A Causal Perspective on Shapley Values

Gabriel Bucur

iCIS | Data Science
Radboud University

The role of causality in XAI

- Humans have a strong tendency to reason about their environment in causal terms.[1] Both causality and XAI are centered on humans, aiming to ensure true usefulness for humans.[2]
- It is often easier for a model to get good predictions for the wrong reasons.[3] Pearl highlights the need to have AI systems that are robust to changes in environment.[4]
- For medical decision support, it is necessary to understand the causality of learned representations, so causal reasoning becomes an important component of explainable AI.[5]

---

1 Sloman, *Causal Models*.

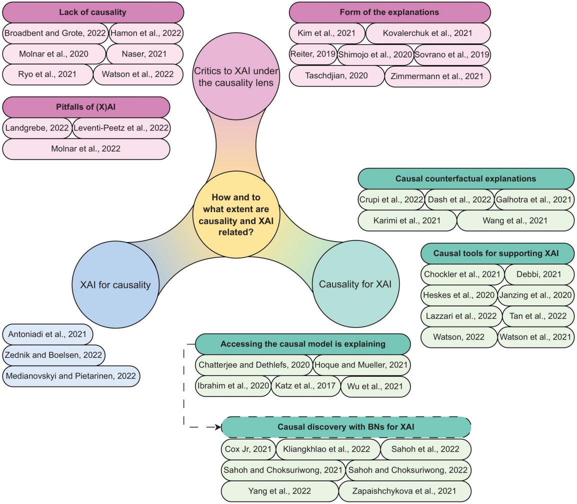
2 Carloni, Berti, and Colantonio, "The Role of Causality in Explainable Artificial Intelligence".

3 Pichler and Hartig, "Machine learning and deep learning: A review for ecologists".

4 Pearl, "The seven tools of causal inference, with reflections on machine learning".

5 Wu et al., "Methods and Applications of Causal Reasoning in Medical Field".

The role of causality in XAI



Source: Carloni, Berti, and Colantonio, "The Role of Causality in Explainable Artificial Intelligence"

Shapley values

Shapley values are based on solid game-theoretic principles and provide a natural way to estimate the contribution of each input feature in a predictive model.

The prediction Y of a model $f(\mathbf{X})$ can be decomposed into:

$$Y = f(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})] + \sum_{i=1}^{n} \phi_i[f(\mathbf{x})]\,, i \in N = \{1, 2, \ldots, n\},$$

where the *Shapley value* of feature i is

$$\phi_i = \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} \left[\upsilon(S \cup \{i\}) - \upsilon(S) \right],$$

for a chosen *payoff* / *value function* υ and *coalition* S.

Value function[a]

[a] Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions"; Aas, Jullum, and Løland, "Explaining individual predictions when features are dependent".

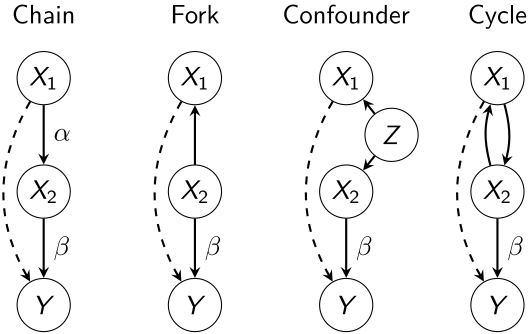
A common choice for the value function involves computing conditional distributions on the observed data:

$$\begin{aligned} v(S) &= \mathbb{E}\left[f(\mathbf{X})|\mathbf{X}_S = \mathbf{x}_S\right] \\ &= \int d\mathbf{X}\, P(\mathbf{X}|\mathbf{X}_S = \mathbf{x}_S) f(\mathbf{X}) \\ &= \int d\mathbf{X}_{\bar{S}}\, P(\mathbf{X}_{\bar{S}}|\mathbf{X}_S = \mathbf{x}_S) f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S)\,, \end{aligned}$$

where S is a coalition of players and $\bar{S} = N \setminus S$ is the set of players outside the coalition.

Why do we need *causal* Shapley values?

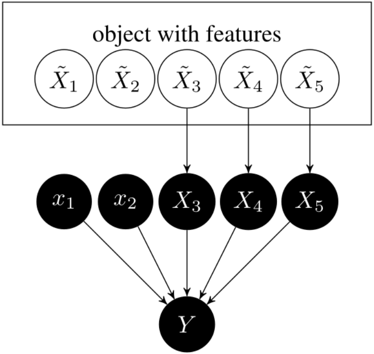
$$\mathbb{E}[X_1] = 0$$
$$\mathbb{E}[X_2] = 0$$
$$\mathbb{E}[X_2|x_1] = \alpha x_1$$
$$Y = \beta x_2$$



	Direct		*Evenly split*		*Root cause*	
	direct	indirect	direct	indirect	direct	indirect
ϕ_1	0	0	0	$\frac{1}{2}\beta\alpha x_1$	0	$\beta\alpha x_1$
ϕ_2	βx_2	0	$\beta x_2 - \frac{1}{2}\beta\alpha x_1$	0	$\beta x_2 - \beta\alpha x_1$	0

Incorporating causality into Shapley values

Idea 1: use marginal distributions for the value function[6]



$$P(\mathbf{X}_{\bar{S}} | \mathbf{X}_S = \mathbf{x}_S) = P(\mathbf{X}_{\bar{S}}) \implies$$

$$v(S) = \int d\mathbf{X}_{\bar{S}} P(\mathbf{X}_{\bar{S}}) f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S)$$

[6] Janzing, Minorics, and Bloebaum, "Feature relevance quantification in explainable AI".

Incorporating causality into Shapley values

Idea 2: choose coalitions based on known causal orderings[7]

For any permutation (arbitrary ordering) of the N variables, we define:

$$\phi_i(\pi) = \upsilon(\{j : j \preceq_\pi i\}) - \upsilon(\{j : j \prec_\pi i\}) \,,$$

with $j \prec_\pi i$ if j precedes i in the permutation π. Then

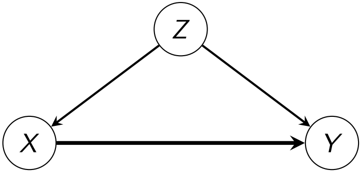
$$\phi_i = \sum_{\pi \in \Pi} \phi_i(\pi) \,,$$

where Π is the set of all permutation consistent with the causal structure between features. These Shapley values are no longer symmetric.

[7] Frye, Rowat, and Feige, "Asymmetric Shapley values".

Incorporating causality into Shapley values

Our idea: apply do-calculus[8]

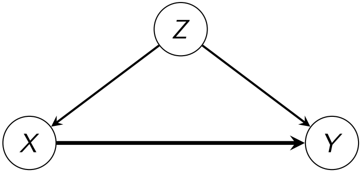


$$P(y|x)$$

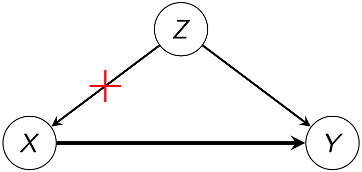
[8] Heskes et al., "Causal Shapley Values".

Incorporating causality into Shapley values

Our idea: apply do-calculus[8]



$P(y|x)$



$P(y|do(x))$

$$P(y|x) \neq \sum_z P(y|x,z)P(z) = P(y|do(x))$$

8 Heskes et al., "Causal Shapley Values".

Causal Shapley values

We define the value function as

$$v(S) = \mathbb{E}\left[f(\mathbf{X}) | do(\mathbf{X}_S = \mathbf{x}_S)\right] = \int d\mathbf{X}_{\bar{S}}\, P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S)\,,$$

where S is a coalition of players and $\bar{S} = N \setminus S$ is the set of players outside the coalition.

Given a complete causal ordering, the interventional distribution is:

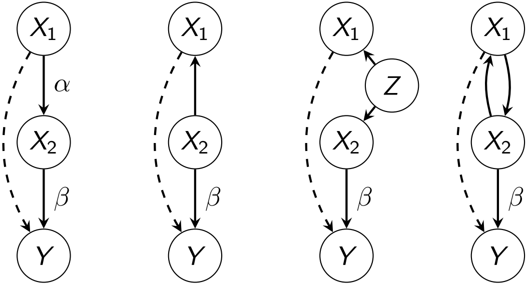
$$P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{j \in \bar{S}} P(X_j | \mathbf{X}_{pa(j) \cap \bar{S}}, \mathbf{x}_{pa(j) \cap S})\,,$$

where $pa(j) \cap S$ are the parents of j that are also part of the coalition S.

Causal Shapley values

$$\begin{aligned}
\mathbb{E}[X_1] &= 0 \\
\mathbb{E}[X_2] &= 0 \\
\mathbb{E}[X_2|x_1] &= \alpha x_1 \\
Y &= \beta x_2
\end{aligned}$$

Chain | Fork | Confounder | Cycle



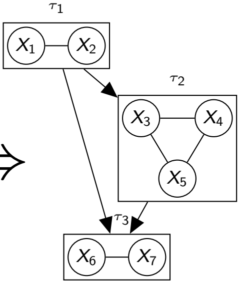
		Chain	Fork	Confounder	Cycle
marginal		D	D	D	D
conditional	symmetric	E	E	E	E
	asymmetric	R	D	E	E
causal	symmetric	E	D	D	E
	asymmetric	R	D	D	E

Causal Shapley values in practice

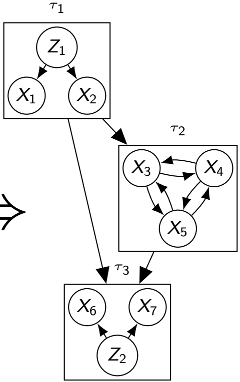
partial causal ordering

$(\{1, 2\}, \{3, 4, 5\}, \{6, 7\})$

$\Longrightarrow$



$\Longrightarrow$

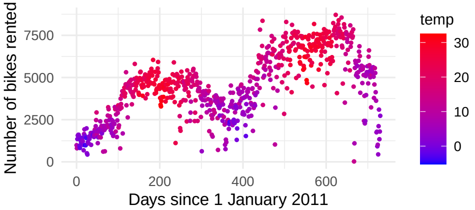


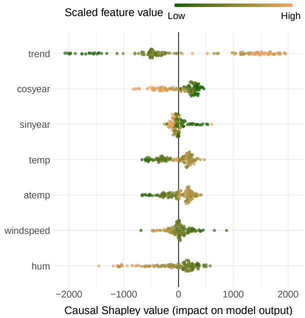
$$P(\mathbf{X}) = \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_\tau | \mathbf{X}_{pa(\tau)})$$

Causal Shapley values in practice

partial causal ordering

$$(\{trend\}, \{cosyear, sinyear\}, \{temp, atemp, windspeed, hum\})$$

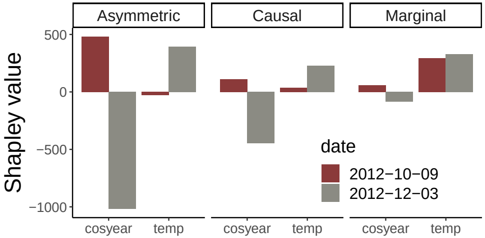


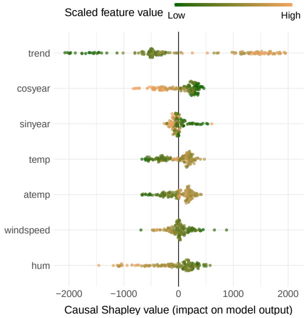


Causal Shapley values in practice

partial causal ordering

$$(\{trend\}, \{cosyear, sinyear\},$$
$$\{temp, atemp, windspeed, hum\})$$





Applications in healthcare

- Banerjee et al.[9] use causal Shapley values to understand the causal connections between socioeconomic metrics and the spread of COVID-19. They consider three plausible partial causal graphs.
- Su et al.[10] use causal Shapley values to address potential biases caused by confounding features in a study on predicting mortality of hemodialysis patients.
- Tyrovolas et al.[11] proposes the use of causal XAI for cancer diagnosis to address the vulnerability of predictive models to biases and spurious correlations.

[9] Banerjee et al., "Causal connections between socioeconomic disparities and COVID-19 in the USA".

[10] Su et al., "Prediction of mortality in hemodialysis patients based on autoencoders".

[11] Tyrovolas et al., "Towards Causal Explainable AI in Cancer Diagnosis".

Extensions

- Watson[12] propose *rational Shapley values*, which extend the methodology to also explain contrastive outcomes by shifting the reference distribution.
- Wang, Zhang, and Fu[13] perform causal discovery (with Direct-LiNGAM) to obtain a fully-specified causal graph that can be used to compute causal Shapley values.
- Ng et al.[14] incorporate causal strengths (estimated with IDA) into the SHAP algorithm by using them to reweigh Shapley values.

[12] Watson, "Rational Shapley Values".

[13] Wang, Zhang, and Fu, "Time series prediction of tunnel boring machine (TBM) performance during excavation using causal explainable artificial intelligence (CX-AI)".

[14] Ng et al., "Causal SHAP".

Conclusions

- When causal information (partial graph, strength estimates) is available, it is useful to incorporate it into your SHAP analysis to achieve a more causally intuitive feature attribution.
- Using the interventional distribution is optimal when one seeks explanations for the causal data-generating processes.[15]
- Causal Shapley values provide a principled way of incorporating causal information via do-calculus[16], that results in a sensible separation of direct and indirect effect contributions.

[15] Watson, "Rational Shapley Values".

[16] Pearl, "The seven tools of causal inference, with reflections on machine learning".

Thank you!

Direct and indirect effects

$$v(S) = \mathbb{E}\left[f(\mathbf{X})|do(\mathbf{X}_S = \mathbf{x}_S)\right] = \int d\mathbf{X}_{\bar{S}}\, P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S))f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S)\,.$$

$$\implies v(S \cup i) - v(S) =$$

$$\begin{aligned}
&= \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{S\cup i})|do(\mathbf{X}_{S\cup i} = \mathbf{x}_{S\cup i})] - \mathbb{E}[f(\mathbf{X}_{\bar{S}\cup i}, \mathbf{x}_S)|do(\mathbf{X}_S = \mathbf{x}_S)] && \text{(total effect)}\\
&= \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{S\cup i})|do(\mathbf{X}_S = \mathbf{x}_S)] - \mathbb{E}[f(\mathbf{X}_{\bar{S}\cup i}, \mathbf{x}_S)|do(\mathbf{X}_S = \mathbf{x}_S)] + && \text{(direct effect)}\\
&\quad \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{S\cup i})|do(\mathbf{X}_{S\cup i} = \mathbf{x}_{S\cup i})] - \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{S\cup i})|do(\mathbf{X}_S = \mathbf{x}_S)] && \text{(indirect effect)}
\end{aligned}$$

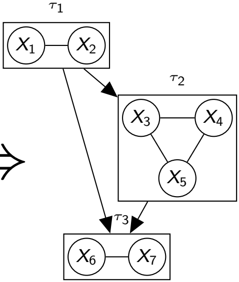
where $S \in N \setminus i$ is an arbitrary coalition and $\bar{S} = N \setminus (S \cup i)$.

Causal Shapley values in practice

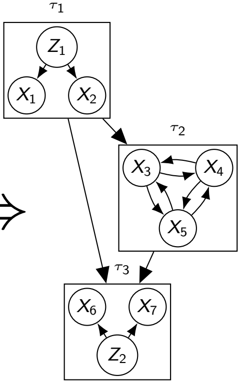
partial causal ordering

$(\{1,2\},\{3,4,5\},\{6,7\})$

$\Rrightarrow$



$\Rrightarrow$



$$
\begin{aligned}
P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) \quad = & \prod_{\tau \in \mathcal{T}_{\mathrm{confounding}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}) \times \\
\times & \prod_{\tau \in \mathcal{T}_{\overline{\mathrm{confounding}}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}, \mathbf{x}_{\tau \cap S})
\end{aligned}
$$

References I

Aas, Kjersti, Martin Jullum, and Anders Løland. "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values". In: *Artificial Intelligence* 298 (Sept. 1, 2021), p. 103502. ISSN: 0004-3702. DOI: 10.1016/j.artint.2021.103502. URL: https://www.sciencedirect.com/science/article/pii/S0004370221000539 (visited on 02/01/2026).

Banerjee, Tannista et al. "Causal connections between socioeconomic disparities and COVID-19 in the USA". In: *Scientific Reports* 12.1 (Sept. 22, 2022). Publisher: Nature Publishing Group, p. 15827. ISSN: 2045-2322. DOI: 10.1038/s41598-022-18725-4. URL: https://www.nature.com/articles/s41598-022-18725-4 (visited on 02/02/2026).

Carloni, Gianluca, Andrea Berti, and Sara Colantonio. "The Role of Causality in Explainable Artificial Intelligence". In: *WIREs Data Mining and Knowledge Discovery* 15.2 (2025). _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.70015, e70015. ISSN: 1942-4795. DOI: 10.1002/widm.70015. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.70015 (visited on 01/30/2026).

Frye, Christopher, Colin Rowat, and Ilya Feige. "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 1229–1239. URL: https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html (visited on 02/01/2026).

Heskes, Tom et al. "Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 4778–4789. URL: https://proceedings.neurips.cc/paper/2020/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html (visited on 01/04/2026).

References II

Janzing, Dominik, Lenon Minorics, and Patrick Bloebaum. "Feature relevance quantification in explainable AI: A causal problem". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, June 3, 2020, pp. 2907–2916. URL: https://proceedings.mlr.press/v108/janzing20a.html (visited on 02/01/2026).

Lundberg, Scott M and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html (visited on 02/01/2026).

Ng, Woon Yee et al. "Causal SHAP: Feature Attribution with Dependency Awareness through Causal Discovery". In: *2025 International Joint Conference on Neural Networks (IJCNN)*. 2025 International Joint Conference on Neural Networks (IJCNN). ISSN: 2161-4407. June 2025, pp. 1–8. DOI: 10.1109/IJCNN64981.2025.11228295. URL: https://ieeexplore.ieee.org/abstract/document/11228295 (visited on 02/03/2026).

Pearl, Judea. "The seven tools of causal inference, with reflections on machine learning". In: *Commun. ACM* 62.3 (Feb. 21, 2019), pp. 54–60. ISSN: 0001-0782. DOI: 10.1145/3241036. URL: https://dl.acm.org/doi/10.1145/3241036 (visited on 01/30/2026).

Pichler, Maximilian and Florian Hartig. "Machine learning and deep learning: A review for ecologists". In: *Methods in Ecology and Evolution* 14.4 (2023). _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14061, pp. 994–1016. ISSN: 2041-210X. DOI: 10.1111/2041-210X.14061. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14061 (visited on 01/30/2026).

References III

Sloman, Steven. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press, Aug. 18, 2005. ISBN: 978-0-19-518311-5. DOI: 10.1093/acprof:oso/9780195183115.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780195183115.001.0001 (visited on 01/30/2026).

Su, Shuzhi et al. "Prediction of mortality in hemodialysis patients based on autoencoders". In: *International Journal of Medical Informatics* 195 (Mar. 1, 2025), p. 105744. ISSN: 1386-5056. DOI: 10.1016/j.ijmedinf.2024.105744. URL: https://www.sciencedirect.com/science/article/pii/S1386505624004076 (visited on 02/03/2026).

Tyrovolas, Marios et al. "Towards Causal Explainable AI in Cancer Diagnosis: Advances, Challenges, and Future Directions*". In: *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). ISSN: 2694-0604. July 2025, pp. 1–7. DOI: 10.1109/EMBC58623.2025.11253261. URL: https://ieeexplore.ieee.org/abstract/document/11253261 (visited on 02/02/2026).

Wang, Kunyu, Limao Zhang, and Xianlei Fu. "Time series prediction of tunnel boring machine (TBM) performance during excavation using causal explainable artificial intelligence (CX-AI)". In: *Automation in Construction* 147 (Mar. 1, 2023), p. 104730. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2022.104730. URL: https://www.sciencedirect.com/science/article/pii/S0926580522006008 (visited on 01/30/2026).

Watson, David. "Rational Shapley Values". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 1083–1094. ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533170. URL: https://dl.acm.org/doi/10.1145/3531146.3533170 (visited on 02/02/2026).

References IV

Wu, Xing et al. "Methods and Applications of Causal Reasoning in Medical Field". In: *2021 7th International Conference on Big Data and Information Analytics (BigDIA)*. 2021 7th International Conference on Big Data and Information Analytics (BigDIA). Oct. 2021, pp. 79–86. DOI: 10.1109/BigDIA53151.2021.9619639. URL: https://ieeexplore.ieee.org/abstract/document/9619639 (visited on 01/30/2026).