Causality and Explainable AI: A Review of the Literature

Leo Jean Paggen
Maastricht University (UM)
Maastricht, The Netherlands
l.paggen@student.maastrichtuniversity.nl

Abstract

While traditional **Explainable Artificial Intelligence (XAI)** methods have enabled researchers to understand why models produce a certain output based on observed data, their ability to provide actionable insights remains limited. Popular XAI techniques focus on statistical correlation between variables, and provide insights about which variables a model found to be the most influential, but do not look at the underlying causal relationships between the variables of a model. A consequence of the nature of such frameworks is that such explanations are insufficient to guide future interventions, or predict the effect of any given intervention. Recognizing these flaws, much of the field of XAI has shifted towards **causal inference**, enabling domain experts to not only understand why a prediction was made, but also which changes in the input would lead to a desired outcome. This review analyzes influential work in **Causal Explainable Artificial Intelligence (CXAI)** to identify advances in the field, the potential adoption of such frameworks, and reliability of these methods.

1 Introduction

The rapid increase in the complexity of machine learning models has prompted a critical need for frameworks to provide transparency to these models in an effort to increase their trustworthiness. This need is particularly crucial in fields such as healthcare, and domains where bias-mitigation is a priority. Two of the most widely used XAI methods, Local Intepretable Model-Agnostic Explanations (LIME) [11] and SHapley Additive exPlanations (SHAP) [8], can quantify the importance of any input variables with respect to an output. These methods make strong assumptions such as **feature independence**, which prevents them from providing insights about the true effect of altering input variables. Such insights fall under the domain of causal inference, pioneered by Judea Pearl. In his foundational work, Pearl (1985) laid the groundwork for **Bayesian Networks** (BN), which use **Directed Acyclic Graphs** (DAG) to represent probabilistic dependencies [6].

In 2009, Pearl would shift from networks built on probability theory only and introduce the **do-calculus**, a set of rules which allows for the prediction of an intervention using only observed data. These rules, applied together with **Structural Causal Graphs** establish the possibility of computing causation from correlation, provided a correct causal graph is available.

Despite the formalization of causal logic, a gap between theoretical frameworks and applied machine learning remained. Standard XAI techniques focused on quantifying the impact of variables after models were trained, but bypassed causal logic entirely. However, recent work has begun integrating some of Pearl's work in the XAI pipelines we commonly use. This review examines five papers on the subject, moving from theoretical to practical frameworks such as **causal Shapley Values** [5].

2 Methodology

The five papers selected for this review are all issued from top venues and conferences in machine learning. The papers were found by querying Arxiv with the following list of keywords: *causal graphs*, *causality*, *causal shapley values*, *causality and explainable AI*. A first batch of papers was selected, and was further pruned by keeping only those papers which appeared in top machine learning conferences. Additional relevant resources were selected by following the bibliography sections of the papers found on Arxiv. Five core papers were then selected, with each one having made a different contribution to the field of causality and XAI.

This review aims to analyze the five aforementioned core papers along the following dimensions:

(1) **Limitations**
(2) **Domain** - theoretical, practical
(3) **Complexity** - possibility of widespread adoption
(4) **Rung on Pearl's ladder of causation**

These four dimensions are chosen due to their relevance in the XAI field. Many papers lay the groundwork for future work but remain entirely theoretical, as is the case for Beckers (2022) [2] and Pearl (1985, 2009) [6, 9]. Other work incorporates these theoretical frameworks into consideration, while making them a viable and practical approach for researchers to use, as observed in the work of Heskes et al. (2020) [5].

3 Concepts in Causality and Causal Inference

Correlation alone is not enough to draw causal relationships: knowing that two variables move together does not allow to draw conclusions about what happens to one if the other is varied. Pearl (1985) [6] introduced the **Bayesian Network**, a DAG which represents conditional dependencies between variables of a model. BNs lacked a crucial component; they can tell us about how two variables depend on each other with a certain probability, but cannot provide information about the factors which influence a certain variable. **Structural Causal Models** (SCM), introduced by Pearl (2009) [9] encode exactly that information, and enable interventions by allowing us to force a variable to a certain value instead of observing it. SCMs extend BNs with functional equations. Each variable is a function of its parents, enabling the use of do-calculus to cut edges and simulate intervention. Do-calculus differs from simply conditioning on a variable X: it simulates an intervention on that variable by setting it to a value and cutting all edges representing incoming causal relations on X. This enables the prediction of interventions from observational data, **provided the causal graph is known**. This strong assumption is a critical flaw of the SCMs, which the field of **Causal Discovery** aims to bypass.

Pearl's work can be summarized using his **ladder of causation**, which consists of three rungs of increasing causal expressiveness.

The first rung, **association**, which covers the conditional dependencies between variables, as represented by BNs. The second rung, **intervention**, examines the effect on a model's output when a variable is changed, represented by do-calculus. The last rung of the ladder is **counterfactuals**, which is concerned with answering the question of what would have happened if a variable had been different.

The core papers in this review each operate on one or more area of the causal ladder, which is summarized in table 1.

Table 1: Pearl's Ladder of Causation and XAI relevance

Rung	Type	Question	XAI Relevance
1	Association	How does seeing X change my belief about Y?	SHAP, LIME
2	Intervention	What happens to Y if I force X?	Causal Shapley Values, Causal Graphs
3	Counterfactual	What would Y have been had X been different?	Counterfactual XAI

4 Comparison of the Five Core Papers

This section compares and summarizes the findings of the five core papers across the three dimensions defined in section 2. The chosen papers are listed along with the journal/venue they appear in, in table 2. Heskes et al. (2020) [5], Frye et al. (2020) [4] appear in *Neural Information Processing Systems* (NeurIPS). Beckers (2022) [2] appears in *Proceedings of Machine Learning Research*, Vowels et al. (2022) [14] appears in *ACM Computing Surveys*, and Karimi et al. (2020) [7] appears in *ACM Conference on Fairness, Accountability, and Transparency*. This review tells a story about each paper's contribution to the field, ordered by their importance and limitations.

4.1 Towards Causal Shapley Values

Slack et al. (2020) [12] demonstrated empirically that both SHAP and LIME can be fooled easily through adversarial attacks. This highlights the key vulnerability of these methods: their assumptions are too simple and fragile when it comes to practice.

We use the terminology employed by Heskes et al. (2020) [5] to distinguish **Marginal** SHAP from **Conditional** SHAP. Marginal SHAP is the weakest version of SHAP when it comes to causality: it assumes feature independence, and as a result ignores causal structures. Aas et al. (2020) [1] improve upon Marginal SHAP by introducing Conditional SHAP, which samples from the *conditional distribution* rather than the marginal one. Intuitively, where marginal SHAP might consider *age* important irrespective of a patient's *weight*, conditional SHAP recognizes these two features can co-vary and samples more realistic combinations of the two. While conditional SHAP certainly is an improvement over marginal SHAP, we must recognize that conditioning on a variable is not equivalent to an intervention on said variable: conditional SHAP does not address causation, and falls onto the first rung of Pearl's ladder of causation.

We define the 4 axioms[1] which Shapley values are grounded by following the terminology of Frye et al. (2020) [4]: **Efficiency**, **Linearity**, **Nullity** (defined as "Null player (dummy)" in Heskes et al. (2020) [5]), and **Symmetry**. Efficiency implies the sum of each individual attribution sums to the total attribution. Linearity means that the Shapley values of a linear ensemble of models can be obtained as a linear combination of its constituent models. Nullity guarantees that a feature which is irrelevant to the model's output receives a 0 Shapley value, and finally, Symmetry requires that for two features which contribute equally as much to the model's output, an identical Shapley value is attributed to both features.

Frye et al. (2020) critique this final axiom of symmetry, highlighting the fact that some features in a model *cause* other features to have certain values. To illustrate why this assumption fails to capture reality, take the following example presented by Heskes et al. (2020) [5]: assume a model predicts bike sales are equally influenced by season and temperature, it should be apparent that season determines temperatures to some degree, not the other way around. In practice, Shapley values would assign an equal score to both season and temperature, ignoring this causal chain. This axiom of symmetry is exactly what Frye et al. (2020) drop in their work, leading to **asymmetric Shapley values** (ASV), which satisfy all but this fourth axiom of symmetry. By relaxing this fourth axiom, ASVs can incorporate causal ordering over features; rather than treating features as peers, we allow features which causally precede others to be evaluated first, giving more importance to root causes. The strength of this contribution is that ASVs do not require full knowledge of the causal graph to be effective, any (limited or full) knowledge about the data-generating process[2] to be incorporated into the model explanation, lowering the barrier to practical adoption. Frye et al. compute ASVs on the *US Census Income* dataset, and find that *sex* (gender) is a more important causal predictor than all other features, because it is the causal ancestor of other features such as *occupation* or *working hours*, a conclusion neither marginal or conditional Shapley values are able to come to due to their nature.

Heskes et al. (2020) [5] would go on to demonstrate that ASVs cannot distinguish between confounders and cycles in a causal graph, limiting their reliability in practice. Additionally, they claim that ASVs can be exploited by *zero strength causal links*, which we explain in the next section.

4.2 Formalization of Causal Shapley Values

Heskes et al. (2020) [5] argue that ASVs are not resilient enough to be considered true causal Shapley values. They state that ASVs are sensitive to *zero strength causal links*, which they support via the following example: consider a model perfectly trained to predict a binary outcome in which two mutually exclusive features contribute equally to said outcome[3]. Suppose we wish to use ASVs and incorporate into the model that feature $X1$ causally precedes feature $X2$, and assign the causal strength of this arrow to zero, meaning the influence of $X1$ on $X2$ is null. A well-behaved model

[1] The original Lundberg and Lee (2017) [8] paper uses a different terminology

[2] This often refers to the graphs which can represent the causal chain which produced the data

[3] They use the example of a neural network used to predict the XOR of two features in their work, this explanation is equivalent

should ignore this, but the ASV framework would instead wrongly attribute 100% of the outcome to $X2$.

Heskes et al. (2020) [5] further claim that ASVs apply conditioning by observation, which is insufficient and unnecessary for causal Shapley values. They argue that conditioning by intervention is both necessary and sufficient to obtain causal Shapley values, which they do through the use of Pearl's do-calculus [9] on causal DAGs. The difference between conditioning on intervention and observation is key: conditioning on $X = x$ asks *what to expect given we observe* $X = x$, whereas $do(X = x)$ asks *what do we expect given we force* X *to take on value* x. The latter effectively cuts all causal influences on X, allowing us to isolate true causal contribution. On the other hand, conditioning on observation conflates the causal effect of a feature with the features influencing it (confounders).

To further illustrate the failure points of marginal and conditional (including asymmetric) Shapley values, we include the four causal graph structures presented in Heskes et al. (2020) [5], which are displayed under figure 1 of the appendix. Marginal Shapley values only provide direct effects for all graphs, meaning they cannot recognize chains and causal precedence in a graph. Symmetric conditional Shapley values can encode more information than marginal values, but assign equal credit to all features influencing the same variable (axiom of symmetry). Once the axiom of symmetry is dropped, we get ASVs, which correctly give all credit to the root feature of a graph, but fail to detect confounders, instead assigning equal importance to two variables influencing the same variable. In the case of the causal Shapley values developed in their work, symmetric and asymmetric causal Shapley values behave similarly, except for the fact that asymmetric Shapley values are able to assign full credit to the root of the chain graph, compared to symmetric values, which split the effect equally across all features in the chain. Heskes et al. (2020) [5] conclude that causal Shapley values are the only type of Shapley values which can provide sensible causal explanations across the four types of causal structures, provided a fully correct causal graph is defined.

Despite their theoretical soundness, causal Shapley values suffer from the same limitations as the SCMs they are built on: the full causal graph must be known and defined correctly. In practice, two options arise: either the graph is defined by a domain expert, a costly and error-prone process, or it is learned through means of *causal discovery*. This assumption of correctness, taken for granted by Heskes et al., is what Beckers (2022) [2] critiques and identifies as a fundamental obstacle in CXAI.

4.3 An Unsolved Challenge in CXAI

In his work, Beckers (2022) [2] claims that the correctness of causal explanations hinges upon the correctness of the causal graph used to generate them. This argument, while seemingly simple and obvious, is stronger than it appears: Beckers does not only highlight the necessity of the causal graph being correct, but also formalizes the conditions under which causal explanations can be considered valid. His main argument is that every paper on CXAI so far has failed to provide a mechanism to validate the assumption that a causal graph is correct, instead blindly relying on an assumption which does not hold in many real-life scenarios, such as healthcare, where the causal graph is rarely ever known.

For Beckers, explaining a model is not the same as explaining the real world mechanisms which the model aims to represent. In his work, he argues that a model may provide valid causal explanations in its own context, but that such explanations may be invalid or misleading if the model is an ill-defined representation of reality. To categorize causal explanation types, Beckers defines the following taxonomy: **sufficient explanations**, which identify the smallest set of features which contribute to the outcome; **direct explanations**, which isolate the immediate cause of an outcome; and **counterfactual explanations**, which ask what the value of the outcome would have been had we changed the value of another feature. Clearly, this highlights a gap in the work of Frye et al. (2020) [4] and Heskes et al. (2020) [5], as neither work provides sufficiency or counterfactual explanations, Shapley values merely assign credit to features which contributed to the outcome. Even in the case where the assumption made by Heskes et al. holds (the causal graph is correct), causal Shapley values have no way of telling us which features are enough to predict the outcome. Beckers raises concerns about the importance of a valid causal graph, the next section covers methods to learn such graphs from observational data.

4.4 Causal Discovery

Causal Discovery is a sub-field of causality which encompasses methods used to extract causal graphs from data. The work of Vowels et al. (2022) [14] encompasses a wide range of causal discovery methods, which we classify into the following families[4]: *score-based* methods, *constraint-based* methods, and *functional methods*. Score-based methods evaluate the correctness of a learned graph against a scoring function S. The most widely-cited score-based algorithm for causal discovery is the *Greedy Equivalence Search* (GES) by Chickering (2002) [3]. The GES algorithm works by adding edges to a graph with the objective of maximizing a scoring function, then greedily removes edges which improve the score. Constraint-based algorithms work by testing conditional independences between variables, removing edges between independent variables. Examples of constraint-based algorithms include the popular *PC* algorithm [13], which requires no hidden confounders. Vowels et al. highlight the fact that the two families of algorithms we mentioned thus far have the important limitation that they produce **Markov Equivalence Classes** (MEC) rather than unique causal graphs. A MEC is a set of graphs which encode the same set of conditional independencies, without providing directional edges, ignoring causal precedence altogether. Clearly, since causal Shapley values require do-calculus, they require a directed graph, which functional methods may produce. Functional methods, which include the *Linear Non-Gaussian Acyclic Model* (LiNGAM), can achieve full identifiability[5] at the cost of stricter assumptions. To illustrate the nature of those assumptions, we briefly explain the assumptions behind LiNGAM: **linearity** states that each variable is a linear function of its parents, plus noise; and *non-Gaussianity*, requires the noise in such relationships to follow a non-Gaussian distribution. LiNGAM achieves full identifiability through the second assumption, as exploiting

[4]They use the terminology: "constraint-based, score-based, those exploiting structural asymmetries, and those exploiting various forms of intervention". We simplify their work for the sake of clarity.

[5]Refers to the ability to determine the true causal graph from the data alone

Paper	Rung	Domain	Key Contribution	Limitations	Practicality
Heskes et al. (2020)	Intervention	Practical	Causal Shapley Values	Requires known causal graph	Medium
Frye et al. (2020)	Intervention	Theoretical	Asymmetric Shapley Values	Partial ordering still assumed	Low
Beckers (2022)	Intervention	Critical	Formal critique of causal XAI assumptions	No implementation proposed	Low
Vowels et al. (2022)	All	Theoretical	Survey of causal discovery methods	Does not address XAI directly	Low
Karimi et al. (2020)	Counterfactual	Practical	SCM-grounded algorithmic recourse	Requires known causal graph	Medium

Table 2: Comparison of the 5 core papers of this review along the dimensions of interest. The *Rung* refers to Pearl's ladder of causation, presented in Pearl (2009)

this asymmetry in the residuals allows it to infer causal precedence. To illustrate the workings of this second assumption, consider the following scenario: if we regress *income* against *age*, we should observe a non-Gaussian distribution of the residuals of said regression only if age truly causes income. Reversing this regression should result in a drastically different distribution, revealing the direction of causality. Conversely, if we regress two variables against each other and observe a Gaussian distribution of the residuals, even after reversing the regression, we cannot say anything about causality without domain knowledge, which is precisely where the critique of Beckers (2022) [2] still holds: we have no standard and generally accepted way of verifying the validity of a causal graph without extensive domain knowledge.

4.5 Counterfactual Explanations

A promising alternative to the often too limited interventional frameworks is that of counterfactual explanations (CFE), which operate on a rung separate from intervention. Karimi et al. (2020)[7] argue CFEs fail in practice because they make the assumptions that features are independent, a conclusion in line with the literature reviewed thus far. Their key contribution to the field is *Recourse through Minimal Interventions* (MINT), which aims to find the minimal set of do-operations which produce a CFE which leads to a desired change in the output. MINT operates on an SCM, and can effectively be thought of as a downstream cause-and-effect waterfall: any intervention (do-operation) in the SCM propagates through the graph, updating all downstream variables according to the defined structural equations of the graph. Concretely, MINT follows the *Abduction-Action-Prediction* steps defined by Pearl (2013) [10]: abduction computes values for all exogenous variables; action modifies the SCM according to the desired interventions via do-operations; and lastly prediction determines values for the endogenous variables by using the exogenous variables obtained in step 1, and the SCM obtained from step 2. We no longer look at an explanation after an intervention, rather we find cheap (least effort) actions to guide the output of a model in any desired direction. Karimi et al. illustrate the power of MINT on the *German Credit* dataset: MINT finds a course of action which is 42% cheaper than what a standard CFE would recommend[6].

Unfortunately, MINT itself is limited by the fact that it relies on SCMs, and therefore assume a valid causal graph is known. This is a problem Karimi et al. acknowledge, and defer to future work.

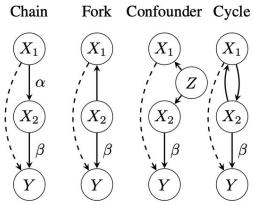
5 Discussion and Conclusion

This review tells a story about recent developments in CXAI by relating five key papers to Pearl's seminal ladder of causation. Research in CXAI has led to important advances, like causal Shapley values, or MINT, but their implementation and wider adoption by researchers remains limited by the strong and unrealistic assumptions they make. We covered methods related to interventions (Second rung of Pearl's ladder), presented a critique of said methods, discussed the field of causal discovery, which aims to respond to this critique, and presented MINT as a fundamental shift from explanation to action, which operates on the third rung of the ladder. A similar conclusion is made across all five papers: a correct causal graph is required for true causal explanations to be valid. Heskes et al. need this requirement for do-calculus, Karimi et al. need it for MINT, and Beckers explains why the absence of this requirement invalidates causal explanations. The analysis of the work of Vowels et al. led us to the conclusion that no causal discovery method which is universally considered valid exists, leading us to the conclusion that while this problem remains unsolved, the gap between theoretical CXAI and its practical adoption will remain open.

Future work must address the two following points: firstly, work must be done to develop a robust causal discovery method which can learn valid causal graphs from observational data alone, and secondly, a set of metrics to evaluate the validity of such a graph must be developed, with a focus on the importance of the researchers understanding said metrics.

[6]The complete example is too mathematically heavy for this review, refer to Karimi et al. (2020) [7] for a complete example

6 Appendix



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

Figure 1: The four causal graph structures presented in Heskes et al. (2020) [5]. For each type of Shapley value discussed in their work, a letter is assigned to represent a method's effectiveness. *D* stands for "direct effects", *E* stands for "indirect effects which are split evenly between two features", and *R* stands for "all credit goes to the root cause".

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2020. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. arXiv:1903.10464 [stat.ML] https://arxiv.org/abs/1903.10464

[2] Sander Beckers. 2022. Causal Explanations and XAI. arXiv:2201.13169 [cs.AI] https://arxiv.org/abs/2201.13169

[3] David Maxwell Chickering. 2003. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3, null (March 2003), 507–554. doi:10.1162/153244303321897717

[4] Christopher Frye, Colin Rowat, and Ilya Feige. 2021. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. arXiv:1910.06358 [stat.ML] https://arxiv.org/abs/1910.06358

[5] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. arXiv:2011.01625 [cs.AI] https://arxiv.org/abs/2011.01625

[6] Pearl Judea. 1985. *Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning*. Technical Report. University of California, Los Angeles.

[7] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic Recourse: from Counterfactual Explanations to Interventions. arXiv:2002.06278 [cs.LG] https://arxiv.org/abs/2002.06278

[8] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI] https://arxiv.org/abs/1705.07874

[9] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, Cambridge.

[10] Judea Pearl. 2013. Structural Counterfactuals: A Brief Introduction. *Cognitive Science* 37, 6 (2013), 977–985. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12065 doi:10.1111/cogs.12065

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG] https://arxiv.org/abs/1602.04938

[12] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. arXiv:1911.02508 [cs.LG] https://arxiv.org/abs/1911.02508

[13] Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search*. Massachusetts Institute of Technology.

[14] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. 2021. D'ya like DAGs? A Survey on Structure Learning and Causal Discovery. arXiv:2103.02582 [cs.LG] https://arxiv.org/abs/2103.02582