



Research on Methods for Improving the Explainability of Artificial Intelligence Based on Causal Reasoning

Lingling He

¹Zhejiang Ronghe Tea Culture Co., Ltd., Hangzhou 310000, Zhejiang, China.

²Yunnan Ronghehao Tea Industry Co., Ltd., Kunming 650000, Yunnan, China.

³Kunming Zhonggu Tea Industry Co., Ltd., Kunming 650506, Yunnan, China.

How to cite this paper: Lingling He. (2025) Research on Methods for Improving the Explainability of Artificial Intelligence Based on Causal Reasoning. *Advances in Computer and Communication*, 6(2), 87-93. DOI: 10.26855/acc.2025.04.006

Received: March 19, 2025

Accepted: April 16, 2025

Published: May 13, 2025

***Corresponding author:** Lingling He, Zhejiang Ronghe Tea Culture Co., Ltd., Hangzhou 310000, Zhejiang, China; Yunnan Ronghehao Tea Industry Co., Ltd., Kunming 650000, Yunnan, China; Kunming Zhonggu Tea Industry Co., Ltd., Kunming 650506, Yunnan, China.

Abstract

In the era of rapid development of AI, although AI systems have made significant progress in performance in many areas, the lack of interpretive capability has been a major obstacle to further popularization and application, especially in key areas such as medicine, finance, and autonomous driving. Causal reasoning aims to study the cause-and-effect relationships between variables, which gives a new look at improving the interpretability of AI. This article discusses ways and means to increase the interpretability of AI through causal reasoning. Initially, the meaning of the respective concepts and their connection with the interpretability of AI and other theoretical foundations is laid out. After that, concrete paths of realization are given, such as cause-and-effect modeling, cause-and-effect analysis, and generation of interpretation of cause-and-effect instructions. Testing specific examples, the effectiveness and practical value of these methods are confirmed in typical use cases. Integrating cause-and-effect connections into an AI system can make the AI decision-making process more transparent and understandable, which is of great importance for improving AI reliability and recognition.

Keywords

Causal reasoning; Artificial intelligence explainability; Causal modeling; Counterfactual analysis; Explanation generation

1. Introduction

Thanks to the constant development of technology, artificial intelligence has infiltrated all areas of life. Machine learning algorithms, such as deep learning models, are effective in processing image recognition, natural language, and data prediction tasks. These models often show the characteristics of a black box, and the decision mechanism in them is difficult to capture. For example, in the key area of medical diagnosis, accuracy and interpretability are necessary, and the unexplained nature of the artificial intelligence system can cause distrust among users, and then lead to serious diseases. It's not impossible. For example, in medicine, if doctors or patients cannot understand the reason for making a certain diagnosis with artificial intelligence, they can abandon it, which leads to limiting the possibility of medical application of artificial intelligence.

The problem of artificial intelligence's interpretability has attracted much attention from academia and industry. Traditional methods, such as analysis of the importance of characteristics and visualization of models, have obvious limitations. These methods usually give only superficial explanations of the behavior of the model, and it is difficult to perform an in-depth analysis of the cause-and-effect mechanism behind decision-making. Causal reasoning focuses

on identifying causal elements that influence an event and the path of its action. Application of cause-and-effect links in the field of artificial intelligence can increase the interpretability of the system in order to achieve deeper and more practical results, allowing users not only to know the content of the decision-making process of artificial intelligence, but also to understand the reasons behind its acceptance.

This study aims to explore effective ways to improve the interpretability of artificial intelligence through cause-and-effect relationships. Discuss the theoretical foundations and practical application of cause-and-effect relationships in the field of artificial intelligence interpretability, and try to propose a number of possible solutions to the "black box" problem of artificial intelligence. The value of this study is reflected in both theory and practice. In theory, he expanded the field of artificial intelligence interpretability research and created a new theoretical base based on cause-and-effect relationships. In practice, this can increase the transparency and completeness of AI systems, reduce the degree of risk in decision-making, and increase user confidence in AI systems, thus promoting the full application of AI in many areas.

2. Theoretical Foundations of Causal Reasoning and AI Explainability

2.1 Relevant Concept Definitions

2.1.1 Causal Reasoning

Causal reasoning refers to the process of inferring causal relationships between events or variables. It involves identifying the causes that lead to a particular effect and understanding the mechanisms by which these causes act. In philosophy, causality has been a long-standing research topic, and different theories of causality have been proposed, such as Hume's regularity theory and counterfactual theory. In the field of statistics and machine learning, causal reasoning has been gradually developed into a set of systematic methods, such as causal graphs, structural causal models (SCMs), and causal inference algorithms. As shown in the following table:

Table 1. Comparison of causal reasoning and AI interpretative concepts

Concept	Define the core	Core objectives	Key methods	Application value
Causal reasoning	To infer the causal relationship between events or variables and reveal the causal mechanism	Identify causal factors and distinguish between causality and correlation	Causal graph, structural causal model (SCM), counterfactual analysis	Provide deep causal basis for AI decision making
AI explanatory	The AI system provides humans with understandable decision interpretation capabilities	Improve model transparency and enhance user trust	Feature importance analysis, model visualization, causal driven interpretation generation	Promote the trusted application of AI in medical, financial and other fields

2.1.2 AI Explainability

The interpretability of artificial intelligence means the ability of an artificial intelligence system to explain its decisions or behavior to people. Interpreted artificial intelligence must clearly present the reasons and framework for decision-making so that users can understand and act correctly. This ability to interpret is usually divided into two aspects: global and local: the former focuses on the general logic of the model's operation, while the latter focuses more on individual decisions made for specific inputs [1].

2.2 Correlation Analysis Between Causal Reasoning and AI Explainability

Cause-and-effect relationships are closely related to the interpretability of artificial intelligence, and the former may be the theoretical basis for the latter. After modeling causality in the data, we will be able to better understand how the input variables affect the output of the artificial intelligence model and based on this we will be able to provide a more meaningful explanation. As with medical prognostic models, cause-and-effect relationships can be used to identify factors that have a key impact on disease outcomes and explain to doctors or patients why they are important. In

addition to that, in order to ensure the interpretability of artificial intelligence, one correlation analysis is not enough, it is necessary that cause-and-effect relationships cross this limit. Many traditional methods focus only on the degree of correlation between characteristics and results, but it is only a simple correlation, not a cause-and-effect relationship. Causal reasoning makes it possible to distinguish a true cause-and-effect relationship from pseudocorrelation, and then helps to obtain more accurate and reliable explanations. As shown in the following table:

Table 2. Comparison of AI interpretative theoretical systems based on causal reasoning

System info	Statistics and machine	Core methods	Application scenarios	Superiority	Boundedness
Structural causal model (SCM)	Based on counterfactual causality theory	Causal graph modeling, intervention analysis, counterfactual reasoning	Causal analysis and decision attribution of complex systems	Systematically model causality and support multi-dimensional interpretation	Causal discovery relies on prior knowledge or high-quality data
Counterfactual explanations	Judea Pearl (causal graph theory)	If-then hypothesis analysis (e.g., "What happens to output Y if variable X changes")	Local interpretation (specific input decision attribution)	Intuitively show the path of influence of variables on output	The computing cost is high in high-dimensional data
Causal effect estimation	Proposer/core idea	Double robust estimation, instrumental variable method, and matching method	Feature importance quantification and decision factor priority ranking	Quantify the degree of causal influence and support the global interpretation	Not affected by unobserved confounding variables

2.3 Review of Existing Theoretical Systems

The field of artificial intelligence interpretability includes many theoretical systems related to cause-and-effect relationships. The structural causal model (SCM) proposed by Pearl occupies a special place in these systems. The model uses cause-and-effect diagrams to show a cause-and-effect relationship between variables, forming the basis for cause-and-effect conclusions and interpretation. There are also some scientists who focus on the counterfactual interpretation of this system and clarify the decision-making process in artificial intelligence models by exploring possible outcomes in the case of different input variables. Counterfactual interpretation is based on causality theory, which allows users to clearly understand the role of each input variable in the output. A clear recognition and understanding of the students about this complex relationship will contribute to a more accurate construction of the model.

In recent years, many results have emerged in the investigation of the combination of cause-and-effect relationships and the interpretability of artificial intelligence. For example, some researchers have attempted to use cause-and-effect diagrams to model decision-making processes in deep learning models to generate cause-and-effect explanations; others have focused on developing causal inference algorithms, trying to find causal factors from the data and using them to analyze model prediction results. This type of research has led to the emergence of new ideas and directions at both theoretical and practical levels.

3. Methods for Improving AI Explainability Based on Causal Reasoning

3.1 Causal Modeling Methods

A key step in improving the interpretability of artificial intelligence is causal modeling. Building a cause-and-effect diagram is the main goal of cause-and-effect modeling. The diagram should represent a cause-and-effect relationship

between input variables, internal model variables, and the output. In causal modeling, all this can be roughly divided into two types of methods: knowledge-oriented methods and data-dependent methods.

Data-based causal simulation uses statistical methods to infer cause-and-effect relationships from data. Cause-and-effect relationship detection algorithms can be used to derive causality diagrams from observational data. Such algorithms analyze statistical dependencies between variables and identify causal structures that are consistent with the data. Knowledge-based causal modeling relies on prior knowledge and expert knowledge from the domain to construct cause-and-effect diagrams. In many applied scenarios, especially in complex areas such as health and finance, a priori knowledge is crucial to accurately modeling causality.

After building a causal diagram, it can be used to analyze the cause-and-effect impact of various variables on the output of an artificial intelligence model. For example, a cause-and-effect graph can be used to determine which input variables are the immediate cause of the output, and which variables are associated with the output only through the intermediate variables, as shown in the following table:

Table 3. Comparison of causal modeling methods

Knowledge-driven modeling	Core technology	Superiority	Applicable scene	Throw down the gauntlet
Data-driven modeling	Cause discovery algorithm (such as PC algorithm, IC algorithm), machine learning inference	No prior knowledge is required, and it is suitable for exploratory analysis	Scenarios where data is abundant but domain knowledge is lacking (such as causal analysis of image data)	The accuracy of causal graph inference decreases in high-dimensional data, and the computational complexity is high
Method type	Domain experts build causal graphs and embed logical rules	Use prior knowledge to improve the accuracy of modeling and reduce data dependence	Strong domain constraint scenarios such as medical and financial (e.g., disease causality path modeling)	Relying on expert experience, the cost of knowledge acquisition is high

3.2 Causal Analysis Methods

The causal-consequential analysis method can be applied to the causal-consequential relationship shown in the causal model to extract valuable information that can be interpreted. The key causal analysis method includes the assessment of causality used to assess the intensity and direct the causal impact of one variable on another variable. This method makes it possible to determine the importance of each input variable in the decision-making process in the artificial intelligence model. For example, in a regression model, we can estimate the cause-and-effect impact of each characteristic on the dependent variable and interpret the results of the Model prediction.

Counterfactual analysis refers to another important method of cause-and-effect analysis. This type of analysis involves asking questions such as "If the input variable changes, what will be the result of the artificial intelligence model?" The answers to these questions can apply counterfactual explanations to model decision making, which helps users understand the sensitivity of the model to different input variables.

3.3 Causally-driven Explanation Generation Methods

The method of generating cause-and-effect interpretation is based on the results of cause-and-effect modeling and analysis to create explanations in natural language that are understandable to humans. Complex cause-and-effect links and statistical information must be converted into brief explanatory content.

Conventional methods use cause-and-effect factors defined by cause-and-effect models to generate explanations. For example, if a causal model shows that a particular medical test result is a key causal factor in the diagnosis of a disease, the method of generating an interpretation can use natural language to display information and explain to the

user that the diagnosis of the disease mainly follows from this test result, as well as to clarify the causality. the mechanism behind it.

Explanations can also be created using pattern-based methods. The pattern is developed according to the architecture of the cause-and-effect model, and then the corresponding cause-and-effect information is placed in it to build an interpretation in natural language. This method can not only provide consistency of explanation, but can also make it more readable. The results obtained by such operations have their unique value in the category of cause-and-effect interpretation.

4. Application and Practice of Causal Reasoning-based AI Explainability Methods

4.1 Analysis of Typical Application Scenarios

4.1.1 Healthcare

In medicine, artificial intelligence technologies are gradually being introduced for the diagnosis of diseases, the preparation of treatment plans, and the preparation of forecasts. In this regard, it is extremely important to ensure the interpretability of artificial intelligence in the medical system. Doctors and patients must have sufficient insight into the causes of decisions made by artificial intelligence to ensure the safety and effectiveness of the treatment process. For example, in the medical diagnostic system of cancer using imaging, causal relationships can be used to identify key image characteristics that are causally associated with a diagnosis and provide an interpretation of why these characteristics are critical [2].

4.1.2 Finance

The application of artificial intelligence in the financial sphere includes credit rating, risk assessment, and investment decision-making. In order to meet regulatory requirements and gain the trust of customers, the interpretability of the artificial intelligence system is a prerequisite. For example, the credit scoring model analyzes cause-and-effect factors that influence the assessment, such as the applicant's income, debt balance, and credit history, as well as explanations of how these factors determine the credit score. As shown in the following table:

Table 4. Summary of the application of causal reasoning methods in typical application scenarios

Domain	AI assignment	Causal inference method	Explain the requirements	Case studies
Medical treatment	Disease diagnosis/prognosis	Causal graph identifies key image features and counterfactual analysis of the effects of treatment plans	Doctors / Patients need to address the rationale for the decision (e.g., "Why is a certain indicator a key factor in cancer diagnosis?")	In the risk prediction system for cardiovascular disease, the direct causal effects of blood pressure and cholesterol were determined by causal effect estimation
Finance	Credit scoring/risk assessment	Cause analysis of the impact of income, liabilities, and other factors on credit	Compliance requirements (such as EU GDPR interpretative provisions), and user trust building	In the credit scoring model, the causal path between "income level" and "credit rating" is explained to eliminate the interference of irrelevant variables, such as gender

4.1.3 Enterprise Management

In the field of enterprise management, artificial intelligence is increasingly applied to core scenarios such as strategic decision-making, human resource management, and supply chain optimization. In this context, ensuring the explainability of AI systems is crucial for improving management efficiency, reducing decision-making risks, and enhancing

trust between management and employees. Enterprises need to clarify the logic behind AI decisions to avoid management misjudgments or resource waste caused by the "black-box" nature of AI.

4.2 Case Study of Practical Applications

To test the effectiveness of the proposed method, a clinical trial was conducted in the field of Medicine. We have developed an AI-based system that uses patient data such as demographic information, medical history, and laboratory test results to predict cardiovascular disease risk.

First, using a data-based cause-and-effect relationship detection algorithm, a cause-and-effect map is made based on the patient's data. This figure shows that variables such as age, blood pressure, and cholesterol levels are direct risk factors for cardiovascular disease, while other variables such as gender and smoking are indirectly related to these direct factors.

Moreover, the importance of each causal factor is determined by the assessment of cause-and-effect relationships. The results showed that blood pressure has the most prominent impact on the risk of heart disease, followed by cholesterol levels and age.

Finally, a pattern-based interpretation generation method is used that provides a natural language interpretation to predict the model. For specific patients, this explanation shows that the high risk of cardiovascular disease is due to hypertension factors. After all, hypertension has a significant cause-and-effect relationship with heart disease. At the same time, this also overlaps with relatively high cholesterol caused by a combination of various factors, such as age.

4.3 Application Challenges and Coping Strategy

Although the application of cause-and-effect methods to improve the interpretability of artificial intelligence has yielded certain results, it still faces many challenges and difficulties. Especially with the increase in the amount of data, causal modeling becomes increasingly complex to use. Algorithms may not be able to accurately construct a suitable cause-and-effect graph structure in complex input data sets, and at the same time, high computational costs have become an inevitable obstacle [3-5].

In addition, in some areas, there is a lack of prior knowledge. In these areas, which lack a priori knowledge, it is quite difficult to construct an exact cause-and-effect model based on knowledge and is limited by data processing conditions. Applying a data-based approach can lead to multiple false conclusions.

Several strategies can be adopted to address these problems. Given the complexity of cause-and-effect modeling, efficient algorithms for detecting cause-and-effect relationships and the introduction of parallel computing technology to reduce computational costs are expected to be developed. Due to a lack of prior knowledge, a combination of data-based and knowledge-based methods, limited prior knowledge can guide the causal modeling process, thus increasing the accuracy of cause-and-effect models [6-8].

5. Conclusion

5.1 Research Conclusions

This article systematically explores methods for improving the interpretability of artificial intelligence through cause-and-effect reasoning and establishes a theoretical link between the definition of the concept of association and cause-and-effect reasoning and interpretability. Three main methods are proposed: causal modeling, causal analysis, and generating interpretability based on causal relationships, and their specific methods of application and application are discussed in detail. Case studies in typical application scenarios show that the effectiveness and practicality of these methods are well reflected.

The study showed that integrating cause-and-effect connections into an artificial intelligence system can significantly improve its interpretability. Causal modeling helps to understand the causal relationship underlying artificial intelligence decision-making. Causal analysis can quantify the causal effects of various variables, and a method of generating interpretations based on generating causal relationships allows people to get that information in a more understandable way.

5.2 Research Limitations

However, this study also has a number of limitations. First of all, the causal link detection algorithm used may not be able to accurately infer the causal link graph in all data processing scenarios, especially in the presence of unnoticed

variables that may lead to confusion. Secondly, the way to explain generation is still relatively simple and it can be difficult to solve very complex cause-and-effect relationships. Moreover, case studies are carried out in certain areas, and their versatility in other areas should be further checked.

5.3 Future Research Prospects

In future studies, we will try to eliminate existing limitations and advance in-depth research. The focus will be on developing more robust causal detection algorithms. In this process, unnoticeable confusing variables and multidimensional data must be taken into account. Synchronous will also be promoted to improve the method of generating interpretations in order to increase its flexibility and obtain more complex explanations. In addition, we face the challenge of popularizing the method in various fields, at the same time conducting large empirical research as a key link to testing the universality of the method.

In short, research into methods for improving the interpretability of artificial intelligence by cause-and-effect relationships is very important and has broad perspectives. Continuous research and innovation in this area enable to achieve the goals of transparency, comprehensibility and trust in artificial intelligence systems, meet the needs to promote healthy development of technology and promote its deeper integration with the real needs of real applications.

References

- [1] Zhang Z. Research on the prediction method of judicial judgments based on causal analysis [dissertation]. Shanxi: Shanxi University; 2023.
- [2] Cai Y. From Bayesian inference to causal inference: a study on Bayesian classification models based on causal networks [dissertation]. Jilin: Jilin University; 2024.
- [3] Li M. Research on the knowledge expression of traditional Chinese medicine practitioners' differential diagnosis and treatment of chronic heart failure based on dynamic uncertain causal graph [dissertation]. Shandong: Shandong University of Traditional Chinese Medicine; 2023.
- [4] Guo Z. A study on interpretative auxiliary diagnosis and causal effect estimation methods [dissertation]. Beijing: Beijing Jiaotong University; 2022.
- [5] Sun Y. Research on transfer reinforcement learning control method based on causal modeling [dissertation]. Jiangsu: Southeast University; 2023.
- [6] Bao C. Lightweight cervical TCT pathological image cell detection model and interpretability study [dissertation]. Shanghai: Donghua University; 2023.
- [7] Moraes DAI, Arrighi L, Junior BS, et al. Explainable artificial intelligence (xAI) applied to deep computer vision of microscopy imaging and spectroscopy for assessment of oleogel stability over storage. *J Food Eng.* 2025;394:112515.
- [8] Kruk M. SHAP-NET, a network based on Shapley values as a new tool to improve the explainability of the XGBoost-SHAP model for the problem of water quality. *Environ Model Softw.* 2025;188:106403.