PE-SHAP: Causally Interpretable Path-Wise Shapley Explanations

Maksym Buleshnyi[1], **Joshua Loftus**[2], **Sakina Hansen**[2]

[1]Department of Computer Science, Ukrainian Catholic University
[2]Department of Statistics, London School of Economics London
maksym.buleshnyi.pn@ucu.edu.ua, j.r.loftus@lse.ac.uk, s.a.hansen1@lse.ac.uk

Abstract

Explainability plays a critical role in ensuring that AI and machine learning models are transparent, trustworthy, and actionable - especially in high-stakes domains. However, many popular explanation techniques, such as Shapley values, focus on predictive rather than causal explanations. This limits their ability to inform decisions or policy. Recently, researchers have introduced variants of causally-aware Shapley values. In this paper, we extend a path-wise causal explanation framework for binary treatment settings, by introducing a new effect designed to better capture mediation. Additionally, we leverage doubly robust estimators to improve both reliability and efficiency. We validate our framework through simulations and real-world case studies, demonstrating its practical utility. We also show how individual-level explanations can be aggregated to estimate population-level effects, which allows broader causal analysis.

Introduction

In recent years, the need for explainability in machine learning has become very important, especially in high-impact domains like law or healthcare. In these settings it is important to understand not only the outputs of predictive algorithms but also how and why decisions are made (Sadeghi et al. 2024).

One of the most widely used model-agnostic frameworks for model explainability is based on Shapley values. Since explanations are intended for human interpretation, it is essential that the resulting attributions have a coherent and logical interpretation. In real-world scenarios features are often interconnected through complex causal connections. Accounting for these causal relationships is crucial for making explanations more interpretable and trustworthy. This need has led to the development of several causal-aware extensions of the classical Shapley value framework.

Traditional Shapley approaches, such as Marginal or Conditional Shapley values have been shown to provide very limited causal interpretability (Rozenfeld 2024; Janzing, Minorics, and Blöbaum 2020), (Chen et al. 2020). More recent methods, like Asymmetric Shapley values (Frye, Rowat, and Feige 2020) and Causal Shapley values (Heskes et al. 2020), aim to incorporate causality but are still limited in providing path-specific explanations. While Causal Shapley values can decompose feature attribution into direct and indirect effects, this level of decomposition often lacks the granularity required to identify biases along individual causal paths. In cases where multiple causal paths contribute differently, the aggregated indirect effect may hide bias, making it difficult to uncover the true sources of influence within the model.

Path-wise interpretability offers a promising approach to uncover such biases by tracing distinct causal pathways. In particular, recent frameworks like Path-Wise Shapley (PW-SHAP) (Ter-Minassian et al. 2023) integrate user-defined causal graphs with Shapley value-based methods to decompose feature effects along specific causal paths. However, the proposed effect may obscure the contribution of a mediator, especially when complex interactions with moderators or downstream processes are involved, potentially leading to unexpected directions of attribution.

To address this, we introduce , an extension of PW-SHAP that is specifically designed to improve causal interpretation in mediation settings. We directly incorporate a doubly robust estimator into the computation of Shapley effects. approach is designed to provide both local, path-wise interpretable explanations for individual predictions and global, population-level causal insights. Our contributions are as follows:

- We propose , an effect that more faithfully captures mediation-related causal properties.
- We propose doubly robust estimators for the effects enabling more direct and efficient estimation.
- Demonstrate how aggregated effects can be used to extract meaningful population-level insights.
- Conduct comparative analysis with existing approaches.

Causality for Model Explainability

Following the notation of Janzing, Minorics, and Blöbaum (2020), let a model f operate on input features $X = \{X_1, X_2, \ldots, X_N\}$, for each input feature corresponds to a real-world causal variable. In practice, these real-world variables may be connected through a causal graph, informed by domain knowledge.

Consider a model that predicts income using *age* (X_1), work experience (X_2), and *English proficiency* (X_3). The model may rely primarily on X_2 and X_3, with little or no direct use of X_1. Some feature attribution methods might then suggest that age has no influence on the prediction. However, both experience and English proficiency can be causally affected by age - making age a source of significant indirect effects on the outcome.

Focusing only on direct effects can thus obscure the broader causal impact of certain features. Causal explainability helps disentangle these relationships by distinguishing between features that affect the model output directly and those that act through causal pathways.

This decomposition can be further extended to analyze specific indirect paths (e.g., Age $\rightarrow$ Experience $\rightarrow$ Income or Age $\rightarrow$ English $\rightarrow$ Income), allowing more detailed and actionable interpretations.

Such causal improvements have been shown to provide actionable insights for achieving desired outcomes and supporting informed decision-making (Albini et al. 2022; Watson 2022).

Shapley Values

Shapley values were originally introduced by Shapley (1951) to fairly allocate a collective payoff among players based on their individual contributions. In the context of machine learning, each input feature is treated as a "player" in a cooperative game, and the objective is to quantify the contribution of each feature to a model's prediction.

Formally, let ϕ_i^f denote the attribution of feature i with respect to a model f. To ensure fairness, Shapley values satisfy four fundamental properties: Efficiency, Dummy Player, Symmetry, and Additivity (see Appendix for formal definitions). These axioms uniquely characterize the Shapley value.

We begin by defining some notation:

- N is the set of all feature indices.
- $S \subseteq N$ represents a coalition (subset) of features, where $\overline{S} = N \setminus S$ is the complementary set.
- $v : 2^N \rightarrow \mathbb{R}$ is a value function that assigns a real-valued payoff to any coalition S, quantifying its contribution to the model output.

The contribution of feature $i \notin S$ to coalition S is defined as:

$$\phi_{i,S}^f = v(S \cup \{i\}) - v(S).$$

Considering all possible permutations π of the feature set N, define

$$S_\pi^i = \{j \in N : j \prec_\pi i\}$$

as the set of features preceding i in permutation π. The Shapley value of feature i is then computed by averaging its marginal contributions over all permutations:

$$\phi_i^f = \sum_{\pi \in \Pi} \frac{1}{|N|!} \left[v(S_\pi^i \cup \{i\}) - v(S_\pi^i) \right].$$

This formulation effectively marginalizes the contribution of feature i across all possible combinations of other features, ensuring a fair attribution. The choice of the value function v plays a critical role in shaping the behavior of Shapley values; we elaborate on this in the following subsection.

Related Work

The interpretability of Shapley values, especially their causal meaning, critically depends on how the value function v is defined.

A common approach is the Marginal Shapley Values, where the value function is computed as the expected model output over the marginal distribution of features outside the coalition, $v(S) = \mathbb{E}[f(x_S, X_{\overline{S}})]$. This ignores dependencies between features and can lead to unrealistic feature combinations, capturing only direct effects.

In contrast, the Conditional Shapley Values conditions on the observed values of features outside the coalition and defines the value function as $v(S) = \mathbb{E}[f(x_S, X_{\overline{S}}) \mid X_S = x_S]$, thereby accounting for statistical dependencies and capturing both direct and indirect effects (Rozenfeld 2024). However, since this relies on correlations rather than causal relationships, it may attribute importance to features correlated with the outcome but not causally responsible.

Several extensions have been proposed to integrate causal information directly into the Shapley value framework. Asymmetric Shapley Values (Frye, Rowat, and Feige 2020) relax the symmetry property to incorporate known causal structures by computing contributions only over permutations that respect a given causal DAG.

Another approach, Causal Shapley Values (Heskes et al. 2020), modifies the value function to use the interventional distribution: $v(S) = \mathbb{E}[f(x_S, X_{\overline{S}}) \mid \text{do}(X_S = x_S)]$, explicitly incorporating the causal graph into the computation. Causal Shapley values allow the decomposition of feature attributions into direct and indirect effects .

These two key extensions, relaxing symmetry and using interventions, offer distinct causal improvements. Combined, they define Asymmetric Causal Shapley Values. Unlike the symmetric variant, which splits indirect effects between a feature and its mediators, the asymmetric approach assigns the full indirect effect to the root cause. See Appendix for theoretical explanation.

Besides focusing on causal properties, many recent studies have concentrated on enhancing robustness and improving estimation methods to make Shapley values more practical. For example, Manifold Restricted Shapley values (Taufiq, Blöbaum, and Minorics 2023) address a issue that most Shapley value methods evaluate model outputs on data points outside the training data distribution, which makes them vulnerable to adversarial attacks (Slack et al. 2020; Frye et al. 2020). The Do-Shapley approach (Jung et al. 2022) proposes estimators for Causal SHAP values when the model is inaccessible, meaning that not every combination is possible to provide in the model.

More recently, the Path-Wise Shapley Values method (Ter-Minassian et al. 2023), an on-manifold Shapley (conditional) values approach, decomposes feature attributions along causal paths. This is achieved by expressing coalition Conditional Shapley values in terms of the propensity score and pseudo-CATE (path-wise effect). By aggregating

these path-wise effects, the method quantifies how much each causal path contributes to the final prediction. For more details on properties and definitions of explained approaches see Appendix.

In this work, we propose , an extension of PWSHAP, a model-agnostic framework designed to provide both local, path-wise interpretable explanations of individual model predictions and global, population-level causal insights.

Our approach is motivated by the decomposition of Conditional Shapley values into weighting and CATE-like components, as introduced in (Ter-Minassian et al. 2023).

We consider a feature X_i as T the treatment (feature of interest), C as $X_{N\setminus i}$ the other covariates, and Y as the outcome of interest. Given a trained machine learning model f^*, our goal is to explain the predicted outcome

$$\hat{Y} = f^*(T, C) + \epsilon, \quad \text{with} \quad \mathbb{E}[\epsilon \mid T, C] = 0,$$

by uncovering the path-wise mechanisms through which T influences $\hat{Y}$.

Coalition Conditional Average Treatment Effect

Core components of the method is the Coalition Conditional Average Treatment Effect (CATE) (see in definition in Appendix). Note that in our definition, we aim to explain the impact of T on $\hat{Y}$ (model prediction) rather than on Y, as our focus is on model explainability rather than ground-truth causal effects.

[Coalition Conditional ATE (C-CATE)] Let $T \in \{t, t'\}$ be a binary treatment, $\hat{Y}$ the observed outcome, and $C_S \subseteq C$ a subset (coalition) of features. The Coalition Conditional Average Treatment Effect conditioned on $C_S = c_S$ is defined as the difference in the expected potential outcomes under treatments t and t', conditional on the coalition:

$$\Delta_t(c_S) = \mathbb{E}[\hat{Y}(t) \mid C_S = c_S] - \mathbb{E}[\hat{Y}(t') \mid C_S = c_S].$$

The expectation is taken over the distribution of features $C_{\bar{S}}$ not in the coalition, conditional on $C_S = c_S$.

Depending on the causal structure and the set of features we control for, this effect can provide a valuable property in a mediation setting.

[C-CATE as Controlled Direct Effect] Let T be a treatment, Y an outcome, and X_S a set of mediators. If C_S is not a descendant of any other variables except for T or any other variable included in C_S, then $\Delta_t(c_S)$ is a Controlled Direct Effect of T on $\hat{Y}$, having C_S fixed at c_S (proof in Appendix) :

$$\Delta_t(c_S) = CDE_{X_S}(t, t', x_S)$$

Path-specific effects

The PW-SHAP introduced in (Ter-Minassian et al. 2023) defines the Path-Wise Shapley Effect as the difference between two pseudo-CATEs: one where all covariates are held fixed, and another where all covariates except the feature of interest are fixed.

[Path-Wise Shapley Effect] The Path-Wise Effect along the causal path $T \to C_i \to \hat{Y}$ is defined as the difference between the "CATE" of T given all covariates, and the "CATE" of T when all covariates except C_i are held fixed:

$$\Psi_{T \to C_i \to \hat{Y}} = \Delta_t(c_{S^*}) - \Delta_t(c_{S^* \setminus i}),$$

where S^* denotes the index set of all covariates.

We extend this definition to a Portion Eliminated Path-Wise Shapley Effect, which, as we show later, can offer better insights in mediation settings. Motivated by the four-way decomposition of the Average Treatment Effect (VanderWeele 2014) we can extract path-wise information. $CDE_M(t, t', m)$ can be interpreted as the effect of the treatment T not mediated by M. Difference between ATE and $CDE_M(t, t', m)$ isolates the combined effects transmitted through the mediator, including pure mediation, interaction between the mediator and treatment (moderation), and their joint mediated interaction. This is the portion of the effect of the exposure that would remain if the mediator were fixed to 0, commonly refered as Portion Eliminated Effect (Robins and Greenland 1992; Hafeman and Schwartz 2009; VanderWeele 2013). Because these components represent distinct causal pathways by which the treatment influences the outcome indirectly via the mediator or through interactions involving the mediator, we consider this difference a path-wise effect. This effect has been shown to be useful in mediation settings (Hafeman and Schwartz 2009), (VanderWeele 2013).

[Portion Eliminated Shapley Effect] The Effect along the path $T \to C_i \to Y$ is defined as the difference between the total effect of treatment t and the effect when the mediator C_i is held fixed at level c_i. If $\Delta_t(C_i = c_i)$ satisfies Property , then:

$$\lambda_{T \to C_i \to \hat{Y}} = \Delta_t(\emptyset) - \Delta_t(C_i = c_i),$$

where $\Delta_t(\emptyset)$ denotes the total effect of treatment t, and $\Delta_t(C_i = c_i)$ represents the Controlled Direct Effect with C_i fixed at c_i.

Intuitively, we subtract from the total treatment effect the effect that is not mediated by C_i, by fixing C_i to c_i, this way isolating local path-wise effect through C_i.

This formulation can be generalized to paths of arbitrary length by fixing the values of all mediators along the path. We subtract the controlled direct effect along the specified path from the average treatment effect to obtain the contribution of that path.

Comparison of PW-SHAP and PE-SHAP Despite structural similarities, these approaches possess crucial differences.

When the PW-SHAP effect is computed, it controls for all variables except the one of interest. However, this can suppress the true impact of that feature, especially when it is involved in complex interactions with moderators or downstream processes. As a result, PW-SHAP may obscure the actual contribution of mediators and can even produce incorrect signs for mediation effects.

Method	NDE Estimate	NIE Estimate
S-learner	**0.0503 (0.002)**	**0.0284 (0.004)**
	0.0525 (0.007)	0.0404 (0.006)

Table 1: Comparison of Estimated Natural Direct Effect and Natural Indirect Effect on $\hat{Y}$ Across Methods. Estimates are reported with Mean Absolute Error and Monte Carlo Error shown in parentheses.

In contrast, PE-SHAP does not condition on all other variables, allowing it to preserve the mediator's role and its interactions.

As a result, PE-SHAP better captures the mediation effect and produces more accurate effect signs. In settings without moderation, the PE-SHAP effect directly corresponds to the pure mediation effect, while PW-SHAP gives a reversed sign. For a detailed derivation using a specific example, see the Appendix.

Additionally, PE-SHAP effects rely on less localized estimations than PW-SHAP, making them more stable and easier to compute - especially in low-density regions where PW-SHAP can struggle.

PE-SHAP effects can also be aggregated to estimate population-level effects, enabling broader and more interpretable analyses.

Population Effect While local effects provide valuable insight into individual-level behavior, it is also important to quantify effects at the population level. Population-level analysis allows access to the broader model fairness and potential biases that may not be evident from local explanations alone.

By aggregating effects over the distribution of feature realizations conditioned on the alternative treatment, we recover the Natural Indirect Effect of treatment T on the predicted outcome $\hat{Y}$ (see definition in Appendix).

[Effect Aggregation for NIE]

If C_S represents the set of all mediators and is not a descendant of any variables other than T or members of C_S itself, then the Natural Indirect Effect on $\hat{Y}$ can be expressed as the expectation of the Portion-Eliminated Shapley effects over the distribution of C_S under $do(T = t')$ (proof in Appendix):

$$\text{NIE} = \int P(C_S = c_S \mid do(T = t')) \cdot \lambda_{T \to C_S \to \hat{Y}} \, dc_S$$

Similarly Natural Direct Effect can be computed directly from weighted aggregation of Coalition CATE effects.

We experimentally compared estimators of the Natural Direct Effect and Natural Indirect Effect with the more classical S-learner approach, demonstrating that produces comparable results (see in Table 2). For wider analysis see Appendix.

Estimation

The proposed approach can integrate any CATE estimator as a foundational component, benefiting from the extensive research on this topic. This flexibility makes it easy to adapt the method to the specific needs of the use case and the nature of the data.

Method	NDE	NIE
S-learner-MLP	**0.0503 (0.002)**	**0.0284 (0.004)**
-GT	0.0546 (0.007)	0.0407 (0.006)
-MLP	0.0525 (0.007)	0.0404 (0.006)
-XGBoost	0.0547 (0.007)	0.0407 (0.006)

Table 2: Comparison of Estimated Natural Direct Effect and Natural Indirect Effect on Y Across Methods. Estimates are reported with Mean Absolute Error and Monte Carlo Error shown in parentheses.

To estimate the effects, we propose using a doubly robust estimator (Bang and Robins 2005), which combines models for both the treatment and outcome. This approach gives consistent and unbiased effect estimates if either the treatment model or the outcome model is misspecified. This greatly improves the method's robustness, which becomes challenging as data and causal connections grow more complex.

Doubly robust estimator demonstrated superior performance compared to the imputer method originally employed in the PW-SHAP implementation (Carpenter, Kenward, and Vansteelandt 2006). While the imputer relies primarily on outcome imputation to estimate missing or counterfactual values, the doubly robust method leverages information from both treatment assignment and outcome models.

[Doubly Robust Estimator of C-CATE]

A doubly robust (DR) estimator of the C-CATE is given by:

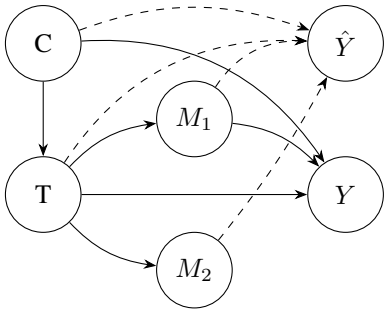
$$\hat{\Delta}_t^{\mathrm{DR}}(c_S) = \mathbb{E}\left[\hat{Y}^{\mathrm{DR}}(t) \mid C_S = c_S\right] - \mathbb{E}\left[\hat{Y}^{\mathrm{DR}}(t') \mid C_S = c_S\right],$$

where:

- $\hat{Y}_j^{\mathrm{DR}}(t) = \hat{\mu}_t(c^{(j)}) + \frac{\mathbb{I}(t^{(j)}=t)}{\hat{e}_t(c^{(j)})}\left(\hat{y}^{(j)} - \hat{\mu}_t(c^{(j)})\right)$ is the doubly robust estimate of the potential outcome,
- $\hat{\mu}_t(c^{(j)}) = \mathbb{E}[\hat{Y} \mid T = t, C = c^{(j)}]$ is the outcome regression model,
- $\hat{e}_t(c^{(j)}) = P(T = t \mid C = c^{(j)})$ is the estimated propensity score.

Note that we use $\hat{y}$ instead of the ground-truth outcome y, as our goal is to explain the model's output rather than the true causal effect.

Using only a single component of the estimator, such as the treatment model alone (IPW estimator) or an individual outcome model estimator like meta-learners (S-learner, T-learner, X-learner, or others), can be advantageous in cases



$$
\begin{aligned}
C &\sim \mathcal{N}(0.3,\, 0.5^2) \\
T &\sim \text{Bernoulli}(0.8 - C) \\
M_1 &= 0.5T + \mathcal{N}(0,\, 0.5^2) \\
M_2 &= 0.7T + \mathcal{N}(0,\, 0.5^2) \\
Y &= C + T - 0.8M_1^2 - 0.5M_1T \\
\hat{Y} &= \hat{f}(C, T, M_1, M_2)
\end{aligned}
$$

Figure 1: Causal DAG and corresponding structural equations

Method	$T \overset{direct}{\rightarrow} \hat{Y}$	$T \rightarrow M_1, M_2 \rightarrow \hat{Y}$	$T \rightarrow M_1 \rightarrow \hat{Y}$	$T \rightarrow M_2 \rightarrow \hat{Y}$
Causal Shapley	0.34(0.35)	-0.16(-0.18)	-	-
PW-SHAP*	0.79(0.78)	0.26(0.26)	0.32(0.33)	0.09(0.08)
PE-SHAP*	0.79(0.78)	-0.26(-0.26)	-0.17(-0.18)	0.06(0.07)
		Sample T = 1, C = 0.2, M_1 = 0.6, M_2 = 1		
Causal Shapley	0.47(0.49)	-0.17(-0.19)	-	-
PW-SHAP*	0.75(0.75)	0.22(0.22)	0.17(0.18)	0.07(0.07)
PE-SHAP*	0.75(0.75)	-0.22(-0.22)	-0.15(-0.15)	-0.05(-0.04)
		Sample T = 1, C = 0.5, M_1 = 0.6, M_2 = 1		
Causal Shapley	0.51(0.54)	-0.17(-0.19)	-	-
PW-SHAP*	0.85(0.85)	0.33(0.32)	0.28(0.28)	0.02(0.01)
PE-SHAP*	0.85(0.85)	-0.33(-0.32)	-0.31(-0.31)	-0.05(-0.04)
		Sample T = 1, C = 0.5, M_1 = 0.3, M_2 = 1		

Table 3: Shapley values comparison: The values indicate the effect on the GT model, with values in brackets showing the corresponding effect on the trained MLP model. Effects marked with '*' are computed using the DR estimator.

where we are confident in the correctness of the outcome or treatment model. These singular estimators may be beneficial, particularly since doubly robust estimators often exhibit higher variance.

Experiments

Path-wise Analysis

To evaluate our proposed method, we design experiments on a synthetic dataset generated from a known causal structure (see Figure 1). We consider a causal DAG with a binary treatment variable T, an outcome variable Y, mediator M_1, independent feature M_2, and a treatment confounder C. We generate samples using structural equations with additive noise, where all variables are continuous and normalized. M_1 not only mediates but also moderates the effect of treatment T on the outcome Y, with a negative influence on the prediction. However we aim to explain effect of T on $\hat{Y}$. Note that M_2 does not impact Y however it is an input of blackbox model $\hat{f}$.

This example demonstrates a critical limitation of Causal Shapley values - they can only capture direct and indirect effects. While the indirect effect is shown as a small positive value, path-wise approaches reveal that most of the effect occurs through M_1, whereas path through M_2 contributes almost no effect. This provides much better resolution of what is actually happening.

Both approaches showed a large effect for M_1 and a very small effect for M_2. While and PWSHAP produced results of similar magnitude, their signs were completely reversed. This difference arises because captures the effect of M_1 when computing the average treatment effect, while Path-wise Shapley values condition on all features, including M_1, removing its influence. PE-SHAP correctly captured the sign of the effect, aligning with expectations. For a more detailed theoretical explanation of this example, see Appendix.

Estimator Comparison

C-CATE Estimator	**MAE (MC)**
Iterative Imputer	0.191 (0.02)
DR Estimator	0.157 (0.03)

Table 4: Comparison of Mean Absolute Error (MAE) with Monte Carlo error (in parentheses) for different C-CATE estimators.

Method	Attribution
Causal SHAP	$+0.040$
Asymmetric Causal SHAP	$+0.040$
Conditional SHAP	-0.001
Asymmetric Cond. SHAP	$+0.019$
Marginal SHAP	$+0.037$

Table 5: Shapley Value Attribution for Gender

Method	Component	Attribution
Causal SHAP	ϕ^{direct}	$+0.037$
	ϕ^{indirect}	$+0.003$
PW-SHAP*	$\lvert\Psi_{T\to X_A\to\hat{Y}}\rvert$	$+0.053$
	$\lvert\Psi_{T\to X_S\to\hat{Y}}\rvert$	$+0.045$
PE-SHAP*	$\lvert\lambda_{T\to X_A\to\hat{Y}}\rvert$	$+0.036$
	$\lvert\lambda_{T\to X_S\to\hat{Y}}\rvert$	$+0.054$

Table 6: Causal Path-wise Attribution for Gender. Effects marked with $*$ are computed using the DR estimator.

To evaluate the robustness and accuracy of different strategies for estimating Coalition Conditional Average Treatment Effects, we compare two distinct approaches:

The original method introduced in the PW-SHAP framework, which estimates C-CATE using Conditional Shapley values combined with iterative imputation to approximate conditional distributions. And a direct estimation approach that uses a doubly robust estimator to compute C-CATE

Table 4 reports the Mean Absolute Error for both method, along with the corresponding Monte Carlo (MC) error in parentheses. The results show that the DR-based C-CATE estimator substantially outperforms the PWSHAP method, achieving lower estimation error.

Additional we made comparison under various model misspecification scenarios (when the outcome or treatment models are incorrectly specified) further demonstrate the improved stability and accuracy of the DR-based C-CATE approach.

The results empirically confirm the theoretical robustness of the Doubly Robust (DR) estimator. It is worth noting that the DR estimator can underperform relative to IPW or regression-based methods when there is high confidence in the correct specification of either the treatment or outcome model. This is primarily due to the higher variance typically associated with DR estimators. See comparison in Table 9 in Appendix.

Real-Data Case Study

In this section, we investigate the presence and propagation of gender-related bias in a credit risk prediction model trained on the German Credit dataset (Hofmann 1994). We focus on the role of the sensitive attribute Gender and how its influence is mediated through other features in the model. While gender is typically not considered confounded at the population level, the imbalance in this dataset indicates a selection effect that could introduce confounding (Arah 2019). The causal structure assumed in our analysis is derived from (Bothmann, Dandl, and Schomaker 2023), where Gender may affect the model prediction both directly and indirectly through variables such as Saving accounts and Credit amount.

We construct a counterfactual comparison between two individuals who are identical in all observed features except for the sensitive attribute Gender. Both individuals are 34 years old, have little savings in their accounts, and are applying for a credit amount of 1569 monetary units. The only difference between the two is their gender.

Our trained model produces different risk predictions for these two individuals, suggesting a potential gender bias. To understand how this disparity arises, we analyze the contribution of Gender to the final prediction using a variety of Shapley-based attribution methods. Specifically, we decompose the overall contribution of Gender into two distinct pathways: (1) the direct effect, representing the influence of Gender on the prediction that is not mediated by any other variables, and (2) the indirect effects, which capture the influence of Gender as it propagates through intermediate features such as Saving accounts and Credit amount.

The Table 5, 6 below summarizes the contribution of Sex to the prediction outcome. The results show that the direct effect of the feature is greater than the indirect effect. Both PE-SHAP and PW-SHAP demonstrated the small yet similar magnitude of effect.

Future Work

The proposed solution opens several promising directions for further investigation. As a next step, we plan to extend this approach to continuous, non-binary treatments. This will enable us to compute the proposed effects for all features and compare them more systematically, similar to the original Shapley visualization framework. This extension will significantly broaden the range of potential applications.

Additionally, we intend to explore how coalition aggregation can enhance this approach by fairly distributing the impact of each feature while accounting for its interactions with other features.

Discussion

We demonstrated that path-wise level explanations provide better resolution and are highly valuable for model interpretability. In this work, we proposed an alternative formulation of path-wise effects that better aligns with the original Shapley values by capturing interactions between features.

We further showed how local explanations can be aggregated to quantify population-level effects, such as the Natural Indirect Effect. Additionally, we introduced an efficient approach for estimating these effects using a doubly robust estimator.

References

Albini, E.; Long, J.; Dervovic, D.; and Magazzeni, D. 2022. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1054–1070.

Arah, O. A. 2019. Analyzing selection bias for credible causal inference: when in doubt, DAG it out. *Epidemiology*, 30(4): 517–520.

Bang, H.; and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4): 962–973.

Bothmann, L.; Dandl, S.; and Schomaker, M. 2023. Causal Fair Machine Learning via Rank-Preserving Interventional Distributions. *arXiv preprint arXiv:2307.12797*.

Carpenter, J. R.; Kenward, M. G.; and Vansteelandt, S. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3): 571–584.

Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Frye, C.; de Mijolla, D.; Begley, T.; Cowton, L.; Stanley, M.; and Feige, I. 2020. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33: 1229–1239.

Hafeman, D. M.; and Schwartz, S. 2009. Opening the Black Box: a motivation for the assessment of mediation. *International journal of epidemiology*, 38(3): 838–845.

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

Hofmann, H. 1994. UCI machine learning repository: Statlog (german credit data) data set. *Institut fur Statistik und "Okonometrie Universit" at Hamburg*.

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

Jung, Y.; Kasiviswanathan, S.; Tian, J.; Janzing, D.; Blöbaum, P.; and Bareinboim, E. 2022. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, 10476–10501. PMLR.

Lendle, S. D.; Subbaraman, M. S.; and van der Laan, M. J. 2013. Identification and efficient estimation of the natural direct effect among the untreated. *Biometrics*, 69(2): 310–317.

Pearl, J. 2010. An introduction to causal inference. *The international journal of biostatistics*, 6(2).

Robins, J. M.; and Greenland, S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2): 143–155.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rozenfeld, I. 2024. Causal Analysis of Shapley Values: Conditional vs. Marginal. *arXiv preprint arXiv:2409.06157*.

Sadeghi, Z.; Alizadehsani, R.; Cifci, M. A.; Kausar, S.; Rehman, R.; Mahanta, P.; Bora, P. K.; Almasri, A.; Alkhawaldeh, R. S.; Hussain, S.; et al. 2024. A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, 118: 109370.

Shapley, L. S. 1951. Notes on the n-person game—ii: The value of an n-person game.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Taufiq, M. F.; Blöbaum, P.; and Minorics, L. 2023. Manifold restricted interventional shapley values. In *International Conference on Artificial Intelligence and Statistics*, 5079–5106. PMLR.

Ter-Minassian, L.; Clivio, O.; Diazordaz, K.; Evans, R. J.; and Holmes, C. C. 2023. PWSHAP: a path-wise explanation model for targeted variables. In *International Conference on Machine Learning*, 34054–34089. PMLR.

VanderWeele, T.; and Vansteelandt, S. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2: 457–468.

VanderWeele, T. J. 2013. Policy-relevant proportions for direct effects. *Epidemiology*, 24(1): 175–176.

VanderWeele, T. J. 2014. A unification of mediation and interaction: a 4-way decomposition. *Epidemiology*, 25(5): 749–761.

Watson, D. 2022. Rational shapley values. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1083–1094.

Appendix

Rules of Do-Calculus

Let $G_{\overline{X}}$ denote the graph obtained by removing all edges directed into the features in X. Let $G_{\underline{X}}$ denote the graph obtained by removing all edges directed out of the features in X. Let $G_{\overline{T(W)}}$ denote the graph obtained by removing all edges directed into the nodes in T that are not ancestors of any node in W.

We will refer to the modified graph as G'.

1. **Observation can be ignored:** If $Y \perp T \mid W, X$ in $G' = G_{\overline{X}}$, then

$$P(y \mid do(x), w, t) = P(y \mid do(x), w)$$

2. **Observation can replace intervention:** If $Y \perp T \mid W, X$ in $G' = G_{\overline{X},\underline{T}}$, then

$$P(y \mid do(t), do(x), w) = P(y \mid t, do(x), w)$$

3. **Intervention can be ignored:** If $Y \perp T \mid W, X$ in $G' = G_{\overline{X},\overline{T(W)}}$, then

$$P(y \mid do(t), do(x), w) = P(y \mid do(x), w)$$

Definitions

Propensity score The propensity score (Rosenbaum and Rubin 1983) measures the probability of the treatment feature T given the observed covariates X_S, where $T \notin S$:

$$\pi(x_S) = P(T = 1 \mid X_S = x_S)$$

In practice, the propensity score can be approximated by fitting a logistic regression model with T as the target variable and X_S as predictors.

Average Treatment Effect The Average Treatment Effect is the expected difference in outcomes between two treatment conditions across the entire population: $\mathbb{E}[Y|do(T = t)] - \mathbb{E}[Y|do(T = t')]$. It captures the overall impact of a treatment, regardless of individual characteristics.

Conditional Average Treatment Effect The Conditional Average Treatment Effect is the expected difference in outcomes between two treatment conditions for a specific subpopulation, given a set of observed covariates: $\mathbb{E}[Y|do(T = t), C = c] - \mathbb{E}[Y|do(T = t'), C = c]$. It captures how the treatment effect varies across different subgroups within the population.

Controlled Direct Effect Controlled Direct Effect is expected difference in outcome while holding the mediator M at fixed value: $\mathbb{E}[Y|do(T = t, M = m)] - \mathbb{E}[Y|do(T = t', M = m)]$ ((VanderWeele and Vansteelandt 2009), (Pearl 2010)).

Natural Direct Effect Natural Indirect Effect is the causal effect that measures the effect transmitted through the mediator, while holding the treatment set and comparing the outcome when the mediator takes the value it would have under treatment t versus the value it would have under alternative treatment t':

$$NDE = \mathbb{E}[Y \mid do(T = t, M = m_{t'})] - \mathbb{E}[Y \mid do(T = t', M = m_{t'})]$$

The value $m_{t'}$ is the potential value that the mediator M would naturaly take if the exposure is set to t'. ((Lendle, Subbaraman, and van der Laan 2013))

Natural Indirect Effect Natural Indirect Effect is the causal effect that measures the portion of the treatment effect transmitted through the mediator, capturing the change in outcome due to the mediator changing from its natural value under treatment t to its natural value under treatment t', while holding the treatment fixed at t.

$$NIE = \mathbb{E}[Y \mid \mathrm{do}(T = t, M = m_t)] - \mathbb{E}[Y \mid \mathrm{do}(T = t, M = m_{t'})]$$

Two-Way Decomposition of the Average Treatment Effect The two-way decomposition of the ATE breaks down the total effect into two main components: the Natural Direct Effect effect and the Natural Indirect Effect.

Mathematically, this is expressed as:

$$\begin{aligned}
\text{ATE} &= \mathbb{E}[Y_{1,M_1} - Y_{0,M_0}] \\
&= \mathbb{E}[(Y_{1,M_1} - Y_{1,M_0}) + (Y_{1,M_0} - Y_{0,M_0})] \\
&= \underbrace{\mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]}_{\text{Natural Direct Effect (NDE)}} + \underbrace{\mathbb{E}[Y_{1,M_1} - Y_{1,M_0}]}_{\text{Natural Indirect Effect (NIE)}}.
\end{aligned}$$

Four-Way Decomposition of the Average Treatment Effect The total causal effect of an exposure X on an outcome Y can be decomposed using the four-way decomposition framework (VanderWeele 2014), which explicitly accounts for mediation and interaction effects involving a mediator M. The four-way decomposition expresses:

$$\begin{aligned}\text{ATE} &= \mathbb{E}[Y_{1,M_1} - Y_{0,M_0}] \\ &= \mathbb{E}[(Y_{1,M_1} - Y_{1,M_0}) + (Y_{1,M_0} - Y_{0,M_0})] \\ &= \underbrace{\mathbb{E}[Y_{1,M_0} - Y_{0,M_0}]}_{\text{Natural Direct Effect (NDE)}} + \underbrace{\mathbb{E}[(Y_{1,M_1} - Y_{1,M_0}) - (Y_{0,M_1} - Y_{0,M_0})]}_{\substack{\text{Reference Interaction (INT}_{\text{ref}}\text{)} \\ \text{and Mediated Interaction (INT}_{\text{med}}\text{)}}} + \underbrace{\mathbb{E}[Y_{0,M_1} - Y_{0,M_0}]}_{\text{Pure Indirect Effect (PIE)}}.\end{aligned}$$

Shapley Values Axioms

Efficiency Total payoff is fully distributed across players:

$$v(N) = \sum_{j \in N} \phi_j^f.$$

Dummy Player A player i that does not contribute to any coalition receives zero attribution:

$$v(S \cup \{i\}) = v(S) \implies \phi_i^f = 0.$$

Symmetry If two players i and j contribute equally to every coalition, they receive the same attribution:

$$\forall S \subseteq N \setminus \{i, j\}, \quad v(S \cup \{i\}) = v(S \cup \{j\}) \implies \phi_i^f = \phi_j^f.$$

Additivity For two games with value functions v and w, the attribution of the combined game $v + w$ is the sum of the attributions for each game:

$$\phi_i^{f,v+w} = \phi_i^{f,v} + \phi_i^{f,w}, \quad \forall i \in N.$$

Marginal Shapley Values

Marginal Shapley values use marginal (i.e., observational) distributions to compute the value function of a coalition S. The value function is defined as:

$$v(S) = E[f(x_S, X_{\overline{S}})] = \int f(x_S, x_{\overline{S}}) P(X_{\overline{S}} = x_{\overline{S}})\, dx_{\overline{S}}$$

where x_S are the values of sample for the subset of features S, and $X_{\overline{S}}$ are the remaining features marginalized over their distribution.

To illustrate the interpretation of marginal Shapley values, consider a linear model with two features:

$$f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2.$$

We compute the marginal Shapley values for feature X_1 with two coalition orders:

$$\begin{aligned}\phi_{1,\emptyset}^{f,\text{marginal}} &= (\beta_1 x_1 + \beta_2 \mathbb{E}[X_2]) - (\beta_1 \mathbb{E}[X_1] + \beta_2 \mathbb{E}[X_2]) \\ &= \beta_1 (x_1 - \mathbb{E}[X_1]). \\ \phi_{1,\{2\}}^{f,\text{marginal}} &= (\beta_1 x_1 + \beta_2 x_2) - (\beta_1 \mathbb{E}[X_1] + \beta_2 x_2) \\ &= \beta_1 (x_1 - \mathbb{E}[X_1]).\end{aligned}$$

The full marginal Shapley attribution for feature X_1 is the average over both permutations:

$$\begin{aligned}\phi_1^{f,\text{marginal}} &= \frac{1}{2}\left(\phi_{1,\emptyset}^{f,\text{marginal}} + \phi_{1,\{2\}}^{f,\text{marginal}}\right) \\ &= \beta_1 (x_1 - \mathbb{E}[X_1]).\end{aligned}$$

This shows that, under marginal Shapley values, the attribution of X_1 only reflects its direct effect on the model output, ignoring any indirect effects through X_2. In particular, if $\beta_1 = 0$, the attribution for X_1 is zero, even if X_1 influences X_2, which in turn affects the outcome. This illustrates a key limitation of marginal Shapley values in capturing causal effects.

Conditional Shapley values

Conditional Shapley Values compute feature attributions using a value function that conditions on the values of in-coalition features. Formally, the value function for a coalition $S \subseteq N$ is defined as:

$$v(S) = \mathbb{E}[f(x_S, X_{\overline{S}}) \mid X_S = x_S] = \int f(x_S, x_{\overline{S}})\, P(X_{\overline{S}} = x_{\overline{S}} \mid X_S = x_S)\, dx_{\overline{S}},$$

where x_S denotes the observed values of features in the coalition, and $\overline{S} = N \setminus S$ represents the remaining features.

This approach is conceptually similar to Marginal Shapley Values, but differs in how the weighting is computed. Instead of averaging over marginal distributions, Conditional Shapley Values incorporate feature dependencies by conditioning on the values in the coalition.

Consider a linear model $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$. The conditional Shapley values for feature X_1 can be derived as follows:

$$\begin{aligned}
\phi_{1,\emptyset}^{f,\text{cond}} &= (\beta_1 x_1 + \beta_2 \cdot \mathbb{E}[X_2 \mid X_1 = x_1]) - (\beta_1 \cdot \mathbb{E}[X_1] + \beta_2 \cdot \mathbb{E}[X_2]) \\
&= \beta_1 (x_1 - \mathbb{E}[X_1]) + \beta_2 \left(\mathbb{E}[X_2 \mid X_1 = x_1] - \mathbb{E}[X_2]\right), \\
\phi_{1,\{2\}}^{f,\text{cond}} &= (\beta_1 x_1 + \beta_2 x_2) - (\beta_1 \cdot \mathbb{E}[X_1 \mid X_2 = x_2] + \beta_2 x_2) \\
&= \beta_1 \left(x_1 - \mathbb{E}[X_1 \mid X_2 = x_2]\right).
\end{aligned}$$

The final attribution for feature X_1 is the average over these two permutations:

$$\begin{aligned}
\phi_1^{f,\text{cond}} &= \frac{1}{2}\left(\phi_{1,\emptyset}^{f,\text{cond}} + \phi_{1,\{2\}}^{f,\text{cond}}\right) \\
&= \frac{1}{2}\Big[\beta_1(x_1 - \mathbb{E}[X_1]) + \beta_1(x_1 - \mathbb{E}[X_1 \mid X_2 = x_2]) \\
&\qquad + \beta_2(\mathbb{E}[X_2 \mid X_1 = x_1] - \mathbb{E}[X_2])\Big].
\end{aligned}$$

Interestingly, even if $\beta_1 = 0$, the attribution for feature X_1 may still be non-zero due to its influence on X_2, captured via the conditional expectation. This reflects the indirect effect of X_1 on the model output through feature dependencies.

It is important to note that, since Conditional Shapley Values rely on conditioning by observation, they capture correlations rather than causal relationships. As a result, they cannot distinguish between correlation and causation.

Asymmetric Shapley Values

Asymmetric Shapley Values (Frye, Rowat, and Feige 2020) relax the Symmetry Axiom of classical Shapley Values to incorporate prior causal knowledge into feature attributions.

When features are causally related (e.g., `age` causes `education level`), treating them symmetrically can obscure important distinctions. For instance, even though both `age` and `education` may provide information about income, we may prefer to attribute more importance to `age` since it causally precedes `education`, and only consider `education` after accounting for `age`.

ASVs define a new permutation weighting function $w(\pi)$ that depends on the causal DAG. Two common weighting strategies are:

- Distal cause weighting $w_{\text{distal}}(\pi)$: Selects only those permutations where all causal ancestors of a feature already appear in coalition.

  Let M be the number of permutations consistent with the causal graph. Then:

  $$w_{\text{distal}}(\pi) = \frac{1}{M} \begin{cases} 1 & \text{if } i \prec_\pi j \text{ for every known ancestor } i \text{ of } j, \\ 0 & \text{otherwise} \end{cases}$$

  That is, for any descendant $j \notin S$, its ancestor i must appear in the coalition $i \in S$ before j is added.

- Proximate cause weighting $w_{\text{proximate}}(\pi)$: Selects only those permutations where features appear before their causal ancestors, anti-causal orderings

  $$w_{\text{proximate}}(\pi) = \frac{1}{M} \begin{cases} 1 & \text{if } i \succ_\pi j \text{ for every known ancestor } i \text{ of } j, \\ 0 & \text{otherwise} \end{cases}$$

In both cases, if no causal information is provided, these weights reduce to the uniform distribution $w(\pi) = \frac{1}{|N|!}$, recovering the classical (symmetric) Shapley Values.

ASVs use the same conditional value function as Conditional Shapley Values:

$$v(S) = \mathbb{E}[f(x_S, X_{\overline{S}}) \mid X_S = x_S]$$

This captures both direct and indirect effects, while the permutation weights control which effects are emphasized.

Similarly to consider the case where feature X_2 causally precedes X_1, i.e., $X_2 \to X_1$. Under a distal weighting, the only valid permutation is $\pi = (2, 1)$, reflecting the causal ordering.

The attribution for feature X_1 under the Asymmetric Conditional Shapley Value is then:

$$\begin{aligned}\phi_1^{f,\text{asym-cond}} &= \phi_{1,\{2\}}^{f,\text{cond}} \\ &= \beta_1 \cdot (x_1 - \mathbb{E}[X_1 | X_2 = x_2])\end{aligned}$$

This attribution does not rely on conditioning that runs against the causal direction.

Causal Shapley Values

Causal Shapley Values (Heskes et al. 2020) incorporate causal knowledge directly into the computation of feature attributions by using interventional (do-calculus) probabilities. Unlike Asymmetric Shapley Values, which inject causal structure into the ordering of permutations, Causal Shapley Values integrate causal models directly into the value function. These two perspectives represent seperate directions of causal enhancement.

The value function in Causal Shapley Values is defined using interventional distributions:

$$\begin{aligned}v(S) &= \mathbb{E}[f(x_S, X_{\overline{S}}) \mid do(X_S = x_S)] \\ &= \int f(x_S, x_{\overline{S}})\, P(X_{\overline{S}} = x_{\overline{S}} \mid do(X_S = x_S))\, dx_{\overline{S}}\end{aligned}$$

Computing Interventional Distributions When the full causal DAG is known, the interventional distribution $P(X_{\overline{S}} \mid do(X_S = x_S))$ can be computed via the truncated factorization formula (see):

$$P(X_{\overline{S}} \mid do(X_S = x_S)) = \prod_{j=1}^{n} P(X_j \mid X_{\text{pa}(j)\cap\overline{S}}, x_{\text{pa}(j)\cap S}) \tag{1}$$

In practice, however, we often lack access to a complete causal graph. To support this, Causal Shapley Values support partial causal ordering. Specifically, the feature set is partitioned into components $\tau = \{X_i, X_j, \dots\}$, with a topological ordering over components such as (τ_1, τ_2, τ_3), while intra-component relationships remain unspecified.

For each component τ, we specify whether dependencies within the component are due to confounding or mutual interaction. Given this, the interventional distribution can be computed as ():

$$\begin{aligned}P(X_{\overline{S}} \mid do(X_S = x_S)) = &\prod_{\tau \in \mathcal{T}_{\text{confounding}}} P(X_{\tau\cap\overline{S}} \mid X_{\text{pa}(\tau)\cap\overline{S}}, x_{\text{pa}(\tau)\cap S}) \times \\ &\prod_{\tau \in \mathcal{T}_{\overline{\text{confounding}}}} P(X_{\tau\cap\overline{S}} \mid X_{\text{pa}(\tau)\cap\overline{S}}, x_{\text{pa}(\tau)\cap S}, x_{\tau\cap S})\end{aligned} \tag{2}$$

Direct and Indirect Effect Decomposition A major advantage of Causal Shapley Values is their ability to decompose attributions into direct and indirect effects, where their sum yields the total effect:

$$\begin{aligned}\phi_{i,S}^{f,\text{total causal}} &= \phi_{i,S}^{f,\text{direct causal}} + \phi_{i,S}^{f,\text{indirect causal}} \\ \phi_{i,S}^{f,\text{direct causal}} &= \mathbb{E}[f(X_{\overline{S}}, x_{S\cup\{i\}}) \mid do(X_S = x_S)] - \mathbb{E}[f(X_{\overline{S}}, x_S) \mid do(X_S = x_S)] \\ \phi_{i,S}^{f,\text{indirect causal}} &= \mathbb{E}[f(X_{\overline{S}\setminus\{i\}}, x_{S\cup\{i\}}) \mid do(X_{S\cup\{i\}} = x_{S\cup\{i\}})] \\ &\quad - \mathbb{E}[f(X_{\overline{S}\setminus\{i\}}, x_{S\cup\{i\}}) \mid do(X_S = x_S)]\end{aligned}$$

Direct effect: Measures the change in output when feature X_i is set to its sample value, without allowing it to influence other features (i.e., controlling for its downstream effects).

Indirect effect: Captures the influence of X_i on the outcome via its effect on other features, by comparing distributions with and without intervening on X_i.

Total effect: Sum of direct and indirect effects.

Example: Linear Model For a linear model $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$, we get a similar decomposition to that of Conditional Shapley Values, with the key difference being that conditioning is now interventional:

$$\begin{aligned}\phi_1^{f,\text{total causal}} = \phi_1^{f,\text{direct causal}} + \phi_1^{f,\text{indirect causal}} \propto \\ \underbrace{\beta_1 \cdot (x_1 - \mathbb{E}[X_1]) + \beta_1 \cdot (x_1 - \mathbb{E}[X_1 \mid do(X_2 = x_2)])}_{\text{Direct effect}} + \\ \underbrace{\beta_2 \cdot (\mathbb{E}[X_2 \mid do(X_1 = x_1)] - \mathbb{E}[X_2])}_{\text{Indirect effect}}\end{aligned}$$

Here, the direct effect captures the isolated influence of x_1, while the indirect effect reflects how x_1 causally impacts x_2, and thus the final prediction.

Asymmetric/Symmetric Causal Shapley values

Let G - be causal DAG, f - any function, $G^\star$ subtree of G with root in feature X_i spanned by all features it causally impacts. Let X_j be feature which mediates impact of feature X_i - $X_i \prec_{G^\star} X_j$. Let if $X_j' \prec_{G^\star} X_j$, then $Y \perp X_j | X_{j'}$, meaning that knowing information about X_j will not give any additional information for prediction knowing already $X_{j'}$, alternatively X_j just passes impact of parents, not adding anything additionally.

When computing Symmetric Causal Shapley values for feature X_j we take into account any permutation π. Let us denote coalition S', where feature X_i is not included; that is, $X_i \notin S'$.

$$\begin{aligned}\phi_{X_j,S'}^{\text{indirect causal}} = \mathbb{E}\left[f\left(X_{\overline{S'}\setminus\{X_j\}}, x_{S'\cup\{X_j\}}\right) \,\middle|\, do\left(X_{S'\cup\{X_j\}} = x_{S'\cup\{X_j\}}\right)\right] \\ - \mathbb{E}\left[f\left(X_{\overline{S'}\setminus\{X_j\}}, x_{S'\cup\{X_j\}}\right) \,\middle|\, do\left(X_{S'} = x_{S'}\right)\right] \neq 0\end{aligned}$$

Since feature X_j is influenced by X_i, impact of X_i will be given to X_j. This way spreading its impact for all features in $G^\star$.

In contrast if we use Asymmetric Shapley value for feature X_j we would not have such permutation π that $X_i \notin S'$. Let $S^\star = X_{j'} \prec_{G^\star} X_j$, then indirect effect of feature X_j:

$$\begin{aligned}\phi_{X_j,S}^{\text{asym. indirect causal}} &= \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_{S\cup\{X_j\}})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_S)\right] \\ &\overset{(1)}{=} \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_{(S\setminus S^\star)\cup\{X_j, pa(X_j)\}})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_{(S\setminus S^\star)\cup pa(X_j)})\right] \\ &\overset{(2)}{=} \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_{(S\setminus S^\star)\cup pa(X_j)})\right] \\ &\quad - \mathbb{E}\left[f(X_{\overline{S}\setminus\{X_j\}}, x_{S\cup\{X_j\}}) \,\middle|\, do(x_{(S\setminus S^\star)\cup pa(X_j)})\right] \\ &= 0\end{aligned}$$

(1) - Using 3 rule of do calculus we can ignore conditioning by intervention on $X_{j'} \prec_{G^\star} X_j$, because we already condition upon X_j. Intuitively we would not get any additional information about features further down in causal graph G from feature that are above X_j as we already know impact of X_j.

(2) - As we stated $Y \perp X_j | pa(X_j)$, thus we can ignore conditioning on X_j.

Manifold Restricted Interventional Shapley Values

Taufiq, Blöbaum, and Minorics (2023) propose Manifold Shapley Values (ManifoldShap) under the assumption that a model's behavior should primarily be characterized on the data manifold. To satisfy this, they define a value function restricted to a local neighborhood $\mathcal{Z} \subseteq \mathcal{X}$, where $\mathbf{x} \in \mathcal{Z}$, as:

$$v_{\mathcal{Z}}(S) = \mathbb{E}\left[f(x_S, X_{\overline{S}}) \mid do(X_S = x_S), \mathbf{x} \in \mathcal{Z}\right]$$

In practice, it is sufficient to estimate an indicator function $\hat{g} \approx \mathbf{1}(\mathbf{x} \in \mathcal{Z})$, which identifies whether a sample lies on the manifold. The value function becomes:

$$v_{\mathcal{Z}}(S) = \frac{\mathbb{E}\left[f(x_S, X_{\overline{S}}) \cdot \hat{g}(X) \mid do(X_S = x_S)\right]}{\mathbb{E}\left[\hat{g}(X) \mid do(X_S = x_S)\right]}$$

This formulation ensures that Shapley attributions are computed only using samples that lie close to the data manifold, avoiding unrealistic or out-of-distribution combinations that can arise in standard interventional approaches.

Path-Wise Shapley values

Path-Wise Shapley values (Ter-Minassian et al. 2023) is designed to explain the local effect of a binary variable T (e.g., treatment) on an outcome Y through a directed acyclic graph representing causal dependencies.

PWSHAP extends traditional on-manifold (Conditional) Shapley values by breaking them down into causally valid contributions. PWSHAP decomposes Coalition-Wise shapley value in Propensity weight and Coalition-wise Shapley Effect

Decomposing Coalition-Wise Shapley Values into Shapley Effects Let i be a treatment feature (i.e., $X_i = T$) and Y the model prediction. The Shapley value associated with the coalition S can be decomposed into two components: a propensity weight and a Coalition-Wise Shapley effect (proof):

$$\begin{aligned}\phi_{i,S}(v) &= v(S \cup \{i\}) - v(S) \\ &= \underbrace{\left(t - P(T = 1 \mid X_S = x_S)\right)}_{\text{Propensity weight } w(x_S,t)} \cdot \underbrace{\left(v(S \cup \{i\})\big|_{T=1} - v(S \cup \{i\})\big|_{T=0}\right)}_{\text{Coalition-Wise Shapley effect } \psi_{T \to Y \mid X_S}(x_S)}\end{aligned}$$

Propensity Weight. The term

$$w(x_S, t) = t - \pi(x_S) = t - P(T = 1 \mid X_S = x_S)$$

is the difference of feature value t and propensity score . Intuitively it can be understood as a measure of how much feature value t is as an outlier.

Coalition-Wise Shapley Effect. Given a treatment feature T and a subset of covariates X_S, the Coalition-Wise Shapley effect isolates the expected change in output from setting $T = 1$ vs. $T = 0$, while conditioning on the coalition X_S:

$$\psi_{T \to Y \mid X_S}(x_S) = v(S \cup \{i\})\big|_{T=1} - v(S \cup \{i\})\big|_{T=0}$$

This can be seen as a generalization of the Conditional Average Treatment Effect . However, unlike CATE, X_S is not necessarily a valid adjustment set.

Path-Wise Shapley Effect Coalition-Wise Shapley effect can be understood as flow from T to Y through set of covariates X_S. To isolate the effect of the treatment through X_j path we subtract from the causal flow through all covariates, the flow through all covariates except for X_j.

Path-wise effect of T on Y through X_j for model f can be determined as difference of two Coalition-wise Shapley effects:

$$\psi^f_{C_i}(c) = \psi^f_{T \to Y \mid C_{S*}}(c_{S*}) - \psi^f_{T \to Y \mid C_{S* \setminus \{i\}}}(c_{S* \setminus \{i\}})$$

Depending on the causal role of the feature in the graph Path-wise Shapley values can be interpreted differently.

Tracing Different Paths This experiment evaluates the ability of Path-Wise Shapley values to attribute model output to specific causal pathways, and compares them to standard Causal Shapley values. To do so, we generate synthetic datasets under four structural settings: 1) neither C_1 nor C_2 are mediators, 2) only C_1 is a mediator, 3) only C_2 is a mediator, and 4) both C_1 and C_2 are mediators.

We define the mediating variables as follows:

- $C_1^{\text{med}} \sim \text{Bern.}((1 - T) \cdot 0.3 + 0.5)$
- $C_2^{\text{med}} \sim \text{Bern.}((1 - T) \cdot 0.6 + 0.2)$

These expressions indicate that the treatment T affects the distribution of C_1 and/or C_2, thus creating a mediated path from T to Y. Non-mediated versions are sampled independently: $C_1 \sim \text{Bern.}(0.5)$, $C_2 \sim \text{Bern.}(0.2)$. In all settings, the outcome is defined as $Y = C_1 + C_2$. The different data distributions are summarized below.

$$\begin{cases} T \sim \text{Bern.}(0.5) \\ C_1 \sim \text{Bern.}(0.5) \\ C_2 \sim \text{Bern.}(0.2) \\ Y = C_1 + C_2 \end{cases} \quad \begin{cases} T \sim \text{Bern.}(0.5) \\ C_1^{\text{med}} \\ C_2 \sim \text{Bern.}(0.2) \\ Y = C_1 + C_2 \end{cases} \quad \begin{cases} T \sim \text{Bern.}(0.5) \\ C_1 \sim \text{Bern.}(0.5) \\ C_2^{\text{med}} \\ Y = C_1 + C_2 \end{cases} \quad \begin{cases} T \sim \text{Bern.}(0.5) \\ C_1^{\text{med}} \\ C_2^{\text{med}} \\ Y = C_1 + C_2 \end{cases}$$

Mediators: $\emptyset$ **Mediators:** $\{C_1\}$ **Mediators:** $\{C_2\}$ **Mediators:** $\{C_1, C_2\}$

The results in Table 7 show that Path-Wise Shapley values can successfully isolate and quantify the contribution of individual mediation paths.

Note that Path-Wise Shapley values are not guaranteed to sum to the model prediction and may operate on a different scale. We use absolute values to mitigate the impact of sign instability in attributions.

Table 7: Path-Wise Shapley vs. Causal Shapley effects on synthetic mediation scenarios.

Structure	$\lvert\psi_{T\to C_1\to Y}\rvert$	$\lvert\psi_{T\to C_2\to Y}\rvert$	$\lvert\phi_{\text{indirect}}\rvert$
C_1, C_2	0	0	0
C_1-M, C_2	0.29	0	0.06
C_1, C_2-M	0	0.51	0.14
C_1-M, C_2-M	0.16	0.32	0.23

Conditioning by Intervention

Proof of conditioning on interventions using observational conditioning:

$$\begin{aligned} P(X_{\overline{S}} = x_{\overline{S}} | do(x_S)) &= \prod_{j\in\overline{S}} P(X_j = x_j | X_{(j' \prec_G j)\cap\overline{S}}, do(x_S)) = \\ &\overset{(1)}{=} \prod_{j\in\overline{S}} P(X_j = x_j | X_{pa(j)\cap\overline{S}}, do(x_S)) = \\ &\overset{(2)}{=} \prod_{j\in\overline{S}} P(X_j = x_j | X_{pa(j)\cap\overline{S}}, do(x_{pa(j)\cap S})) = \\ &\overset{(3)}{=} \prod_{j\in\overline{S}} P(X_j = x_j | X_{pa(j)\cap\overline{S}}, x_{pa(j)\cap S}) \end{aligned} \tag{3}$$

(1) - Using rule 1 of do-calculus, we can remove conditioning by observation for all nodes that causally precedes parents of X_j.

(2) - Using rule 3 of do-calculus, we can ignore conditioning by interventions upon nodes that are further in the causal graph and causally precedes parents of feature X_j.

(3) - Using rule 2 of do-calculus, we can change conditioning by intervention on conditioning by observation upon variables that are higher up in causal structure.

Proof of conditioning on interventions using observational conditioning for partial causal ordering:

$$\begin{aligned} P(X_{\overline{S}} = x_{\overline{S}} | do(X_S = x_S)) &= P(X | do(X_S = x_S)) = \\ &= \prod_{\tau\in T} P(X_\tau | X_{\tau' \prec_G} X_\tau, do(X_S = x_S)) \\ &\overset{(1)}{=} \prod_{\tau\in T} P(X_\tau | X_{pa(\tau)}, do(X_S = x_S)) \\ &= \prod_{\tau\in T} P(X_{\tau\cap\overline{S}} | X_{pa(\tau)\cap\overline{S}}, do(X_S = x_S)) \\ &\overset{(2)}{=} \prod_{\tau\in T} P(X_{\tau\cap\overline{S}} | X_{pa(\tau)\cap\overline{S}}, do(x_{(\tau' \prec_G \tau)\cap S}), do(x_{\tau\cap S})) \\ &\overset{(3)}{=} \prod_{\tau\in T} P(X_{\tau\cap\overline{S}} | X_{pa(\tau)\cap\overline{S}}, do(x_{pa(\tau)\cap S}), do(x_{\tau\cap S})) \\ &\overset{(4)}{=} \prod_{\tau\in T} P(X_{\tau\cap\overline{S}} | X_{pa(\tau)\cap\overline{S}}, x_{pa(\tau)\cap S}, do(x_{\tau\cap S})) \end{aligned} \tag{4}$$

(1) - Using rule 1 of do-calculus, we can remove conditioning by observation, as $X_\tau \perp X_{\tau' \prec_G (pa(\tau))} | , X_{pa(\tau)}, X_S$ in graph $G_{\overline{X_S}}$. Intuitively, knowing all relevant information about the parents of a variable, makes additional information about parents' parents redundant to determine impact on the variable.

(2) - X_S can be decomposed as $X_S = \{X_{(\tau' \prec_G \tau)\cap S}, X_{(\tau' \succ_G \tau)\cap S}, X_{\tau\cap S}\}$. Using rule 3 of do-calculus, we can ignore conditioning by intervention for all components further down in causal chain, $X_{\tau\cap\overline{S}} \perp X_{(\tau' \succ_G \tau)\cap S} | X_{pa(\tau)\cap\overline{S}}, X_{(\tau' \preceq_G \tau)\cap S}$ in updated graph G'.

(3) - Using 3 rule again, we can ignore conditioning by interventions for all components before $X_{pa(\tau)}$, as $X_{\tau\cap\overline{S}} \perp X_{\tau' \prec_G pa(\tau)} | X_{pa(\tau)}, X_{\tau\cap S}$ in updated graph G'.

(4) - Using rule 2, we can change conditioning on intervention on conditioning by observations.
Using 3 rule of do-calculus, for component with dependencies induced by common confounder:

$$P(X_{\tau\cap\overline{S}}|X_{pa(\tau)\cap\overline{S}}, x_{pa(\tau)\cap S}, do(x_{\tau\cap S})) = \\ = P(X_{\tau\cap\overline{S}}|X_{pa(\tau)\cap\overline{S}}, x_{pa(\tau)\cap S})$$

Again using 2 rule of do-calculus, for component with dependencies induced by mutual interactions:

$$P(X_{\tau\cap\overline{S}}|X_{pa(\tau)\cap\overline{S}}, x_{pa(\tau)\cap S}, do(x_{\tau\cap S})) = \\ = P(X_{\tau\cap\overline{S}}|X_{pa(\tau)\cap\overline{S}}, x_{pa(\tau)\cap S}, x_{\tau\cap S})$$

Path-wise Shapley Value Decomposition

Proof of Property

Let X_i be a treatment feature and $x_i \in \{0, 1\}$ its observed value. We start by expressing the value function using conditional expectations:

$$\begin{aligned} v(S) &= \mathbb{E}[f(x_S, X_i, X_{\overline{S}\setminus\{i\}}) \mid X_S = x_S] \\ &= P(X_i = x_i \mid X_S = x_S) \cdot \mathbb{E}[f(x_S, x_i, X_{\overline{S}\setminus\{i\}}) \mid X_i = x_i, X_S = x_S] \\ &\quad + P(X_i = 1 - x_i \mid X_S = x_S) \cdot \mathbb{E}[f(x_S, 1 - x_i, X_{\overline{S}\setminus\{i\}}) \mid X_i = 1 - x_i, X_S = x_S] \\ &= P(X_i = x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = x_i} \\ &\quad + P(X_i = 1 - x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = 1 - x_i} \end{aligned}$$

We now compute the marginal contribution of X_i to the coalition S:

$$\begin{aligned} v(S \cup \{i\}) - v(S) &= P(X_i = x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = x_i} \\ &\quad + P(X_i = 1 - x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = x_i} \\ &\quad - P(X_i = x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = x_i} \\ &\quad - P(X_i = 1 - x_i \mid X_S = x_S) \cdot v(S \cup \{i\})\big|_{X_i = 1 - x_i} \\ &= P(X_i = 1 - x_i \mid X_S = x_S) \cdot \left[v(S \cup \{i\})\big|_{X_i = x_i} - v(S \cup \{i\})\big|_{X_i = 1 - x_i} \right] \end{aligned}$$

Proof of Property C-CATE as Controlled Direct Effect

Let T be the treatment, Y the outcome and X_S a set of mediators. The Controlled Direct Effect of T on Y, with mediators X_S fixed to value x_S, is defined as:

$$CDE_{X_S}(t, t', x_S) = \mathbb{E}[Y|do(T = t, X_S = x_S)] - E[Y|do(T = t', X_S = x_S)]$$

Then:

$$CDE_{X_S}(t, t', x_S) = \mathbb{E}[Y|do(T = t, X_S = x_S)] - \mathbb{E}[Y|do(T = t', X_S = x_S)] \overset{(1)}{=} \\ \mathbb{E}[Y|do(T = t), X_S = x_S] - \mathbb{E}[Y|do(T = t'), X_S = x_S] = \Delta_t(x_S)$$

(1) - According to Rule 2 of do-calculus, an intervention can be replace by observation if $Y \perp X_S|T$ in $G' = G_{\overline{T},\underline{X_S}}$. This condition hold if all of the variables in X_S are not a descendant of any other variable besides T and any other variable included in X_S.

Proof of Property PE-SHAP Effect Aggregation for ATE

Let T be the treatment, Y the outcome and X_S a minimal adjustment set. From definition of Average Treatment Effect:

$$\begin{aligned}
&\int \Delta_t(x_S) \cdot P(X_S = x_S \mid \mathrm{do}(T = t'))\, dx_S \\
&\overset{(1)}{=} \int \Delta_t(x_S) \cdot P(X_S = x_S)\, dx_S \\
&= \mathbb{E}_{X_S}[\Delta_t(x_S)] \\
&= \mathbb{E}_{X_S}\left[\mathbb{E}[Y \mid \mathrm{do}(T = t), X_S = x_S] - \mathbb{E}[Y \mid \mathrm{do}(T = t'), X_S = x_S]\right] \\
&= \mathbb{E}[Y \mid \mathrm{do}(T = t)] - \mathbb{E}[Y \mid \mathrm{do}(T = t')] \\
&= \text{ATE}
\end{aligned}$$

(1) - By the third rule of do-calculus, since X_S is minimal adjustment set that causally precedes T, the distribution of X_S is unaffected by the intervention $do(T = t')$.

Proof of Property PE-SHAP Effect Aggregation for NDE

Let T be the treatment, Y the outcome and X_S - subset of mediators, not a descendant of any other variables except for T or any other variable included in X_S. Then:

$$\begin{aligned}
&\int \Delta_t(x_S) \cdot P(X_S = x_S \mid \mathrm{do}(T = t'))\, dx_S \\
&\overset{(1)}{=} \int CDE(t, t', x_S) \cdot P(X_S = x_S \mid \mathrm{do}(T = t'))\, dx_S \\
&= \mathbb{E}[Y \mid \mathrm{do}(T = t, X_S = x_{t'})] - \mathbb{E}[Y \mid \mathrm{do}(T = t', X_S = x_{t'})] \\
&= \text{NDE}
\end{aligned}$$

(1) - This equality follows from Property . Since X_S is not a descendant of any other variables except for T or any other variable included in X_S, $CDE(t, t', x_S)$ equals to $\Delta_t(x_S)$.

Proof of Property PE-SHAP Effect Aggregation for NIE

Let T be the treatment, Y the outcome, and X_S a subset of mediators such that no element of X_S is a descendant of any variable other than T or variables already included in X_S. Then:

$$\begin{aligned}
&\int P(X_S = x_S \mid \mathrm{do}(T = t')) \cdot \lambda_{T(t) \to X_S(x_S) \to Y}\, dx_S \\
&= \int P(X_S = x_S \mid \mathrm{do}(T = t')) \cdot (ATE - \Delta_t(x_S))\; dx_S \\
&= ATE - \int \Delta_t(x_S) \cdot P(X_S = x_S \mid \mathrm{do}(T = t'))\, dx_S \\
&\overset{(1)}{=} ATE - NDE \\
&\overset{(2)}{=} NIE
\end{aligned}$$

(1) - Follows from Property about NDE aggregation, since X_S is a subset of mediators that are not descendants of any variable except for T or variables within X_S.

(2) - Follows from the two-way decomposition of the Average Treatment Effect .

Experiments: Population Effects

One important application of the proposed approach is its ability to aggregate local explanations into meaningful population-level causal effect estimates, such as the Natural Direct Effect (NDE) and Natural Indirect Effect (NIE). Our method, , leverages this property to provide interpretable decompositions of the total effect into these causal components. Table 8 compares the estimated NDE and NIE across different methods. The S-learner-MLP, a common baseline, produces the lowest NDE estimate, yielding slightly better results with smaller Monte Carlo error compared to the aggregated estimates. However, despite this slight advantage, performs very well, producing robust and consistent effect estimates that closely align with the baseline. This highlights 's capability to provide reliable causal effect decompositions while also offering enhanced interpretability. Notably,

Method	NDE Estimate (SE)	NIE Estimate (SE)
S-learner-MLP	**0.0503 (0.002)**	**0.0284 (0.004)**
-GT	0.0546 (0.007)	0.0407 (0.006)
-MLP	0.0525 (0.007)	0.0404 (0.006)
-XGBoost	0.0547 (0.007)	0.0407 (0.006)
-Linear	0.1565 (0.003)	0.1340 (0.003)

Table 8: Comparison of Estimated Natural Direct Effect and Natural Indirect Effect on $\hat{Y}$ Across Methods. Estimates are reported with Mean Absolute Error and Monte Carlo Error shown in parentheses.

the -Linear method produces substantially larger estimates for both NDE and NIE, which is expected given the differences in model structure and data representation, particularly due to the inclusion of squared terms.

Overall, while the classical S-learner achieves the smallest effect magnitudes in this setting, demonstrates comparable performance.

Theoretical Details of Shapley Values Comparison Example

Consider a black-box predictive model $\hat{f}$ defined as follows:

$$\hat{Y} = \hat{f}(T, C, M_1, M_2) = C + \beta_T T + \beta_{M_1} M_1 + \beta_{TM_1}(T \times M_1) + \beta_{M_2} M_2 + \varepsilon,$$

where T represents the treatment variable, C is a covariate, and M_1, M_2 are mediators influenced by the treatment through the relationships:

$$M_1 = \gamma_1 T + \eta_1, \quad M_2 = \gamma_2 T + \eta_2,$$

with $\varepsilon, \eta_1, \eta_2$ representing error terms.

We define the total average treatment effect and controlled direct effect on the outcome $\hat{Y}$ as:

$$\begin{aligned}
\Delta_T(\emptyset) = \text{ATE} &= \beta_T + (\beta_{M_1} + \beta_{TM_1})\gamma_1 + \beta_{M_2}\gamma_2, \\
\Delta_T(M_1 = m_1) = \text{CDE}(M_1 = m_1) &= \beta_T + \beta_{TM_1} m_1 + \beta_{M_2}\gamma_2 \\
\Delta_T(M_1 = m_1, M_2 = m_2) = \text{CDE}(M_1 = m_1, M_2 = m_2) &= \beta_T + \beta_{TM_1} m_1, \\
\Delta_T(M_2 = m_2) = \text{CDE}(M_2 = m_2) &= \beta_T + \beta_{M_1}\gamma_1 + \beta_{TM_1}\gamma_1
\end{aligned}$$

By subtracting the controlled direct effect from the total effect, we isolate the path-specific effect of the treatment T mediated through M_1:

$$\begin{aligned}
\lambda_{T \to M_1(m_1) \to \hat{Y}} &= \Delta_T(\emptyset) - \Delta_T(M_1 = m_1) = (\beta_{M_1} + \beta_{TM_1})\gamma_1 - \beta_{TM_1} m_1 \\
\Psi_{T \to M_1 \to \hat{Y}} &= \Delta_T(M_1 = m_1, M_2 = m_2) - \Delta_T(M_2 = m_2) = -\beta_{M_1}\gamma_1
\end{aligned}$$

Importantly, when the interaction term $\beta_{TM_1} = 0$, this path-specific effect simplifies to $\beta_{M_1}\gamma_1$, which represents the direct mediation effect of M_1 without any treatment-mediator interaction.

Experiments: Estimators comparison

Estimators	Both Correct	Incorrect Outcome	Incorrect Treatment	Incorrect Both
Doubly Robust	0.0734(0.0110)	0.0736(0.0109)	0.0734(0.0110)	**0.1146(0.0126)**
Regression	0.0679(0.0111)	0.2844 (0.0254)	**0.0679(0.0111)**	0.2844 (0.0254)
IPW	**0.0657(0.0111)**	**0.0657(0.0111)**	0.1838 (0.0174)	0.1838 (0.0174)

Table 9: Mean Absolute Error with Monte Carlo Error (in parentheses) across different model specification scenarios.

To assess the empirical performance and robustness of the proposed CATE estimators, we evaluate them under four distinct model specification scenarios that reflect varying degrees of misspecification in both the outcome and treatment models:

1. **Both outcome and treatment models are correctly specified.**
2. **Outcome model is misspecified** - we assume a linear functional form, whereas the true data-generating process includes non-linear dependencies.
3. **Treatment model is misspecified** - due to omission of a confounding variable.
4. **Both models are misspecified.**

Table 9 presents the Mean Absolute Error (MAE) along with the Monte Carlo standard error (in parentheses) across all four scenarios. Lower MAE values indicate better estimator accuracy in recovering the true CATE.

The results empirically confirm the theoretical robustness of the Doubly Robust (DR) estimator: it achieves stable performance when either the outcome model or the treatment model is misspecified. However, its performance degrades when both models are misspecified - though it still generally outperforms alternative methods in such settings. Regression-based estimation performs exceptionally well when the outcome model is correctly specified but degrades significantly under outcome model misspecification. In contrast, IPW is highly sensitive to misspecification of the treatment model, as expected, given its reliance on accurate estimation of the propensity scores. It is worth noting that the DR estimator can underperform relative to IPW or regression-based methods when there is high confidence in the correct specification of either the treatment or outcome model. This is primarily due to the higher variance typically associated with DR estimators.