mathbf{u}}}[\mathrm{GDE}^{\pi_{<X}}(X, Y | \tilde{\mathbf{u}}' \to \tilde{\mathbf{u}})] = 0, \tag{40}$$

where $\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}}), \tilde{\mathbf{u}}' \sim P(\mathbf{U}), \tilde{\mathbf{u}} \sim P(\mathbf{U} | \mathbf{V} = \mathbf{v})$. Combining the above equations, we may rewrite our characteristic constraint as a weighted function of other GDEs.

$$\mathrm{GDE}^{\mathbf{Z}}(X_i, Y | \mathbf{u}' \to \mathbf{u}) = \sum_{\substack{\tilde{\mathbf{Z}} \subseteq \mathbf{X} \setminus \{X_i\} \\ \tilde{\mathbf{u}}, \tilde{\mathbf{u}}' \in \mathcal{D}_{\mathbf{U}}}} w(\tilde{\mathbf{Z}}, \tilde{\mathbf{u}}', \tilde{\mathbf{u}}) \, \mathrm{GDE}^{\tilde{\mathbf{Z}}}(X_i, Y | \tilde{\mathbf{u}}' \to \tilde{\mathbf{u}}) \neq 0 \tag{41}$$

where

$$w(\tilde{\mathbf{Z}}, \tilde{\mathbf{u}}', \tilde{\mathbf{u}}) := \frac{P(\pi, \tilde{\mathbf{u}}', \tilde{\mathbf{u}} | (\pi_{<X}, \tilde{\mathbf{u}}', \tilde{\mathbf{u}}) \neq (\mathbf{Z}, \mathbf{u}', \mathbf{u}))}{P(\pi_{<X} = \mathbf{Z}) P(\mathbf{U} = \mathbf{u}') P(\mathbf{U} = \mathbf{u} | \mathbf{V} = \mathbf{v})} \tag{42}$$

These weights are well-defined as long as our positivity assumptions hold. We may view the space of SCMs as characterized by the variable-specific GDE basis,

$$\mathcal{B}_X^{\mathrm{GDE}}(\mathbf{v}) = \{\mathrm{GDE}^{\mathbf{Z}}(X, Y | \mathbf{u}' \to \mathbf{u}) : (\mathbf{Z}, \mathbf{u}', \mathbf{u}) \in I\}, \tag{43}$$

for index set $I = \{(\mathbf{Z}, \mathbf{u}', \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{u}', \mathbf{u} \in \mathcal{D}_{\mathbf{U}}, \mathbf{V}(\mathbf{u}) = \mathbf{v}\}$. Given that the basis is isomorphic to $\mathbb{R}^I$, we view $\phi_i^{L_3}(w)$ as a measurable function on the $\mathcal{D}_{\mathcal{B}_X^{\mathrm{GDE}}}$, the space of all possible GDE settings. Using any atomless measures μ_i on each coordinate $i \in I$, such as the Lebesgue measure, we construct the product measure $\mu : \mathbb{R}^I \to \mathbb{R}$, which is also atomless by construction. We construct the pushforward measure $g := \mu \circ (\phi_i^{L_3})^{-1}$. Due to non-atomicity, $(\phi_i^{L_3})^{-1}(\{0\})$ has zero measure; in other words, $g(\{0\}) = 0$, and the space of SCMs where $\phi_i^{L_3}(w) = 0$ is measure zero.

Therefore, L_3 SVs satisfy weak causal explanatory power. □

A.3.3 Causal Normality

In this section, we discuss the normality property and prove that L_3 Shapley values satisfy normality. App. C.4 contains an extended discussion on the property of causal normality (Prop. 3).

Definition 8 (Comparative abnormality)**.** *X is more of an abnormal cause in $\mathcal{M}_2$ than $\mathcal{M}_1$ in the positive direction iff $\mathcal{D}_X := \mathcal{D}_{X_1} = \mathcal{D}_{X_2}$ and for all $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$, $\mathbf{E}(\mathbf{u}) = \mathbf{e} \wedge X_1(\mathbf{u}) = X_2(\mathbf{u})$ implies that for all $\mathbf{z}' \in \mathcal{D}_{\mathbf{X}\setminus\{X\}}$ and all $x_1' = x_2' \in \mathcal{D}_X$,*

$$Y_{\mathbf{z}', x_1'}(\mathbf{u}) = Y_{\mathbf{z}', x_2'}(\mathbf{u}) \tag{44}$$

$$\wedge (Y_{\mathbf{z}', x_1'}(\mathbf{u}) < Y_{\mathbf{z}'}(\mathbf{u}) \implies P(x_1') \geq P(x_2')) \tag{45}$$

$$\wedge (Y_{\mathbf{z}', x_1'}(\mathbf{u}) \geq Y_{\mathbf{z}'}(\mathbf{u}) \implies P(x_1') \leq P(x_2')) \tag{46}$$

and there exists $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ for which there is a positive change and probability mass decreases from $\mathcal{M}_1 \to \mathcal{M}_2$, or for which there is a negative change and probability mass increases from $\mathcal{M}_1 \to \mathcal{M}_2$:

$$(Y_{\mathbf{z}', x_1'}(\mathbf{u}) < Y_{\mathbf{z}'}(\mathbf{u}) \wedge P(x_1') > P(x_2')) \tag{47}$$

$$\vee (Y_{\mathbf{z}', x_1'}(\mathbf{u}) > Y_{\mathbf{z}'}(\mathbf{u}) \wedge P(x_1') < P(x_2')) \tag{48}$$

□

Property 3 (Causal Normality)**.** *EVA $\phi \in \Phi_n$ satisfies causal normality if, given why query $w \in \mathbb{W}$ and valid corresponding SCMs $\mathcal{M}_1, \mathcal{M}_2$ with identical observed variables $\mathbf{V}$, when some $X \in \mathbf{V}$ is a more of an abnormal cause in the positive direction in $\mathcal{M}_1$ than $\mathcal{M}_2$ but affects Y identically in both SCMs, $\phi_X(\mathcal{M}_1, w) > \phi_X(\mathcal{M}_2, w)$.* □

Lemma 5 (Causal Normality)**.** *L_3 SVs satisfy causal normality.* □

Proof. Let $X \in \mathbf{V}$ be more of an abnormal cause in the positive direction in $\mathcal{M}_2$ than $\mathcal{M}_1$. We first observe that

$$\mathrm{GDE}^{\mathbf{Z}}_{\mathcal{M}_2}(X, Y|\mathbf{u}' \to \mathbf{u}) = \mathrm{GDE}^{\mathbf{Z}}_{\mathcal{M}_1}(X, Y|\mathbf{u}' \to \mathbf{u}), \tag{49}$$

for all elements of the GDE basis on $\mathrm{Why}(y|\mathbf{x})$, due to Eq. (44). Without loss of generality, consider $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ such that

$$Y_{\mathbf{z}', x_1'}(\mathbf{u}) < Y_{\mathbf{z}'}(\mathbf{u}) \wedge P(x_1') > P(x_2') \tag{50}$$

Let $\mathbf{u}'$ induce $\mathbf{Z}(\mathbf{u}') = \mathbf{z}', X_1(\mathbf{u}') = x_1'$. We know that $Y(\mathbf{u})$ is identical between $\mathcal{M}_1(\mathbf{u}), \mathcal{M}_2(\mathbf{u})$. Therefore,

$$\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) > 0 \tag{51}$$

$$P_{\mathcal{M}_1}(\mathbf{u}') > P_{\mathcal{M}_2}(\mathbf{u}'). \tag{52}$$

Following Eqs. (45) and (46), it follows that there is an increase in both probability mass and value of $\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u})$. Given that the non-zero elements of the expectation

$$\phi_X^{L_3}(w) = \mathbb{E}_{\pi, \mathbf{u}', \mathbf{u}}[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{u}' \to \mathbf{u})] \tag{53}$$

along with their weights $P(\mathbf{u}')$ have increased or remained constant from $\mathcal{M}_1 \to \mathcal{M}_2$, $\phi_X^{L_3}(w)$ must increase from $\mathcal{M}_1 \to \mathcal{M}_2$. Therefore, L_3 SVs satisfy normality. □

A.3.4 Formal Statement

Theorem 4 (L_3 SVs satisfy desiderata)**.** *L_3 SVs satisfy causal admissibility, weak causal explanatory power, and normality.* □

Proof. The statement directly follows from Lems. 3, 4 and 5. □

A.4 L_3^* SVs satisfy causal explanation properties

In this section, we prove that L_3^* Shapley values satisfy the desired properties for causal explanations. An extended discussion of the properties is also present in App. C.5.

Definition 9 (Extended Counterfactual Shapley value (L_3^* SV))**.** *The extended counterfactual Shapley value for $X \in \mathbf{X}$ is defined as*

$$\phi_X^{L_3^*}(w) = \langle \mathbb{E}_{\pi, \mathbf{u}', \mathbf{u}} \left[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v}) \right], \mathbb{V}_{\pi, \mathbf{u}', \mathbf{u}} \left[\mathrm{GDE}^{\pi_{<X}}(X, Y|\mathbf{v}) \right] \rangle \tag{54}$$

where $\pi \sim \mathrm{Unif}(\Pi_{\mathbf{X}}), \mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U}|\mathbf{v})$; $\Pi_{\mathbf{X}}$ denotes the set of orderings on $\mathbf{X}$; and $\pi_{<X}$ denotes the variables prior to X in π. □

Lemma 6 (Strong Causal Explanatory Power)**.** *L_3^* SVs satisfy strong causal explanatory power.* □

Proof. This proof is much less involved than that of Lem. 4. When $\phi_X^{L_3^*}(w) = \langle 0, 0 \rangle$, we know that every element of the GDE basis must be zero: if the expectation and variance of a random variable are zero, it must be degenerate with mean zero. If an element of the GDE basis with positive probability mass is non-zero, then $\phi_X^{L_3^*}(w)$ cannot be zero by the contrapositive of the previous statement. If an admissible attribution $\phi'(w) \neq 0$, then there must be a non-zero GDE with positive mass, which implies that $\phi_X^{L_3^*}(w)$ is non-zero. Therefore, $\phi_X^{L_3^*}(w)$ satisfies strong causal explanatory power. □

Theorem 6 (L_3^* SVs satisfy properties)**.** *L_3^* SVs satisfy causal admissibility, strong causal explanatory power, and causal normality.* □

Proof. As shown in the proof of Lem. 3, L_3 SVs satisfy admissibility because the precondition implies that all relevant GDEs are zero. This implies that the mean and variance of the GDE also must be zero, as captured in $\phi_X^{L_3^*}(w) = \langle 0, 0 \rangle$. Therefore, causal admissibility is satisfied.

Strong causal explanatory power is satisfied, following Lem. 6.

Normality is satisfied if the $>$ operator on $\phi_X^{L_3^*}(w)$ is defined to operate only on the first element of the attribution, following Lem. 5. □

A.5 Causal Explanation Framework

In this section, we introduce the global explanatory basis and GDE basis; we prove that the global explanatory basis is equivalent to the SCM, while the GDE basis is equivalent to the NTE basis. An extended discussion of these results can be found in App. D.

Definition 10 (Global Explanatory Basis for SCMs)**.** *Given SCM $\mathcal{M}$, the global explanatory basis of $\mathcal{M}$ may be written as the tuple $\mathcal{C}(\mathcal{M}) := \langle \mathbf{V}_*^{\mathcal{M}}, P^{\mathcal{M}}(\mathbf{U}) \rangle$, where the global counterfactual basis $\mathbf{V}_*^{\mathcal{M}}$ is defined as*

$$\mathbf{V}_*^{\mathcal{M}}(\mathbf{u}) := \bigcup_{V \in \mathbf{V}^{\mathcal{M}}} V_*^{\mathcal{M}}(\mathbf{u}) \tag{55}$$

$$V_*^{\mathcal{M}}(\mathbf{u}) := \left\{ V_{\mathbf{z}}^{\mathcal{M}}(\mathbf{u}) : \mathbf{Z} \subseteq \mathbf{V}, \mathbf{z} \in \mathcal{D}_{\mathbf{Z}} \right\}. \tag{56}$$

□

Theorem 7 (Expressivity of the global basis)**.** *For any SCM $\mathcal{M}$ and intervention $\mathbf{X} \subseteq \mathbf{V}, \mathbf{x} \in \mathcal{D}_{\mathbf{X}}$, the global basis $\mathcal{C}(\mathcal{M})$ identifies submodel $\mathcal{M}_{\mathbf{x}}$.* □

Proof. Consider the SCM and intervention above. Given that $P(\mathbf{U}), \mathbf{U}, \mathbf{V}$ are trivially identified by $\mathcal{C}(\mathcal{M})$, we prove this theorem by constructing $\mathcal{F}_{\mathbf{x}}$. Consider $f_{V_{\mathbf{x}}} \in \mathcal{F}_{\mathbf{x}}$. Then we construct $f_{V_{\mathbf{x}}}(\mathbf{pa}_{V_{\mathbf{x}}}, \mathbf{u}_V) = V_{\mathbf{x}, \mathbf{pa}_{V_{\mathbf{x}}}}(\mathbf{u})$ for any $\mathbf{u} \supseteq \mathbf{u}_V$, where $\mathbf{pa}_{V_{\mathbf{x}}}$ denotes the parents of $V_{\mathbf{x}}$ in $\mathcal{M}_{\mathbf{x}}$. □

Definition 11 (Generalized Direct Effect (GDE) Basis)**.** *Consider SCM $\mathcal{M}$,* Why *query w, event counterfactual basis Y_*, explanatory variable subset $\mathbf{Z} \subseteq \mathbf{X}$, and explanatory variable of interest $X \in \mathbf{X} \setminus \mathbf{Z}$. The unit-level generalized direct effect of X on Y with adjustment set $\mathbf{Z}$, baseline $\mathbf{u}'$, and knowledge $\mathbf{u}$ is defined as*

$$\begin{aligned} \mathrm{GDE}^{\mathbf{Z}}(X, Y | \mathbf{u}' \to \mathbf{u}) &= \mathrm{NTE}(\mathbf{Z} \cup \{X\}, Y | \mathbf{u}' \to \mathbf{u}) - \mathrm{NTE}(\mathbf{Z}, Y | \mathbf{u}' \to \mathbf{u}) && (57) \\ &= Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}'), X(\mathbf{u}')}(\mathbf{u}). && (58) \end{aligned}$$

Assume that for any $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there is $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ inducing $\mathbf{V}(\mathbf{u}') = \mathbf{v}$. Then the GDE *basis of $\mathcal{M}$ for why query w is defined as*

$$\mathcal{B}_{\mathrm{GDE}}^{\mathcal{M}, w}(\mathbf{u}) = \{ \mathrm{GDE}^{\mathbf{Z}}(X, Y | \mathbf{u}' \to \mathbf{u}) : X \in \mathbf{X}, \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{u}' \in \mathcal{D}_{\mathbf{U}}) \} \tag{59}$$

□

Theorem 8 (NTE-GDE equivalence)**.** *The* GDE*s basis uniquely determines and is uniquely determined by the* NTE *basis.* □

Proof. Following Thm. 2, there is a mapping from the GDE basis to the value of every NTE in $\mathcal{B}^{\mathrm{NTE}}$. By definition of the GDE (Def. 3), there is a mapping from the NTE basis to every GDE in the $\mathcal{B}^{\mathrm{GDE}}$. This concludes our proof. □

A.6 Explanatory Impossibility Theorem and Soundness of Bounding

In this section, we prove the explanatory impossibility theorem. An extended discussion of this result is present in App. E.1.

Definition 12 (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 $\mathcal{M} \in \Omega'$,*

$$a \leq f(\mathcal{M}) \leq b. \tag{60}$$

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

Theorem 9 (Explanatory Impossibility Theorem)**.** *Consider positive observational distribution $P(\mathbf{V})$ with binary variables, Ω', the set of all SCMs $\mathcal{M}$ that induce the distribution, and why query $\mathrm{Why}(y|\mathbf{x})$, where $\mathbf{V} = \mathbf{X} \cup \{Y\}$. Ω' yields no information about any L_3 Shapley value $\phi_X^{L_3}$ for $X \in \mathbf{X}$.* □

Proof. Without loss of generality, let $\mathbf{X} = \mathbf{1}, Y = 1$. We show that there exists SCM $\mathcal{M}_1 \in \Omega'$ with maximal $\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = 1$ for all $\mathbf{Z} \subseteq \mathbf{X}$ and $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ where $X(\mathbf{u}') \neq X(\mathbf{u})$; the GDE is zero, otherwise. We next show that there exists SCM $\mathcal{M}_2 \in \Omega'$ with minimal $\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = -1$ for all $\varnothing \subset \mathbf{Z} \subseteq \mathbf{X}$ and $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ such that $Z(\mathbf{u}') \neq Z(\mathbf{u}), X(\mathbf{u}') \neq X(\mathbf{u})$; the GDE is again zero, otherwise. As a result, the induced L_3 SV $\phi_X^{L_3}$ takes its minimum and maximum possible values, respectively, and is not reduced from Ω; in other words, $P(\mathbf{V})$ yields no information about $\phi_X^{L_3}$.

We first construct $\mathcal{M}_1$:

$$\mathcal{F} = \begin{cases} \mathbf{X} & = \mathbf{U_X} \\ Y & = \begin{cases} 0 & U_Y = 0 \\ 1 - X & U_Y = 1 \\ X & U_Y = 2 \\ 1 & U_Y = 3 \end{cases} \end{cases} \tag{61}$$

$$P(\mathbf{U}) = \begin{cases} \mathbf{U_X} & \sim P(\mathbf{X}) \\ U_Y|U_X = 0 & \sim \mathrm{Categorical}(P(Y = 0|\mathbf{X} = \mathbf{u_x}), P(Y = 1|\mathbf{X} = \mathbf{u_x}), 0, 0]) \\ U_Y|U_X = 1 & \sim \mathrm{Categorical}(P(Y = 0|\mathbf{X} = \mathbf{u_x}), 0, P(Y = 1|\mathbf{X} = \mathbf{u_x}), 0]) \end{cases} \tag{62}$$

We confirm that

$$P^{\mathcal{M}_1}(\mathbf{v}) = P^{\mathcal{M}_1}(\mathbf{x})P^{\mathcal{M}_1}(y|\mathbf{x}) = P(\mathbf{x})P(y|\mathbf{x}) = P(\mathbf{v}). \tag{63}$$

Furthermore,

$$\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = \mathrm{GDE}^{\varnothing}(X, Y|\mathbf{u}' \to \mathbf{u}) = 1 \tag{64}$$

because intervening on $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$ has no effect on Y, by construction. Next, we observe that conditional on $X = 1, Y = 1$, $\mathrm{GDE}^{\varnothing}(X, Y|\mathbf{u}' \to \mathbf{u}) = 1$ for any $\mathbf{u}'$ inducing $X(\mathbf{u}') = 0$. Therefore, $\phi_X^{L_3}(w)$ takes its maximal value in $\mathcal{M}_1$, because all of its composing GDEs take their maximal values.

Next, we construct $\mathcal{M}_2$:

$$\mathcal{F} = \begin{cases} \mathbf{X} & = \mathbf{U_X} \\ Y & = \begin{cases} 1 & \mathbf{X} \setminus \{X\} = \mathbf{1} \wedge u_Y = 2 \\ 1 - X & \mathbf{X} \setminus \{X\} \neq \mathbf{1} \wedge u_Y = 2 \\ u_Y & u_Y \in \{0, 1\} \end{cases} \end{cases} \tag{65}$$

$$P(\mathbf{U}) = \begin{cases} \mathbf{U_X} & \sim P(\mathbf{X}) \\ U_Y|\mathbf{u_X} = \mathbf{1} & \sim \mathrm{Categorical}([0, P(Y = 1|\mathbf{X} = \mathbf{u_x}), P(Y = 0|\mathbf{X} = \mathbf{u_x})]) \\ U_Y|\mathbf{u_X} \neq \mathbf{1} & \sim \mathrm{Categorical}([P(Y = 0|\mathbf{X} = \mathbf{u_x}), P(Y = 1|\mathbf{X} = \mathbf{u_x}), 0]) \end{cases} \tag{66}$$

where $\mathbf{1}$ denotes the one vector. Here, it is evident that when $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ induces $\mathbf{Z}(\mathbf{u}') \neq \mathbf{Z}(\mathbf{u}), X(\mathbf{u}') \neq X(\mathbf{u})$,

$$\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}'), X(\mathbf{u}')}(\mathbf{u}) \tag{67}$$

$$= 0 - 1 = -1 \tag{68}$$

Given that $\mathcal{M}_1, \mathcal{M}_2$ are SCMs in which the values of $\mathrm{GDE}^{\mathbf{Z}}(X, Y|\mathbf{v})$ are maximized and minimized, respectively, the induced L_3 SVs also take their respective maximum and minimum values; they are not reduced from Ω. Therefore, Ω' yields no information about any L_3 Shapley value $\phi_X^{L_3}$ for $X \in \mathbf{X}$. □

Table 1: Linking examples to technical results.

Example	Reference	Result
Forest fire	Ex. 1	But-for causation and total effects (Def. 2)
Sprinkler	Ex. 2	Causal explanatory power (weak, Prop. 2)
Forest fire	Ex. 3	Normality Prop. 3
L_1, L_2 SVs	Ex. 4	Admissibility (Prop. 1)
Shark attacks	Ex. 5	Admissibility (Prop. 1)
Flu shot	Ex. 6	Admissibility (Prop. 1)
Exam scores	Ex. 7	Admissibility (Prop. 1)
Cancellation	Ex. 8	Causal explanatory power (strong, Prop. 2)
Binary four-variable basis	Ex. 9	Global explanatory basis Def. 10
Patient hospitalization	Ex. 10	Generalized why queries (Def. 14)
Forest fire	Ex. 11	Generalized why queries (Def. 14)
Forest fire	Ex. 12	Event explanatory basis (Def. 6)
Interview	Ex. 13	Necessity of the NTE (Lem. 2)
Variable interaction effects	Ex. 14	Interpretation of the GDE (Def. 11)
Binary Markovian chain	Ex. 15	Explanatory Impossibility Theorem (Thm. 9)

B Examples

B.1 Overview

Throughout this appendix, we introduce several examples aimed towards growing intuition for each theoretical result. We summarize these examples in Table 1.

B.2 Conjunction

Example 1 (Forest fire)**.** *A lightning strike hits a tree ($X_1 = 1$) in a rain forest, and the forest is arid, or dry ($X_2 = 1$). The strike sparks a forest fire ($Y = 1$). A possible SCM for this setting follows:*

$$P(\mathbf{U}) = \{U_1 \sim \text{Bern}(0.01), U_2 \sim \text{Bern}(0.5), U_Y \sim \text{Bern}(0.05)\} \tag{69}$$

$$\mathcal{F} = \begin{cases} X_1 & = U_1 \\ X_2 & = U_2 \\ Y & = (X_1 \wedge X_2) \vee U_Y \end{cases} \tag{70}$$

Human intuition indicates that the lightning $X_1 = 1$ and the forest's aridity $X_2 = 1$ caused the forest fire $Y = 1$. Depending on its prevalence, $P(U_2)$, dryness may be a better or worse cause than lightning: for instance, in a tropical rain forest, dryness is highly uncommon, and abnormal aridity would be a good explanation for a forest fire. □

From the SCM, we can infer this by observing that the total effect of the lightning on the fire is non-zero:

$$\text{TE}_{0,1}(y) = Y_{X_1=1}(\mathbf{u}) - Y_{X_1=0}(\mathbf{u}) = 1 - 0 = 1. \tag{71}$$

Specifically, the non-zero total effect of X_1 on Y implies that X_1 is a but-for cause of Y.

B.3 Disjunction

While the notion of but-for causation has a strong intuitive appeal and is undoubtedly useful in a number of cases, it actually may provide misleading answers in rather simple settings:

Example 2 (Rain & Sprinkler)**.** *It is raining outside ($X_1 = 1$) and the sprinkler is on ($X_2 = 1$), and the grass is currently wet ($Y = 1$). The mechanism of the Y variable is given by $Y \leftarrow X_1 \vee X_2$. For the described event, we can see that*

$$Y_{X_1=0} = Y_{X_2=0} = Y = 1. \tag{72}$$

In words, "had it not been raining," or "had the sprinkler not been on," the grass would have still be wet. Therefore, univariate changes to X_1 or X_2 cannot change the outcome Y, and but-for causation labels neither X_1 nor X_2 as the cause of Y. □

This example illustrates the importance of multivariate causes: univariate causes can be insufficient to ascertain causation.

B.4 Normality

Example 3 (Forest fire (cont.))**.** *A lightning strike ($X_1 = 1$), in conjunction with the presence of oxygen in the air ($X_2 = 1$), causes a fire ($Y = 1$). The SCM is defined as follows:*

$$\mathcal{F} := \begin{cases} X_1 & := U_L \\ X_2 & := U_O \\ Y & := X_1 \wedge X_2 \end{cases} \qquad P(\mathbf{U}) := \{U_L \sim \text{Bern}(\epsilon), U_O \sim \text{Bern}(1-\epsilon)\} \tag{73}$$

Technically, both the lightning and the oxygen are causes according to human intuition. However, lightning seems to be much better a cause than oxygen. □

B.5 Necessity and Sufficiency of the Natural Total Effect

See App. D.3 for a formal statement of the necessity and sufficiency of the NTE (Thm. 5).

C Explanatory Variable Attributions

C.1 General Explanatory Attributions

The aim of introducing the explanatory variable attribution (EVA, Def. 5) was to compare our method to others. However, our current notion of EVAs attribute vectors in $\mathbb{R}^n$, to explanatory variables in $\mathbf{X}$, to the general explanatory attribution, which attribute vectors in $\mathbb{R}^n$ to arbitrary explanatory objects. Some examples of such objects are subsets of variables [10] and causal pathways [39, 37].

Definition 13 (Explanatory Generalized Attribution)**.** *A generalized explanatory attribution is a mapping $\phi : \Omega \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}^n$, where Ω is the set of SCMs, and attribution dimensionality n is arbitrary.* □

Future work may focus on enumerating and justifying desirable properties for explanatory generalized attributions (EGA).

As a summary of causes, an explanation is expected to be *parsimonious* - to contain a low enough quantity of information that a human can parse it. In the context of EVAs, it may be formulated as a constraint on the attribution's dimensionality n.

Property 4 (Parsimony)**.** *An EVA $\phi : \mathcal{M} \times \mathbb{W} \times \mathbf{X} \to \mathbb{R}^n$ satisfies the parsimony property iff n is constant.* □

In words, explanation methods that attribute importance to the system's variables are considered parsimonious. Methods that attempt to attribute importance to subsets of the observables are not considered parsimonious, since the number of subsets grows exponentially with the number of variables.

C.2 Overview of prior methods

We begin by illustrating that explanatory attributions in the literature [27] fall under the umbrella of the EVA. Prominent in this literature are the set of feature attribution methods [33, 25, 36], which include SHAP values, a type of feature attribution method that summarizes conditional expectations of a model prediction using the Shapley value summarization technique [34]. This set of explanations is generally oblivious to causality in explanations and typically assumes that the event explanandum is the output of an ML model and therefore deterministic with respect to its inputs. In contrast, some work approaches this problem from the perspective of the literature on actual causation [10, 2, 3, 11],

Table 2: Literature analyzed through the desiderata.

Method(s)	Admissibility	Explanatory Power	Normality
Integrated Gradients [36]	✓	✗	✗
LIME [33]	✓	✗	✗
SHAP, asymmetric Shapley [25, 7]	✗	✓*	✓*
Causal Shapley values [16, 12, 17]	✗	✓*	✓*
Actual causation explanations [10, 2, 3, 11]	✓	✓	✗
PN, PS, PNS	✓	✗	✗
L_3 Shapley (this work)	✓	✓	✓

which studies how to logically define a cause within a formal causal model given full information and views an explanation as a single communicated cause; however, this branch of work lacks the ability to reason about probabilities of causation from data. A third branch attempts to approach the problem of explanations from the SCM framework, but it lacks the ability to handle explanations resulting from the effects of the interactions of multiple variables [29, 8]. Finally, some work attempts to incorporate interventional reasoning [7, 12, 16, 17] into a Shapley-style summary. This branch of work suffers from misalignment with definitions of causation; as a result, the notion of explanations as summaries of causes is not upheld. [39, 37] consider edge- and path-specific attributions respectively; both works assume the absence of unobserved confounders, and also do not consider connections to existing definitions of causation. Therefore, there is a gap in the literature at the intersection of key components of the event-specific explanation: probabilistic reasoning, actual causation, and explanatory variable attributions. In this section, we aim to make this gap explicit.

We summarize the groupings of methods in Table 2.

C.3 Admissibility

We motivate our discussion of admissibility with a simple example.

Example 4 (Connections to prior quantities)**.** *Consider a two-variable setting with variables X, Y. Then for the query $w = \text{Why}(y|x)$,*

$$\phi_X^{L_1}(w) = \mathbb{E}_{x' \sim P(X)}[\underbrace{\mathbb{E}[Y_x] - \mathbb{E}[Y_{x'}]}_{\text{TE}_{x',x}(y)} + \underbrace{\mathbb{E}[Y|x] - \mathbb{E}[Y_x]}_{\text{Exp-SE}_x(y)} - \underbrace{(\mathbb{E}[Y|x'] - \mathbb{E}[Y_{x'}])}_{\text{Exp-SE}_{x'}(y)}] \tag{74}$$

$$\phi_X^{L_2}(w) = \mathbb{E}_{x' \sim P(X)}[\underbrace{\mathbb{E}[Y_x] - \mathbb{E}[Y'_x]}_{\text{TE}_{x',x}(y|x')}] \tag{75}$$

$$\phi_X^{L_3}(w) = \mathbb{E}_{x' \sim P(X)}[\underbrace{\mathbb{E}[Y_x|x,y] - \mathbb{E}[Y_{x'}|x,y]}_{\text{TE}_{x',x}(y|x,y)}] \tag{76}$$

where $\text{TE}_{x',x}(y|\mathbf{e})$ is the total effect of changing $x' \to x$ on y, conditional on $\mathbf{e}$; and $\text{Exp-SE}_x(y)$ is the spurious effect of x on y. The differences and additions when transitioning from L_1, L_2 SVs to L_3 are highlighted in red and green, respectively. In words, L_1 SVs capture the total effect of X on Y for all units. In particular, when X is binary with $X = Y = 1$, we have:

$$\phi_X^{L_3}(w) = P(x') \underbrace{P(y'_{x'}|x,y)}_{PN(x,y)} \tag{77}$$

□

The differences and additions when transitioning from L_1, L_2 SVs to L_3 are highlighted in red and green, respectively. We provide examples that highlight the implications of such differences.

First, an example introduced by [16] indicates that SHAP can yield non-zero feature attributions given the absence of event causation.

Example 5 (Shark attacks)**.** *A team of business analysts is interested in improving ice cream sales. They collect data with many variables, including X_1 which corresponds to the monthly number of shark attacks, X_2 which corresponds to monthly ice cream sales, and Y, daily profit, a deterministic*

function of X_2. All variables are binary, with 1 representing a high value and 0 representing a low value. The SCM $\mathcal{M}^*$ of the underlying system, unknown to the analysts, is given by:

$$\mathcal{F} := \begin{cases} X_1 & := U_{12} \\ X_2 & := U_{12} \\ Y & := X_2 \end{cases} \tag{78}$$
$$P(\mathbf{U}) := \{P(U_{12} = 1) = \frac{1}{2}\}$$

and the causal diagram is given in Fig. 9a. One month, shark attacks, ice cream sales, and ice cream profit are each high, implying an observed event $\mathbf{e} = \{X_1 = 1, X_2 = 1, Y = 1\}$. The business analytics team is interested in explaining why profitability was high this month, corresponding to a why query $\text{Why}(y|\mathbf{e})$. □

Clearly, the fact that shark attacks are high is not a good explanation for why profitability is high. We can observe that in this example, L_1 SVs yield a non-zero attribution for shark attacks' contribution to ice cream sales,

$$\phi_{X_1}^{L_1} = \frac{1}{2}\left(\mathbb{E}[Y|x_1] - \mathbb{E}[Y]\right) + \frac{1}{2}\left(\mathbb{E}[Y|x_1, x_2] - \mathbb{E}[Y|x_2]\right) \qquad = \frac{1}{4}, \tag{79}$$

while L_2 SVs yield a zero attribution to this contribution,

$$\phi_{X_1}^{L_2} = \frac{1}{2}\left(\mathbb{E}[Y_{x_1}] - \mathbb{E}[Y]\right) + \frac{1}{2}\left(\mathbb{E}[Y_{x_1,x_2}] - \mathbb{E}[Y_{x_2}]\right) \qquad = 0. \tag{80}$$

However, L_2 SVs do not fix the underlying issue that actual causes in settings consistent with $\mathbf{e}$ are not captured. This is particularly the case when there is causation between variables $\mathbf{X}$. To illustrate this, we introduce a simple example without unobserved confounding that shows that L_2 Shapley values can also fail in this setting, in addition to linear regression, LIME, SHAP, and other L_1 methods.

However, L_2 SVs do not fix the underlying issue that actual causes in settings consistent with $\mathbf{e}$ are not captured. This is particularly the case when there is causation between variables $\mathbf{X}$. To illustrate this, we introduce a simple example without unobserved confounding that shows that L_2 Shapley values can also fail in this setting, in addition to linear regression, LIME, SHAP, and other L_1 methods.

Example 6 (Flu shot)**.** *Alice gets the flu shot ($X = 1$), but still gets the flu later in the year ($Y = 0$). The flu shot is known to never cause the flu, and it sometimes prevents the flu. Alice is curious why she got the flu.*

A possible SCM describing the relationship between the flu shot and the flu follows:

$$P(X, Y_0, Y_1) = \begin{cases} P(X = 1) & = 0.1 \\ P(Y_0, Y_1) & = \begin{cases} P(Y_0 = 0, Y_1 = 0) & = 0.09 \\ P(Y_0 = 0, Y_1 = 1) & = 0 \\ P(Y_0 = 1, Y_1 = 0) & = 0.9 \\ P(Y_0 = 1, Y_1 = 1) & = 0.01 \end{cases} \end{cases} \tag{81}$$

□

In the example above, we don't know why Alice has the flu, but we do know that it is not due to the flu shot. This can be observed by using the probability of necessity [29]:

$$PN(x, y) = P(y'_{x'}|x, y) = \frac{P(y'_{x'}, y_x, x)}{P(x, y)} = 0. \tag{82}$$

On the other hand, we can show that L_1 and L_2 SVs, which are identical due to the absence of unobserved confounders, both assign the flu shot a negative attribution when explaining the output of

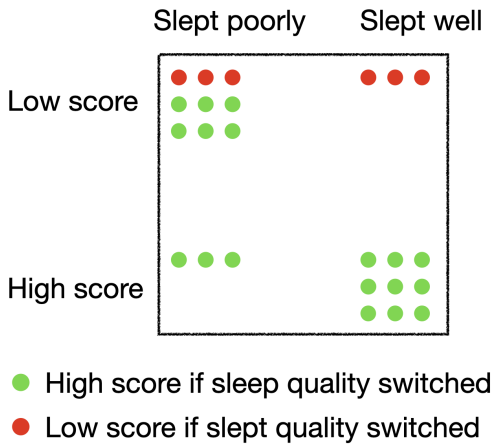


Figure 6: A sample of 24 students from Alice's class. For students like Alice who have high scores and slept well (in the bottom right), reducing sleep quality would not affect their score, indicating high quality sleep is not a cause of their high performance. On the other hand, improving sleep quality would improve performance for some students who had low scores and slept poorly; for these students, poor sleep quality is a cause of their poor performance.

the predictor:[7]

$$\phi_X^{L_1} = \phi_X^{L_2} = \mathbb{E}[Y_{X=1}] - \mathbb{E}[Y] \tag{83}$$

$$= P(X=0)\left(\mathbb{E}[Y_{X=1}] - \mathbb{E}[Y_{X=0}]\right) \qquad = 0.9\,(0.01 - 0.91) = -0.81. \tag{84}$$

We introduce a second counterexample to both L_1 and L_2 SVs, this time using a confounder.

Example 7 (Exam scores)**.** *A team of data scientists interested in academic performance is attempting to understand how sleeping habits might affect test scores. They collect data where X_1 corresponds to whether a student sleeps well the night before their exam, X_2 whether they scores high on the exam, and Y is a deterministic function of whether or not the student does well in class – a mirror of whether or not they do well on the final exam. Whether the student has good study habits is unobserved and described by the exogenous variable U. The SCM $\mathcal{M}^*$ of the underlying system is given by:*

$$\begin{aligned} \mathcal{F} &:= \begin{cases} X_1 & := U_{12} \\ X_2 & := \mathbf{1}[U_2 = 4] \vee (\mathbf{1}[U_2 \neq 1] \wedge (X_1 \vee U_{12})) \\ Y & := X_2 \end{cases} \\ P(\mathbf{U}) &:= \{P(U_{12} = 1) = \frac{1}{2}, U_2 \sim \text{Unif}(\{1,2,3,4\})\} \\ \mathbf{V} &:= \{X_1 = 1, X_2 = 1, Y = 1\} \end{aligned} \tag{85}$$

and the causal diagram is given in Fig. 9c. A sample set of students is given in Fig. 6. A student called Alice sleeps well, scores high, and does well in her class, which implies an observed event $\mathbf{e} = \{X_1 = 1, X_2 = 1, Y = 1\}$. The data science team is interested in explaining why Alice did well in this class - formally, the why query $w = \text{Why}(Y = 1 | X_1 = 1, X_2 = 1)$. □

What should be expected in terms of attributions to sleep quality and exam score? Visually, in the bottom right corner in Fig. 6, there are only green units, meaning that modifying X_1 would have no effect on any of the students like Alice: it should be assigned an attribution of zero. We would then expect the entirety of the attribution to fall to exam score and none to sleep quality. We can also arrive at the same conclusion by examining the mechanisms $\mathcal{F}$ determining a student's performance f_Y (Eq. (85)): since $X_1 = 1$, we can infer that $U_{12} = 1$, implying that X_1 has no effect on X_2 or Y, even when intervening on other observed variables. This intuition translates to the fact that for Alice,

[7] Given that L_1 and L_2 SVs are used to explain deterministic predictors f, we may introduce a function $f_{\hat{Y}}(X, Y) = Y$ as our explanation target. This does not change the valuation of the PN, and halves L_1, L_2 attributions but does not change the fact that they are both nonzero.

sleeping well the night before her exam contributed nothing to her performance in the class. Rather, it was the fact that she had good study habits, which led to her high exam score, and subsequently resulted in strong class performance.

We may compute L_1 and L_2 Shapley values to see if this intuition holds for these EVAs:

$$\begin{aligned}\phi_1^{L_1}(w) &= \frac{1}{2}((\mathbb{E}[Y|x_1,x_2]-\mathbb{E}[Y|x_2]) \\ &\quad + (\mathbb{E}[Y|x_1]-\mathbb{E}[Y])) \\ &= \frac{1}{2}((1-1)+(0.75-0.5)) \\ &= 0.125.\end{aligned} \tag{86}$$

$$\begin{aligned}\phi_1^{L_2}(w) &= \frac{1}{2}((\mathbb{E}[Y|do(x_1,x_2)]-\mathbb{E}[Y|do(x_2)]) \\ &\quad + (\mathbb{E}[Y|do(x_1)]-\mathbb{E}[Y])) \\ &= \frac{1}{2}((1-1)+(0.75-0.5)) \\ &= 0.125.\end{aligned} \tag{87}$$

We can see that LIME, SHAP values, and interventional Shapley values all attribute a non-zero effect to the variable X_1 in this case, contrary to the desired behavior. Note that because there is a non-zero association between X_1 and Y, the attribution for LIME and for linear regression also yield a non-zero attribution for X_1, although for LIME, the exact value depends on the strength of the ridge regression parameter.

As alluded to earlier, Lewis [22] argues that explaining an event is to provide information about its causal history. Both kinds of methods discussed so far – linear regression and methods in the Shapley-based attribution line of work – fail on this account. Crucially, the failure of these methods is not a coincidence but a fundamental corollary of the Causal Hierarchy Theorem (CHT, [1, Thm. 1]), which states that observational or interventional measures alone, in the absence of appropriate causal assumptions, are insufficient for reasoning about counterfactual relationships between variables. The failure of existing methods with respect to causality motivates the first formal property that allows the assessment on whether an explanation satisfies this causal requirement.

Property 1 (Causal Admissibility)**.** *Consider why query $w = \text{Why}(y|\mathbf{X};\mathbf{e}) \in \mathbb{W}$. Let there be an absence of causation from $X_i \in \mathbf{X}$ to Y in consistent worlds if for all units $\mathbf{u} \in \mathcal{D}_\mathbf{U}$, there is no setting $\mathbf{z}' \in \mathcal{D}_{\mathbf{X}\setminus\{X_i\}}$ and $x_i' \in \mathcal{D}_{X_i}$ such that*

$$\mathbf{E}(\mathbf{u}) = \mathbf{e} \wedge Y_{\mathbf{z}',x_i'}(\mathbf{u}) \neq Y_{\mathbf{z}',x_i}(\mathbf{u}). \tag{10}$$

In words, there is an absence of causation from X_i to Y if it is impossible to change Y by changing X_i, under any unit consistent with the observations. Then, EVA $\phi_n \in \Phi$ is causally admissible if for all why queries w and variables $X_i \in \mathbf{X}$, the absence of causation from X_i to Y in consistent worlds implies a zero attribution $\phi_i(w) = \mathbf{0}$. □

In English, causal admissibility states that, considering settings consistent with our observations $\mathbf{E} = \mathbf{e}$, if there is no way to change X_i so that the value of Y changes under any circumstances $\mathbf{Z} = \mathbf{z}'$, then X_i should be assigned a zero attribution.

Returning to Ex. 7, we see that there is an absence of causation from X_1 to Y: there is no way to change X_1 (holding fixed any set of variables) such that Y changes, given the current setting of $\mathbf{u}$. However, we clearly see non-zero attributions for L_1 and L_2 SVs; therefore, L_1, L_2 SVs violate admissibility.

C.4 Normality

There are two considerations when constructing a property for normality: first, what describes an abnormal cause; and second, what implications should such abnormality have on an EVA?

We take the view that the magnitude of a causal effect should be amplified if it is more abnormal, regardless of its sign: a more abnormal negative effect should lead to a more negative attribution, and

a more abnormal positive effect should lead to a more positive attribution. In addition, such changes to abnormality should only be considered, keeping all else the same. There is of course the potential concern of effect cancellation: we address this in App. C.5. In this section, to this end, we introduce the following definition of direction causal abnormality.

Definition 8 (Comparative abnormality)**.** *X is more of an abnormal cause in $\mathcal{M}_2$ than $\mathcal{M}_1$ in the positive direction iff $\mathcal{D}_X := \mathcal{D}_{X_1} = \mathcal{D}_{X_2}$ and for all $\mathbf{u} \in \mathcal{D}_\mathbf{U}$, $\mathbf{E}(\mathbf{u}) = \mathbf{e} \wedge X_1(\mathbf{u}) = X_2(\mathbf{u})$ implies that for all $\mathbf{z}' \in \mathcal{D}_{\mathbf{X}\setminus\{X\}}$ and all $x_1' = x_2' \in \mathcal{D}_X$,*

$$Y_{\mathbf{z}',x_1'}(\mathbf{u}) = Y_{\mathbf{z}',x_2'}(\mathbf{u}) \tag{44}$$

$$\wedge(Y_{\mathbf{z}',x_1'}(\mathbf{u}) < Y_{\mathbf{z}'}(\mathbf{u}) \implies P(x_1') \geq P(x_2')) \tag{45}$$

$$\wedge(Y_{\mathbf{z}',x_1'}(\mathbf{u}) \geq Y_{\mathbf{z}'}(\mathbf{u}) \implies P(x_1') \leq P(x_2')) \tag{46}$$

and there exists $\mathbf{u} \in \mathcal{D}_\mathbf{U}$ for which there is a positive change and probability mass decreases from $\mathcal{M}_1 \to \mathcal{M}_2$, or for which there is a negative change and probability mass increases from $\mathcal{M}_1 \to \mathcal{M}_2$:

$$(Y_{\mathbf{z}',x_1'}(\mathbf{u}) < Y_{\mathbf{z}'}(\mathbf{u}) \wedge P(x_1') > P(x_2')) \tag{47}$$

$$\vee(Y_{\mathbf{z}',x_1'}(\mathbf{u}) > Y_{\mathbf{z}'}(\mathbf{u}) \wedge P(x_1') < P(x_2')) \tag{48}$$

□

Using this definition, we can now fully describe the causal normality property.

Property 3 (Causal Normality)**.** *EVA $\phi \in \Phi_n$ satisfies causal normality if, given why query $w \in \mathbb{W}$ and valid corresponding SCMs $\mathcal{M}_1, \mathcal{M}_2$ with identical observed variables $\mathbf{V}$, when some $X \in \mathbf{V}$ is a more of an abnormal cause in the positive direction in $\mathcal{M}_1$ than $\mathcal{M}_2$ but affects Y identically in both SCMs, $\phi_X(\mathcal{M}_1, w) > \phi_X(\mathcal{M}_2, w)$.* □

C.5 Causal Explanatory Power

In general, it is not possible to guarantee strong explanatory power when constructing an attribution that is a weighted average of effects with potentially mixed signs.

Example 8 (Cancellation)**.** *Consider a two-variable linear SCM $\mathcal{M}$ with variables X_1, X_2, Y and structural function*

$$\mathcal{F} = \begin{cases} X_1 &= u_1, \\ X_2 &= \gamma x_1 + u_2, \\ Y &= \alpha x_1 + \beta x_2 + u_3, \end{cases} \tag{88}$$

where U_1, U_2, U_3 are independent exogenous noise variables with zero mean. Observe that L_3 SVs for X_1 are computed as

$$\phi_{X_1}^{L_3} = \frac{1}{2}\left(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]\right) + \frac{1}{2}\left(\mathbb{E}[Y_{P(X_2)}|\mathbf{v}] - \mathbb{E}[Y_{P(X_1,X_2)} \mid \mathbf{v}]\right) \tag{89}$$

$$= \frac{1}{2}(\alpha + \beta\gamma)x_1 + \frac{1}{2}\alpha x_1 \tag{90}$$

$$= \frac{2\alpha + \beta\gamma}{2}x_1. \tag{91}$$

Specifically, when $\alpha = -\frac{\beta\gamma}{2}$, L_3 SVs will always be zero, despite both the total and direct effects of X_1 being non-zero. This is an example of a cancellation, an instance in the measure-zero space of SCMs where strong explanatory power is violated: there is clearly an effect, but the attribution is zero. □

Indeed, this issue is raised as a fundamental issue of Shapley values [14]. To address this issue, we introduce extended counterfactual Shapley values, which we refer to as L_3^* SVs. We solve the issue by including both expectation and variance of the GDE with respect to its dimension of variation, rather than just the expectation.

Definition 9 (Extended Counterfactual Shapley value (L_3^* SV))**.** *The extended counterfactual Shapley value for $X \in \mathbf{X}$ is defined as*

$$\phi_X^{L_3^*}(w) = \langle \mathbb{E}_{\pi,\mathbf{u}',\mathbf{u}}\left[\mathrm{GDE}^{\pi<X}(X, Y|\mathbf{v})\right], \mathbb{V}_{\pi,\mathbf{u}',\mathbf{u}}\left[\mathrm{GDE}^{\pi<X}(X, Y|\mathbf{v})\right]\rangle \tag{54}$$

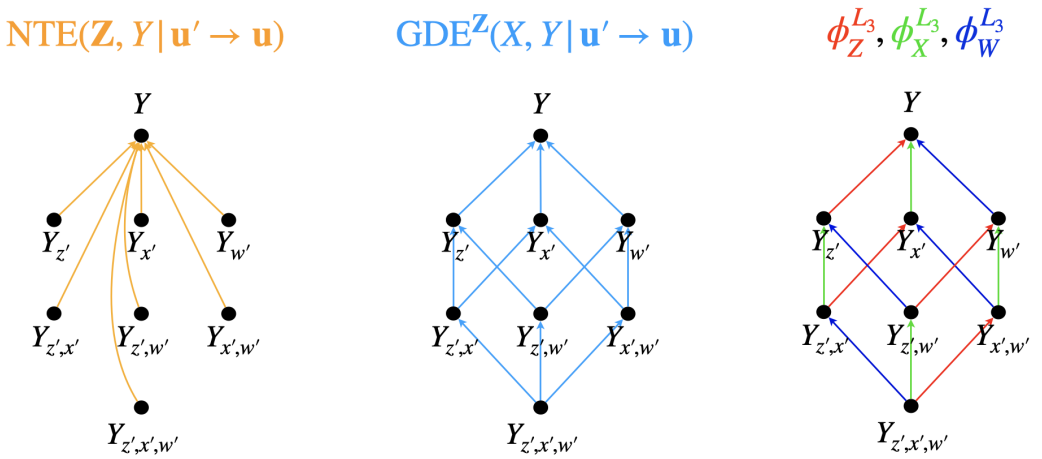


Figure 7: Key counterfactual quantities, illustrated under the query $\text{Why}(Y|\{Z, X, W\}; \cdot)$. Black nodes represent the event counterfactual basis (Def. 6). Left: The orange arrows denote the NTE basis (Def. 7), where each arrow represents an NTE (Def. 2), a difference between two counterfactuals. Center: The light blue arrows represent the GDE basis (Def. 11), where each arrow is a GDE (Def. 3), expressing variable-specific contributions to potential cause $\mathbf{Z} \cup \{X\}$. Right: L_3 Shapley values (L_3 SVs, Def. 4) average GDEs, weighting each GDE with the prior probability of its baseline.

where $\pi \sim \text{Unif}(\Pi_{\mathbf{X}}), \mathbf{u}' \sim P(\mathbf{U}), \mathbf{u} \sim P(\mathbf{U}|\mathbf{v})$*;* $\Pi_{\mathbf{X}}$ *denotes the set of orderings on* $\mathbf{X}$*; and* $\pi_{<X}$ *denotes the variables prior to* X *in* π*.* □

Intuitively, the extended values will not cause violations of admissibility or normality still satisfy admissibility. However, the extension allows the values to satisfy strong, rather than weak, causal explanatory power.

Lemma 6 (Strong Causal Explanatory Power)**.** *L_3^* SVs satisfy strong causal explanatory power.* □

This allows us to prove a strengthened set of causal explanation properties for L_3^* SVs.

Theorem 6 (L_3^* SVs satisfy properties)**.** *L_3^* SVs satisfy causal admissibility, strong causal explanatory power, and causal normality.* □

D A Causal Framework for Explanations

In this section, we describe complete reasoning motivating the NTE, GDE, and L_3 SVs as quantities that summarize an event's causal history. We first introduce the global explanatory basis, an equivalent representation of the SCM that replaces functions with unit-level counterfactuals. Because we are interested in explaining specific events, or outcomes, we introduce the Why query, a technically precise formulation of the natural language why question, and demonstrate how the global explanatory basis can be reduced to the event explanatory basis, a set of counterfactual and probabilistic information we argue is sufficient to answer Why queries. We transform this information into the set of natural total effects (NTE) to describe how any set of variables affects the outcome, and we show the NTE basis - the set of NTEs - is informationally equivalent to the event counterfactual basis. To describe a single variable's contribution to any NTE, we introduce the generalized direct effect GDE. Finally, we introduce L_3 Shapley values and extend them to satisfy all properties (Props. 1 to 3), including strong explanatory power.

D.1 Global explanatory basis

We begin our discussion with an example introducing our basic data structure for reasoning. It illustrates a new way to represent SCMs using a mapping from SCM to unit-level counterfactuals that it induces.

Example 9 (Binary four-variable SCM basis)**.** *Consider a four-variable SCM $\mathcal{M}$ with topologically ordered binary variables $\mathbf{V} = \{Z, X, W, Y\}$ with mechanisms*

$$\mathcal{F} = \{f_Z(U_Z), f_X(Z, U_X), f_W(Z, X, U_W), f_Y(Z, X, U_Y) \tag{92}$$

and exogenous variable distribution $P(\mathbf{U}) = P(U_Z, U_X, U_W, U_Y)$. Under any full specification of $\mathbf{u} = \{u_Z, u_X, u_W, u_Y\}$, we can evaluate settings of the SCM's counterfactuals. For instance, if $\mathbf{u} = \{U_Z = 1, U_X = 1, U_W = 1, U_Y = 1\}$, the SCM induces unit counterfactuals

$$Z(\mathbf{u}), X(\mathbf{u}), W(\mathbf{u}), Y(\mathbf{u}). \tag{93}$$

It also induces unit counterfactuals in all interventional submodels $\mathcal{M}_{\mathbf{z}}$ for $\mathbf{z} \subseteq \{z', x', w', y'\}$ for any binary settings z', x', w', y'. For example, for the variable Y, the SCM would induce the counterfactuals

$$Y_* := \{Y, Y_{z'}, Y_{x'}, Y_{w'}, Y_{z'x'}, Y_{z'w'}, Y_{w'x'}, Y_{z'w'x'}, Y_{y'}, Y_{y',z'}, \dots\}. \tag{94}$$

More generally, we view the full set of counterfactuals as the union of all individual sets,

$$\mathbf{V}_* := Z_* \cup X_* \cup W_* \cup Y_* \tag{95}$$

We will call this set of counterfactuals the global counterfactual basis. We can also reverse this mapping, identifying each function's behavior using the settings of its counterfactuals. For example to identify functional behavior in the interventional submodel $\mathcal{M}_{X=1}$, we can use our topological ordering to identify our functions as:

$$f_{Z_{X=1}}(\mathbf{u}) = Z(\mathbf{u}) \tag{96}$$
$$f_{X_{X=1}}(\mathbf{u}) = 1 \tag{97}$$
$$f_{W_{X=1}}(z, \mathbf{u}) = W_{Z=z, X=1}(\mathbf{u}) \tag{98}$$
$$f_{Y_{X=1}}(z, w, \mathbf{u}) = Y_{Z=z, X=1, W=w}(\mathbf{u}) \tag{99}$$

□

The intuition of representing an SCM with its counterfactuals generalizes to arbitrary numbers of variables with arbitrary domains.

Definition 10 (Global Explanatory Basis for SCMs)**.** *Given SCM $\mathcal{M}$, the global explanatory basis of $\mathcal{M}$ may be written as the tuple $\mathcal{C}(\mathcal{M}) := \langle \mathbf{V}_*^{\mathcal{M}}, P^{\mathcal{M}}(\mathbf{U}) \rangle$, where the global counterfactual basis $\mathbf{V}_*^{\mathcal{M}}$ is defined as*

$$\mathbf{V}_*^{\mathcal{M}}(\mathbf{u}) := \bigcup_{V \in \mathbf{V}^{\mathcal{M}}} V_*^{\mathcal{M}}(\mathbf{u}) \tag{55}$$
$$V_*^{\mathcal{M}}(\mathbf{u}) := \left\{ V_{\mathbf{z}}^{\mathcal{M}}(\mathbf{u}) : \mathbf{Z} \subseteq \mathbf{V}, \mathbf{z} \in \mathcal{D}_{\mathbf{Z}} \right\}. \tag{56}$$

□

We exclude $\mathcal{M}$ when clear from context. Intuitively, the information in the basis is equivalent to all information that can be derived from the SCM and interventions upon the SCM. We formalize this intuition below.

Theorem 7 (Expressivity of the global basis)**.** *For any SCM $\mathcal{M}$ and intervention $\mathbf{X} \subseteq \mathbf{V}, \mathbf{x} \in \mathcal{D}_{\mathbf{X}}$, the global basis $\mathcal{C}(\mathcal{M})$ identifies submodel $\mathcal{M}_{\mathbf{x}}$.* □

In the remainder of this section, we use the global basis as our starting point to ground our quantities measuring causal effects.

D.2 The Why query

To explain specific events, we must first define the *why query*. A precise formalization is necessary: in the absence of contextual information, "why" questions in natural language suffer from ambiguity in terms of the information requested [18]. For instance, even "Why?" alone can be a valid question and requests different information depending on context. For precision, we will refer to a natural language why question as a "why question" and to the precise, formal object representing a why question as a "Why query."

We discuss what information we need to formally represent why questions, such as "Why did the event $Y = y$ occur?" Our first step is to include the **observed events** $\mathbf{E} = \mathbf{e}$, as well as the **event explanandum** $Y = y \in \mathbf{e}$, in the Why query. Explicitly specifying the explanandum $Y = y$ prevents the ambiguity inherent in questions such as "Why?" regarding what is being explained. Doing so also specifies possible **event foils** $Y = y' \neq y$, which may also be ambiguous in why question.[8]

We assume that the underlying model that generates $\mathbf{E} = \mathbf{e}$ is an unobserved SCM $\mathcal{M}$ [30] with endogenous variables $\mathbf{V} \supseteq \mathbf{E}$, and we precisely specify **explanatory variables** $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$, which we use to explain our event explanandum.[9]

Finally, an explanation is a *transfer of information* from one agent with evidence about the world $(\mathcal{M}, \mathbf{E} = \mathbf{e})$ to an agent with potentially different evidence about the world $\mathbf{E}' = \mathbf{e}'$ [10].

Example 10 (Patient hospitalization)**.** *A patient has asthma ($X_1 = 1$), performs a challenging cardio workout ($X_2 = 1$), and is hospitalized ($Y = 1$) as a result. The patient's doctor knows that the patient has asthma and that they were hospitalized $\mathbf{e}_1 = \{x_1, y\}$, while the patient's gym partner knows only they performed the workout and were hospitalized $\mathbf{e}_2 = \{x_2, y\}$. Both are interested in the question: "Why was the patient hospitalized ($Y = 1$)?".* □

In the example above, the best explanation to the doctor would be the fact that the patient performed a challenging workout, given that this information was not known. On the other hand, the best explanation to the patient's gym partner would be the fact that the patient had asthma.

Thus, the knowledge of the explainee is an essential factor in constructing an explanation and determines the distribution over **explanatory variable baselines** - alternate settings for $\mathbf{U}$, which determine alternate, interventional values of $\mathbf{X}$.

Definition 14 (Why Query (general))**.** *Given SCM $\mathcal{M}$, a why query is a tuple $(Y = y, \mathbf{X}, \mathbf{E} = \mathbf{e}, \mathbf{E}' = \mathbf{e}') \in \mathcal{W}$, the space of why queries. The tuple consists of the underlying SCM $\mathcal{M}$ with observed variables $\mathbf{V}$, observed evidence $\mathbf{E} = \mathbf{e}$, where $\mathbf{E} \subseteq \mathbf{V}$, event explanandum $Y = y$ implied by $\mathbf{e}$, explanatory variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$, and the prior evidence of the explainee $\mathbf{E}' = \mathbf{e}'$ for $\mathbf{E}' \subseteq \mathbf{V}$.*

We denote this tuple $\mathrm{Why}(y|\mathbf{e}' \to \mathbf{e}; \mathbf{X})$. □

In this work, we primarily consider cases where the set of evidence $\mathbf{E}$ contains settings for all endogenous variables $\mathbf{V}$, and the explainee has no prior knowledge; that, is $\mathbf{V} = \mathbf{E} = \mathbf{X} \cup \{Y\}$ and $\mathbf{E}' = \varnothing$. If only one of the two is specified, assume $\mathbf{E}' = \varnothing$. For concision, when the SCM $\mathcal{M}$ is unambiguous, we denote the why query as $\mathrm{Why}(y|\mathbf{x})$.

We provide further examples of natural language why questions to Why queries below.

Example 11 (Bivariate forest fire)**.** *A lightning strike hits a tree ($X_1 = 1$) in a rainforest. The tree is dry ($X_2 = 1$). The strike sparks a forest fire ($Y = 1$). A possible SCM for this setting follows:*

$$P(\mathbf{U}) = \{U_1 \sim \mathrm{Bern}(0.01), U_1 \sim \mathrm{Bern}(0.5), U_Y \sim \mathrm{Bern}(0.05)\} \tag{100}$$

$$\mathcal{F} = \begin{cases} X_1 & = U_1 \\ X_2 & = U_2 \\ Y & = (X_1 \wedge X_2) \vee U_Y \end{cases} \tag{101}$$

A common Why *query in this setting may be* $\mathrm{Why}(y|x_1, x_2)$*, which translates to the English request "Given that lightning struck and the tree was dry, why did the tree catch fire? Explain your answer in terms of the lightning strike and the tree's dryness." In this case, the English answer "Because lightning struck and the tree was dry" may be a valid answer to this query.*

There are many dimensions of variation of the Why *query.*

First, we may be interested in restricting or changing the explanatory variables. For instance, $\mathrm{Why}(y|\{X_1\}; \mathbf{e}' \to \mathbf{e})$ *would be a valid query, requesting explanation only in terms of the lighting*

[8] A classic example, reproduced in [26], is the question "Why is the door open?" This question does not clearly specify a foil and could take many contextually dependent meanings, including "Why is the door open *rather than closed*?" or "Why is the door open *rather than the window*?"

[9] The distinction between an explanatory variable $X \in \mathbf{X} \subseteq \mathbf{V}$ and an observed variable $E \in \mathbf{E} \subseteq \mathbf{V}$ is that while both variables can be observed and intervened upon, only variables in $\mathbf{E}$ are actually observed, while only variables in $\mathbf{X}$ are intervened upon to construct explanations.

strike. In this case, "Because of the lightning strike." might be a valid answer. Similarly, "Because of the tree's dryness" may be a valid answer to $\text{Why}(y|\{X_2\};\mathbf{e})$*. A good deal of the challenge of answering an English why question lies in determining an appropriate set of explanatory variables [18]; for a technically precise* Why *query, we require this set to be specified.*

We may also vary the event explanandum. For instance, we could ask $\text{Why}(X_1|\{X_2, Y\};\mathbf{e})$ *- "Why did the lightning strike?" In this case, we would not have enough information to answer the question. A good explanation would indicate that neither the tree's dryness nor the forest fire caused the lightning strike, and that we do not know why the lightning struck.*

A third dimension of variation is the set of events $\mathbf{e}$ *observed by the explainer. We could similarly ask* $\text{Why}(y|\{X_1, X_2\})$*: "Without prior knowledge, why did the tree catch fire?" In this case, "With a small chance, because lightning struck. Other reasons not in the explanatory variable set (e.g., a campfire ran over, vegetation spontaneously combusted) are also likely." would be a reasonable answer.*

The final dimension of variation is the set of events $\mathbf{e}'$ *known by the explainee. Say the explainee already knew that the lightning strike had occurred:* $\text{Why}(y|\{X_1, X_2\};\mathbf{e}' = \{x_1\} \rightarrow \mathbf{e})$*. A good answer to the same why question, "Why did the tree catch fire?" would likely prioritize dryness in the explanation. Conversely, if the explainee knew only that the tree was dry (*$\text{Why}(y|\{X_1, X_2\};\mathbf{e}' = \{x_2\} \rightarrow \mathbf{e})$*), a good explanation would likely prioritize the novel information of the lightning strike.* □

The example above touches upon each of the dimensions of variation of the Why query, given an SCM. In this work, we primarily focus on cases where the set of evidence $\mathbf{E}$ contains settings for all endogenous variables $\mathbf{V}$, and the explainee has no prior knowledge; that, is $\mathbf{V} = \mathbf{E} = \mathbf{X} \cup \{Y\}$ and $\mathbf{E}' = \varnothing$. Thus, for concision, when the SCM $\mathcal{M}$ is unambiguous, we denote the why query as $\text{Why}(y|\mathbf{x})$.

Given a Why query, we aim to select only information in the global explanatory basis $\mathcal{C}(\mathcal{M})$ necessary to answer the query.

Example 12 (Bivariate forest fire (cont.))**.** *Consider the forest fire example. Let observations* $\mathbf{e} = \{X_1 = 1, X_2 = 1, Y = 1\}$*, and consider the why query* $w = \text{Why}(y|x_1, x_2)$*. We construct a minimal set of information from the global explanatory basis* $\mathcal{C}(\mathcal{M})$ *sufficient to infer the causes of* y*. The effects of the explanatory variables* x_1, x_2 *on* Y *may be written as what we will call the event counterfactual basis* $Y_{\mathbf{X}^*}(\mathbf{u})$*, the set of variables*

$$\{Y(\mathbf{u}), Y_{x_1'}(\mathbf{u}), Y_{x_1}(\mathbf{u}), Y_{x_2'}(\mathbf{u}), Y_{x_2}(\mathbf{u}), Y_{x_1', x_2'}(\mathbf{u}), Y_{x_1', x_2}(\mathbf{u}), Y_{x_1, x_2'}(\mathbf{u}), Y_{x_1, x_2}(\mathbf{u})\}. \tag{102}$$

There are two settings of $\mathbf{U}$ *consistent with observations* $\mathbf{e}$*:* $\mathbf{u}_1 = \{u_1, u_2, u_Y\}$*, and* $\mathbf{u}_2 = \{u_1, u_2, u_Y'\}$*. Under each of these settings, we have*

$$Y_{\mathbf{X}^*}(\mathbf{u}_1) = \{y, y_{x_1'}, y_{x_1}, y_{x_2'}, y_{x_2}, y_{x_1', x_2'}, y_{x_1', x_2}, y_{x_1, x_2'}, y_{x_1, x_2}\}. \tag{103}$$

$$Y_{\mathbf{X}^*}(\mathbf{u}_2) = \{y, y'_{x_1'}, y_{x_1}, y'_{x_2'}, y_{x_2}, y'_{x_1', x_2'}, y'_{x_1', x_2}, y'_{x_1, x_2'}, y_{x_1, x_2}\}. \tag{104}$$

From Eqs. (103) *and* (104)*, we can construct quantities that align with human intuition for causes of* y*. For instance, given that no change to elements of* $\mathbf{X}$ *can induce a change in* Y *in* $\mathbf{u}_1$*, we can conclude that in this world, neither* x_1, x_2 *are causes of* y*. On the contrary, given that that setting* $X_1 = x_1'$ *changes* $Y(\mathbf{u}_2) = 1$ *to* $Y_{x_1'}(\mathbf{u}_2) = 0$*, or more concisely,*

$$Y(\mathbf{u}_2) - Y_{x_1'}(\mathbf{u}_2) \neq 0, \tag{105}$$

we can infer x_1 *is a cause of* y *in the world induced by* $\mathbf{u}_2$*. Similarly, given that*

$$Y(\mathbf{u}_2) - Y_{x_2'}(\mathbf{u}_2) \neq 0, \tag{106}$$

we can infer that x_2 *is a cause of* y *in the world induced by* $\mathbf{u}_2$*. Since* Y *does not change upon intervention in* u_1*, we can infer that in this world there are no causes of* Y *among the explanatory variables.*

The prior probability distribution $P(\mathbf{U}|\mathbf{e})$ *is necessary to determine the probability and degree to which each explanatory variable causes the outcome. Specifically, we know that* $P(\mathbf{U}|\mathbf{e})$ *assigns non-zero probability mass to only* $\mathbf{u}_1, \mathbf{u}_2$*, with* $P(\mathbf{u}_1|\mathbf{e}) = 0.05$ *and* $P(\mathbf{u}_2|\mathbf{e}) = 0.95$*. We can*

conclude that there is a $P(\mathbf{u}_2|\mathbf{e}) = 0.95$ probability that each of x_1, x_2 is a cause of y, given that neither is a cause under $\mathbf{u}_1$ and both are causes under $\mathbf{u}_2$.

Next, we require knowledge of the settings of explanatory variables $\mathbf{X}(\mathbf{u})$ under each setting of $\mathbf{u}$ consistent with $\mathbf{e}'$, the knowledge of the explainee. In this case, since $\mathbf{e}' = \varnothing$, we have the set of counterfactuals:

$$\mathbf{X}(\mathbf{u}') = \{X_1 = U_1, X_2 = U_2\}, \tag{107}$$

where each $\mathbf{u}' \sim P(\mathbf{U}|\mathbf{e}')$. From this information, we can infer that the probability of lightning not striking $P(X(\mathbf{u}') = x_1') = 0.99$, and the probability of a tree in the rainforest being wet $P(x_2') = 0.5$. From this information, we see an asymmetry between x_1, x_2 that means that the lightning strike x_1 is a more abnormal cause and therefore better suited for an explanation. □

We argue that the four pieces of information $\langle Y_{\mathbf{X}^*}, \mathbf{X}(\mathbf{u}), P(\mathbf{U}|\mathbf{e}'), P(\mathbf{U}|\mathbf{e})\rangle$ are sufficient to answer all Why queries insofar as they can be answered from an SCM; specifically, we argue that they are sufficient to capture the causal history [21] of the event explanandum $Y = y$. We term this subset of information the *event explanatory basis* of $\mathcal{M}$ for why query w, and provide a general definition below.

Definition 6 (Event Explanatory Basis)**.** *The event basis of SCM $\mathcal{M}$ with respect to query $\mathrm{Why}(y|\mathbf{X}; \mathbf{e}' \to \mathbf{e})$ may be written as $\mathcal{C}(\mathcal{M}, w) := \langle Y_{\mathbf{X}^*}, \mathbf{X}, P(\mathbf{U}|\mathbf{e}'), P(\mathbf{U}|\mathbf{e})\rangle$, where $Y_{\mathbf{X}^*}$ is the event counterfactual basis, defined as:*

$$Y_{\mathbf{X}^*}(\mathbf{u}) := \{Y_{\mathbf{z}}(\mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X}, \mathbf{z} \in \mathcal{D}_{\mathbf{Z}}\}. \tag{12}$$

Note that $\mathbf{X}$ above denotes the function $\mathbf{X}(\mathbf{u}) := \{X(\mathbf{u}) : X \in \mathbf{X}\}$. □

We argue that the event explanatory basis contains all of the information we need from the SCM to answer our Why query: the counterfactuals of Y, observational settings of $\mathbf{X}$, and conditional distributions of the exogenous variable given the observations of the explainer $\mathbf{e}$ and explainee $\mathbf{e}'$.

D.3 Multivariate causes

Equipped with the Why query and the event counterfactual basis, we return to our core argument that an explanation of an event is a summary of its causes. We revisit important aspects of each cause of the event: whether or not it is present, the magnitude of the causal effect, and the likelihood of the cause.

We showed briefly in Ex. 1 how to use differences in elements of the event counterfactual basis to infer causation in specific worlds $(\mathcal{M}, \mathbf{u})$. We generalize this notion of differencing effects by introducing the *natural total effect* (NTE) of a set of variables on the outcome. Intuitively, the NTE is the effect of changing variables $\mathbf{Z}$ on Y, setting them to their values $\mathbf{Z}(\mathbf{u}')$ in world $\mathbf{u}'$ rather than their actual values $\mathbf{Z}(\mathbf{u})$. Using the NTE, we can construct the NTE basis, which we show is equivalent to the event counterfactual basis, given an observed value of $Y(\mathbf{u})$.

Definition 7 (Natural Total Effect (NTE) Basis)**.** *Consider SCM $\mathcal{M}$, Why query w, event counterfactual basis Y_*, and explanatory variable subset $\mathbf{Z} \subseteq \mathbf{X}$. The unit-level natural total effect of $\mathbf{Z}$ on Y with respect to baseline $\mathbf{u}'$ and knowledge $\mathbf{u}$ is defined as*

$$\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = Y_{\mathbf{Z}(\mathbf{u})}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}). \tag{13}$$

Assume that for any $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there is $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ inducing $\mathbf{V}(\mathbf{u}') = \mathbf{v}$. Then the NTE basis of $\mathcal{M}$ for why query w is defined as

$$\mathcal{B}_{\mathrm{NTE}}^{\mathcal{M},w}(\mathbf{u}) := \{\mathrm{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) : \mathbf{Z} \subseteq \mathbf{X}, \mathbf{u}' \in \mathcal{D}_{\mathbf{U}}\}. \tag{14}$$

□

Fig. 7, left, illustrates the NTE basis in a simple setting with $|\mathbf{X}| = 3$. We first illustrate that the NTE basis is informationally equivalent to the event counterfactual basis, assuming the actual value of the event explanandum $Y(\mathbf{u})$ is known: we do not lose information by transitioning to the NTE basis.

Lemma 1 (NTE basis equivalence)**.** *Consider SCM $\mathcal{M}$ and why query w. Assume that for every setting of observed variables $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there exists unobserved setting $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ such that $\mathbf{V}(\mathbf{u}) = \mathbf{v}$. Then, the NTE basis $\mathcal{B}_{\mathrm{NTE}}^{\mathcal{M},w}$ and the value $Y(\mathbf{u})$ uniquely determine and are uniquely determined by the event counterfactual basis $Y_{\mathbf{X}^*}$.* □

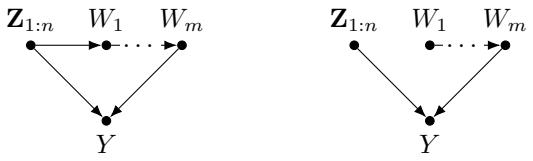


Figure 8: Interview graphs for companies A (left) and B (right).

We show that every NTE is necessary for the purpose of inferring causation in the example below by showing that each NTE is needed to distinguish between two causal worlds $(\mathcal{M}_1, \mathbf{u}_1)$ and $(\mathcal{M}_2, \mathbf{u}_2)$ where our perception of the causes of event Y differ. The causal graph for the example below is shown in Fig. 8.

Example 13 (Interview)**.** *Alice is searching for a job at companies A and B. At each company, there are $n > 0$ senior managers ($\mathbf{Z}_{1:n}$) and $m \geq 0$ stages in the interview process ($\mathbf{W}_{1:m}$). Alice will receive a job offer from a company if any manager refers her ($Z_i = 1$ for any $i \in [n]$) or if she passes her final stage interview ($W_m = 1$) with the company, assuming $m > 0$.*

In actuality, Alice is an exceptional candidate, and at both companies A and B, all senior managers refer her for the job ($\mathbf{Z} = \mathbf{1}_n$), and she passes all interview stages ($\mathbf{W} = \mathbf{1}_m$). As a result, she receives job offers from both companies ($Y = 1$).

However, the companies differ slightly in their interview procedures. At company A, a senior manager referral is needed for Alice to pass her first interview ($W_1 = 1$). On the other hand, at company B, Alice applies to the company separately from her referrals and passes her first interview. The functions corresponding to the SCMs for each company, $\mathcal{M}_A, \mathcal{M}_B$ are shown below. We consider the actual setting $\mathbf{u} = \{\mathbf{U_Z} = \mathbf{1}_n, \mathbf{U_W} = 2 \cdot \mathbf{1}_m\}$. Note that the functions are parameterized by company $C \in \{A, B\}$, and functional differences between the companies are shown in red and green, respectively.

$$\mathcal{F}_C(n,m) = \begin{cases} Z_i & = u_Z^i & & \forall i \in [n] \\ W_j & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0] & C = A \wedge u_W^j = 2 \wedge j = 1 \\ 1 & C = B \wedge u_W^j = 2 \wedge j = 1 \\ W_{j-1} & u_W^j = 2 \wedge j > 1 \\ u_{W^j} & u_W^j \in \{0,1\} \end{cases} & & \forall j \in [m] \\ Y & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0 \vee W_m = 1] & m > 0 \\ \mathbf{1}[\mathbf{Z} \neq 0] & m = 0 \end{cases} & & \end{cases} \tag{108}$$

We argue that the causes (and therefore causal histories) of Y at companies A and B are substantively different. Specifically, at company A, only Alice's n referrals are causes of her job offer; the interview process wouldn't have occurred without them. On the other hand, at company B, Alice's n referrals are also causes of her job offer, but every step of the interview process is also a cause of her job offer, given that she interviewed independently from receiving referrals. □

We generalize the intuition explored in Ex. 13 to all NTEs, illustrating that every NTE is necessary to distinguish $\mathcal{M}_A$ from $\mathcal{M}_B$, where our assessments of causation differ.

Lemma 2 (NTE necessity)**.** *Consider SCMs $\mathcal{M}_A, \mathcal{M}_B$, described by the following functions, for $C \in \{A, B\}$ and $n \in \mathbb{Z}^+$, $m \in \mathbb{Z}^{\geq 0}$*

$$\mathcal{F}_C(n,m) = \begin{cases} Z_i & = u_Z^i & & \forall i \in [n] \\ W_j & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0] & C = A \wedge u_W^j = 2 \wedge j = 1 \\ 1 & C = B \wedge u_W^j = 2 \wedge j = 1 \\ W_{j-1} & u_W^j = 2 \wedge j > 1 \\ u_{W^j} & u_W^j \in \{0,1\} \end{cases} & & \forall j \in [m] \\ Y & = \begin{cases} \mathbf{1}[\mathbf{Z} \neq 0 \vee W_m = 1] & m > 0 \\ \mathbf{1}[\mathbf{Z} \neq 0] & m = 0 \end{cases} & & \end{cases} \tag{18}$$

and why query $w = \text{Why}(Y = 1|\mathbf{Z} = \mathbf{1}_n, \mathbf{W} = \mathbf{1}_m)$. *Every element of the* NTE *basis* $\mathcal{B}_{\text{NTE}}^{\mathcal{M},w}(\mathbf{u})$ *for units* $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ *consistent with observed events is necessary to distinguish between* $\mathcal{M}_A(n, m)$ *and* $\mathcal{M}_B(n, m)$ *under at least one setting of* n, m. □

For clarity, our argument following Lem. 2 is not that every NTE is necessary to distinguish every SCM from at least one SCM in the set of SCMs for which human evaluations of causation differs. Rather, we argue that every NTE is necessary to distinguish at least two SCMs in which human evaluations of causation differ: $\mathcal{M}_{\text{A}}$ and $\mathcal{M}_{\text{B}}$.

This leads us to our formal statement of the necessity and sufficiency of the NTE.

Theorem 5 (NTE necessity and sufficiency)**.** *Consider SCM* $\mathcal{M}$ *and why query* $w = \text{Why}(y|\mathbf{x})$. *Every element of the* NTE *basis* $\mathcal{B}_{\text{NTE}}^{\mathcal{M},w}(\mathbf{u})$ *for actual units* $\mathbf{u} \in \mathcal{D}_{\mathbf{U}}$ *consistent with observed events is necessary to distinguish between* $\mathcal{M}_A(n, m)$ *and* $\mathcal{M}_B(n, m)$ *under at least one setting of* n, m. *Furthermore, the* NTE *basis is sufficient to describe the causal history of* $Y = y$ *contained in the event counterfactual basis* $\mathbf{Y}_{\mathbf{X}^*}$. □

In this subsection, we have introduced the NTE and NTE basis to describe the effects of any subset of explanatory variables on the outcome. We have shown that the NTE basis preserves all information in the event counterfactual basis given the event explanandum $Y(\mathbf{u})$ and, furthermore, that every single NTE in the basis is necessary to infer causation in at least one SCM.

D.4 Generalized direct effects

Although the NTE basis is conceptually useful for ascertaining causation, it is exponential in size with respect to the number of input variables. Thus, we turn our attention towards the concept of a variable-specific contribution to an NTE. We motivate our definition with an example.

Example 14 (Variable interaction effects)**.** *Consider the following world.*

$$\mathbf{U} = \{U_Z = 0, U_X = 0, U_W = 0\} \tag{109}$$

$$\mathcal{F} = \begin{cases} V := U_V, \forall V \in \{Z, X, W\} \\ Y := \alpha Z + \beta X + \gamma W - \delta XW \end{cases} \tag{110}$$

Consider $\mathbf{u}' = \{U = -1 : U \in \mathbf{U}\}$. *We first note a general formula for the* NTE*, where* $\mathbf{Z} \subseteq \{Z, X, W\}$

$$\text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) = \alpha\mathbf{1}[Z \in \mathbf{Z}] + \beta\mathbf{1}[X \in \mathbf{Z}] + \gamma\mathbf{1}[W \in \mathbf{Z}] + \delta\mathbf{1}[\{X, W\} \subseteq \mathbf{Z}] \tag{111}$$

Below, we compute the four NTE*s containing* X *and the four excluding* X*:*

$$\text{NTE}(\{X\}, Y|\cdot) = \beta \qquad \text{NTE}(\{\}, Y|\cdot) = 0 \tag{112}$$

$$\text{NTE}(\{X, Z\}, Y|\cdot) = \alpha + \beta \qquad \text{NTE}(\{Z\}, Y|\cdot) = \alpha \tag{113}$$

$$\text{NTE}(\{X, W\}, Y|\cdot) = \beta + \gamma + \delta \qquad \text{NTE}(\{W\}, Y|\cdot) = \gamma \tag{114}$$

$$\text{NTE}(\{X, Z, W\}, Y|\cdot) = \alpha + \beta + \gamma + \delta \qquad \text{NTE}(\{Z, W\}, Y|\cdot) = \alpha + \gamma \tag{115}$$

Lems. 1 and 2 suggest it would be incorrect to take only $\text{NTE}(X, Y|\mathbf{u}' \to \mathbf{u}) = \beta$ *to represent the full effect of* X *on* Y*. This is confirmed upon examination by the fact that* δ *affects how* X *affects* Y *differs from its expectation but is not included in the expression. Therefore, we consider the set of all* NTE*s and aim to determine the contribution of* X *to each of the four* NTE*s containing it in the left column; to do so, we compute the difference between each* NTE *and the corresponding* NTE *excluding* X*. We call this difference the generalized direct effect (*GDE*) of* X *on* Y *with respect to some subset* $\mathbf{Z}$*.*

$$\text{GDE}^{\mathbf{Z}=\varnothing}(X, Y|\cdot) = \text{NTE}(\{X\}, Y|\cdot) - \text{NTE}(\{\}, Y|\cdot) = \beta \tag{116}$$

$$\text{GDE}^{\mathbf{Z}=\{Z\}}(X, Y|\cdot) = \text{NTE}(\{X, Z\}, Y|\cdot) - \text{NTE}(\{Z\}, Y|\cdot) = \beta \tag{117}$$

$$\text{GDE}^{\mathbf{Z}=\{W\}}(X, Y|\cdot) = \text{NTE}(\{X, W\}, Y|\cdot) - \text{NTE}(\{W\}, Y|\cdot) = \beta + \delta \tag{118}$$

$$\text{GDE}^{\mathbf{Z}=\{Z,W\}}(X, Y|\cdot) = \text{NTE}(\{X, Z, W\}, Y|\cdot) - \text{NTE}(\{Z, W\}, Y|\cdot) = \beta + \delta \tag{119}$$

We observe that in cases when $W \in \mathbf{Z}$*, the* GDE *captures interaction effects between* X *and* W*. This supports our claim that each* $\text{GDE}^{\mathbf{Z}}(X, Y|\cdot)$ *captures the contribution of* X *to the effect of the corresponding subset containing it* $\mathbf{Z} \cup \{X\}$ *on* Y*,* $\text{NTE}(\mathbf{Z} \cup \{X\}, Y\cdot)$. □

We formally define the *generalized direct effect* (GDE) below.

Definition 11 (Generalized Direct Effect (GDE) Basis)**.** *Consider SCM $\mathcal{M}$,* Why *query w, event counterfactual basis Y_*, explanatory variable subset $\mathbf{Z} \subseteq \mathbf{X}$, and explanatory variable of interest $X \in \mathbf{X} \setminus \mathbf{Z}$. The unit-level generalized direct effect of X on Y with adjustment set $\mathbf{Z}$, baseline $\mathbf{u}'$, and knowledge $\mathbf{u}$ is defined as*

$$\begin{aligned}\text{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) &= \text{NTE}(\mathbf{Z} \cup \{X\}, Y|\mathbf{u}' \to \mathbf{u}) - \text{NTE}(\mathbf{Z}, Y|\mathbf{u}' \to \mathbf{u}) &(57)\\ &= Y_{\mathbf{Z}(\mathbf{u}')}(\mathbf{u}) - Y_{\mathbf{Z}(\mathbf{u}'),X(\mathbf{u}')}(\mathbf{u}). &(58)\end{aligned}$$

Assume that for any $\mathbf{v} \in \mathcal{D}_{\mathbf{V}}$, there is $\mathbf{u}' \in \mathcal{D}_{\mathbf{U}}$ inducing $\mathbf{V}(\mathbf{u}') = \mathbf{v}$. Then the GDE *basis of $\mathcal{M}$ for why query w is defined as*

$$\mathcal{B}^{\mathcal{M},w}_{\text{GDE}}(\mathbf{u}) = \{\text{GDE}^{\mathbf{Z}}(X, Y|\mathbf{u}' \to \mathbf{u}) : X \in \mathbf{X}, \mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}, \mathbf{u}' \in \mathcal{D}_{\mathbf{U}})\} \quad (59)$$

□

Fig. 7, center, illustrates the GDE basis in a simple setting with $|\mathbf{X}| = 3$. Highlighted in red (Z), green (X), and dark blue (W), respectively, are the GDEs expressing the contribution of the corresponding variable to each NTE whose subset contains that value.

We next observe that the NTE of all variables on the outcome may be decomposed into GDEs. In this way, we clearly illustrate that the set of GDEs captures all variation in the set of NTEs. Revisiting Ex. 14, consider the ordering $\{Z, X, W\}$; we show that the following equality holds:

$$\begin{aligned}\text{NTE}(\{Z, X, Y\}, Y|\mathbf{u} \to \mathbf{u}') &= (\alpha) + (\beta) + (\gamma + \delta) &(120)\\ &= \text{GDE}(Z, Y|\mathbf{u} \to \mathbf{u}') + \text{GDE}(Z, Y|\mathbf{u} \to \mathbf{u}') + \text{GDE}(Z, Y|\mathbf{u} \to \mathbf{u}') &(121)\end{aligned}$$

Following the intuition that variable-specific GDEs decompose the NTE, we prove that the GDE preserves all information contained in the set of NTEs; in other words, no information is lost when transitioning to this variable-specific representation of causal effects on Y. To this end, we first note that, following Thm. 2, the NTE can be decomposed into a sum of GDEs. This leads to the implication that the NTE and GDE bases are equivalent.

Theorem 8 (NTE-GDE equivalence)**.** *The* GDE*s basis uniquely determines and is uniquely determined by the* NTE *basis.* □

In this section, we have introduced the global explanatory basis, the technical Why query and its corresponding event explanatory basis, and the counterfactual NTE and GDE queries as tools to ascertain the causes of an event explanandum. We have argued via Thm. 8 that the sets are both necessary and sufficient to infer causation from the underlying SCM.

E Experiments

E.1 Methodology

In this section, we prove the Explanatory Impossibility Theorem (Thm. 9): we prove that substantial causal assumptions are needed to infer L_3 SVs from data. Following this motivation, we introduce Alg. 1 to bound L_3 SVs from data and assumptions in the form of a causal diagram.

E.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 [38].

Definition 12 (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 $\mathcal{M} \in \Omega'$,*

$$a \leq f(\mathcal{M}) \leq b. \quad (60)$$

$[a, b]$ 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 [29] 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:

$$\begin{aligned}\mathrm{NTE}(X, Y|\mathbf{e}) &= \mathbb{E}[Y|\mathbf{e}] - \mathbb{E}[Y_{P(X)}|\mathbf{e}] && (122)\\ &= 1 - (P(x)P(y|\mathbf{e}) + P(x')P(y_{x'}|\mathbf{e})) && (123)\\ &= P(x')PN(x, y|\mathbf{e}). && (124)\end{aligned}$$

This implies that in certain binary settings, we may use existing bounds on the PN [38] to bound counterfactual Shapley values. Particularly, when Markovianity holds, we have:

$$\max\left(0, 1 - \frac{P(y_{x'})}{P(y_x)}\right) \leq PN(x, y) \leq \min\left(1, \frac{P(y'_{x'})}{P(y_x)}\right). \tag{125}$$

Below, we illustrate an application of these bounds.

Example 15 (Two-variable binary Markovian chain)**.** *Consider a binary Markovian SCM with observed variables $\{X_1, X_2, Y\}$ with an observational distribution factorizing as:*

$$P(\mathbf{v}) = P(x_1)P(x_2|x_1)\mathbf{1}[y = x_2]. \tag{126}$$

We cannot infer any bounds on ϕ^{L_3} when the assumption of Markovianity is removed [38]. 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 Eq. (125)*, observing that:*

$$\begin{aligned}\phi_1^{L_1} = \phi_1^{L_2} &= P(x'_1)(P(y_{x_1}) - P(y_{x'_1})) && (127)\\ &\leq P(x'_1)\max\left(0, 1 - \frac{P(y_{x'})}{P(y_x)}\right) && (128)\\ &\leq \phi_1^{L_3}, && (129)\end{aligned}$$

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 $\mathbf{V} = \{X_1 = 1, X_2 = 1, Y = 1\}$:

$$\mathcal{M}_1 = \begin{cases} U_1, U_2 & \sim \mathrm{Bern}(0.5) \\ X_1 & = U_1 \\ X_2 & = X_1 \oplus U_2 \\ Y & = X_2 \end{cases} \tag{130}$$

$$\mathcal{M}_2 = \begin{cases} U_1, U_2 & \sim \mathrm{Bern}(0.5) \\ X_1 & = U_1 \\ X_2 & = U_2 \\ Y & = X_2 \end{cases} \tag{131}$$

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:

$$\begin{aligned}\phi_1^{L_3}(\mathcal{M}_1) &= \frac{1}{2}(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]) && (132)\\ &= \frac{1}{2}(1 - 0) = \frac{1}{2} \\ \phi_1^{L_3}(\mathcal{M}_2) &= \frac{1}{2}(\mathbb{E}[Y|\mathbf{v}] - \mathbb{E}[Y_{P(X_1)}|\mathbf{v}]) && (133)\\ &= \frac{1}{2}(1 - 1) = 0\end{aligned}$$

This corresponds to the bounds obtained in Eq. (125)*, 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.

Theorem 9 (Explanatory Impossibility Theorem)**.** *Consider positive observational distribution* $P(\mathbf{V})$ *with binary variables,* Ω'*, the set of all SCMs* $\mathcal{M}$ *that induce the distribution, and why query* $\text{Why}(y|\mathbf{x})$*, where* $\mathbf{V} = \mathbf{X} \cup \{Y\}$*.* Ω' *yields no information about any* L_3 *Shapley value* $\phi_X^{L_3}$ *for* $X \in \mathbf{X}$*.* □

In this subsection, we have motivated the problem of bounding explanatory variable attributions in Thm. 9, 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.

E.1.2 Bounding Counterfactual Shapley values

In this subsection, we introduce Alg. 1, a method to bound counterfactual Shapley values, extending the counterfactual identification algorithm of [40]. 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 1: Bounding counterfactual Shapley values

Input: Query $q : \Omega \to \mathbb{R}$, variable of interest $X \in \mathbf{V}$, L_2 datasets $\mathbb{Z}(\mathcal{M})$, and causal diagram $\mathcal{G}$.

Output: 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 $\phi_X^{\min}, \phi_X^{\max}$

E.2 Experimental Setup

In this appendix, we provide additional details on our experimental setup and approach, complementing the experiments described in Sec. 4 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. E.2.1-E.2.3). 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. E.2.4).

E.2.1 Color MNIST – Ground Truth

We first described the ground truth SCM for the color MNIST experiment, based on [28]. 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 28x28 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:

$$P(U) = \begin{cases} u_Y & \sim \text{Unif}(\{0, \dots, 9\}) \\ u_X & \sim \text{Unif}(0, 1) \\ u_I^i & \sim \text{MNIST}(i) \end{cases} \tag{134}$$

$$\mathcal{F}_{\beta,f} = \begin{cases} X & = \left(\frac{u_Y}{9} + 0.5\Phi(u_X) + \beta\right) \mod 1 \\ Y & = u_Y \\ I & = \text{hsv\_to\_rgb}\left(u_I^{Y=y}, \frac{u_Y}{9}, X\right) \\ \hat{Y} & = \hat{f}(I) \end{cases} . \tag{135}$$

Here, β represents a hue shift parameter, f represents an image classifier, $\text{MNIST}(i)$ 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 28x28 RGB image. The causal diagram for this SCM is shown in Fig. 3a.

The is aim to explain two LeNet [20] 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 $\text{Why}(\hat{y}|x, y, i)$. The detailed interpretation of the different explanation methods is described in the main text (see Sec. 4.1).

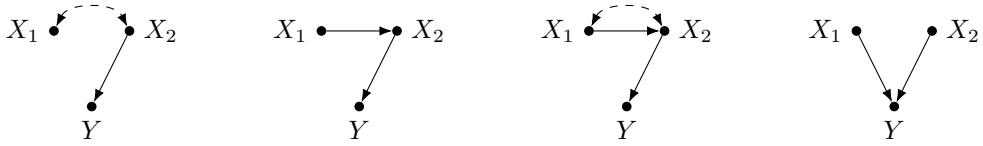
E.2.2 CelebA – Ground Truth

We next describe the ground truth SCM for the CelebA experiment, based on [19]. 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:

$$\mathcal{F} = \begin{cases} S & = U_S \\ M & = \begin{cases} 0 & U_M = 0 \\ s & U_M = 1 \\ 1 & U_M = 2 \end{cases} \\ I & = U_I^{s,m} \\ \hat{M} & = f_{\hat{M}}(I) \end{cases} \tag{136}$$

$$P(\mathbf{U}) = \begin{cases} U_S & \sim \text{Bern}(0.5) \\ U_M & \sim \text{Categorical}([0.05, 0.9, 0.05]) \\ U_I & \sim \text{CelebA-HQ}(\texttt{Smiling, Mouth\_Slightly\_Open}) \end{cases} . \tag{137}$$

Here, $\text{CelebA-HQ}(\texttt{Smiling, Mouth\_Slightly\_Open})$ denotes a distribution over a list of four CelebA-HQ images, such that $U_I^{s,m}$ denotes an image where $\texttt{Smiling}\ S = s$ and



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

Figure 9: Causal diagrams for toy experiments.

`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. 4.

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 $\text{Why}(\hat{m}|s, m, i)$. The detailed interpretation of the different explanation methods is described in the main text (see Sec. 4.2).

E.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 $\text{Bern}(0.5)$. $Y = 1$ 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 Fig. 9. We next discuss each SCM in order.

Spurious SCM (Shark Attacks, Ex. 5 and Fig. 9a) 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 Fig. 9a. The ground truth SCM is given by:

$$\mathcal{M}_1 = \begin{cases} X_1 & := U_{12} \\ X_2 & := U_{12} \vee U_2 \\ Y & := X_2 \end{cases} \tag{138}$$

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. 10 (first column, blue bar), we observe that L_1 Shapley values give X_1 a non-zero attribution, violating the property of causal admissibility (Prop. 1). Conversely, L_2, L_3 Shapley values satisfy admissibility in this example, giving a zero attribution to the variable X_1.

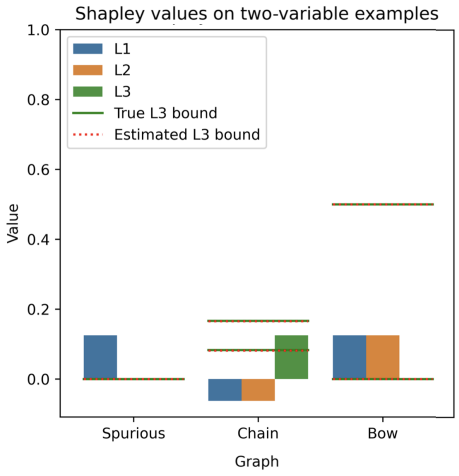


Figure 10: SVs for X_1 in toy experiment SCMs. Green and dotted red lines denote true and estimated bounds. Error bars are negligible.

Chain SCM (Fig. 9b) The causal diagram for the chain SCM is shown in Fig. 9b, and the SCM is given by:

$$\mathcal{M}_2 = \begin{cases} X_1 & := U_1 \\ X_2 & := (U_2^1 \wedge X_1) \vee (\neg U_2^1 \wedge (U_2^2 \vee \neg X_1)) \\ Y & := X_2 \end{cases} \tag{139}$$

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. 10 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 (Fig. 9c) The causal diagram for the bow SCM is shown in Fig. 9c, and the SCM is given by

$$\mathcal{M}_3 = \begin{cases} X_1 & := U_{12} \\ X_2 & := (U_2^1 \wedge U_2^2) \vee ((U_2^1 \vee U_2^2) \wedge (X_1 \vee U_{12})) \\ Y & := X_2 \end{cases} \tag{140}$$

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. 10). 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.

E.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. E.1.2. Specifically, we train a conditional diffusion model with 4000 steps [13] to model $P(I|X, Y, U_Y)$ for 150 epochs. At inference time, we reduce this to 25 steps [35], still achieving realistic results. The bounds obtained on the SVs are shown in Fig. 3d 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. E.1.2), 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 Tian and Pearl [38]. 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 Fig. 10, 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.