International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

A Concise Review of Causal and Counterfactual Explanations in Explainable Artificial Intelligence

Shruti A. Gadewar[1], Dr. Salim Y. Amdani[2]
[1]*PhD Scholar, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India*
[2]*Asssociate Professor, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India*

DOI: 10.5281/zenodo.19252680

ABSTRACT

Explainable Artificial Intelligence (XAI) has become a critical requirement for deploying machine learning models in high-stakes domains. While popular post-hoc explanation techniques primarily rely on correlation-based feature attribution, they often fail to provide faithful insights into the underlying decision-making process of complex models. To address these limitations, recent research has increasingly focused on causal and counterfactual explanations, which aim to reveal cause–effect relationships and minimal input changes that alter model predictions, respectively. This paper presents a compact review of causal and counterfactual explanation methods, with an emphasis on their theoretical foundations, methodological differences, and applicability to modern machine learning systems, including natural language processing tasks. We provide a structured taxonomy of existing approaches, highlight representative methods, and offer a comparative analysis of causal and counterfactual explanations in terms of interpretability, faithfulness, and practical usability. Furthermore, we discuss key challenges such as lack of causal grounding, evaluation complexity, and scalability, which remain open research problems. This review aims to serve as a concise reference for researchers and practitioners seeking to understand and advance causally grounded explainability techniques.

Keywords: - Explainable Artificial Intelligence, Causal Inference, Counterfactual Explanation, Interpretability, Machine Learning, Natural Language Processing.

1. INTRODUCTION

Recent advances in machine learning (ML) and deep learning (DL) have led to the widespread adoption of complex predictive models across diverse application domains such as healthcare, finance, legal decision-making, and natural language processing (NLP). Despite their strong predictive performance, these models are often criticized for their *black-box* nature, which limits transparency, trust, and accountability. As a result, Explainable Artificial Intelligence (XAI) has emerged as a critical research area aimed at making model decisions understandable to humans [1], [2].

Early XAI techniques primarily focused on correlation-based feature attribution, including post-hoc methods such as LIME and SHAP, which explain predictions by estimating the influence of input features on model outputs [3], [4]. Although widely adopted, such methods often fail to capture the true decision-making process of models, as they rely on statistical associations rather than cause–effect relationships. This limitation becomes particularly problematic in high-stakes scenarios, where misleading explanations may result in incorrect or unsafe decisions [5].

To overcome these shortcomings, recent research has increasingly emphasized the role of causal reasoning in explainability. Causal explanations aim to identify how changes in one variable directly influence an outcome by modeling underlying data-generating mechanisms, often using Structural Causal Models (SCMs) and intervention-based reasoning [6], [7]. By moving beyond correlation, causal explanations provide more faithful and robust insights into model behavior, especially under distribution shifts and unseen conditions.

In parallel, counterfactual explanations have gained significant attention as an intuitive and human-centered explanation paradigm. Counterfactual explanations answer *"what-if"* questions by identifying the minimal changes to input features that would alter a model's prediction [8]. Due to their contrastive nature, counterfactuals are particularly appealing in decision-support systems and NLP applications such as sentiment analysis, fairness evaluation, and bias detection [9], [10]. However, many counterfactual methods lack explicit causal grounding, leading to unrealistic or spurious explanations that may not reflect plausible real-world scenarios [11].

Despite rapid progress, the literature on causal and counterfactual explanations remains fragmented, with limited efforts to systematically compare these approaches, particularly under practical constraints such as interpretability, faithfulness, and applicability to modern ML and NLP models. Moreover, the relationship between causal and counterfactual explanations—whether complementary or competing—remains an open research question.

Motivated by these challenges, this paper presents a concise review of causal and counterfactual explanation methods within the context of XAI. The main contributions of this review are threefold:

1. it provides a compact overview of foundational concepts and representative methods;
2. it offers a comparative analysis of causal and counterfactual explanations in terms of interpretability and causal validity; and
3. it highlights key open challenges that motivate future research toward causally grounded explainability frameworks.

2. BACKGROUND

Explainable Artificial Intelligence (XAI) aims to improve the transparency and interpretability of machine learning models, particularly those that operate as black boxes [1], [2]. Existing XAI techniques are commonly categorized as intrinsic or post-hoc, and as local or global explanations, depending on whether interpretability is built into the model or generated after training, and whether explanations target individual predictions or overall model behaviour [3], [4], [5]. Although post-hoc, local explanation methods such as LIME and SHAP are widely used due to their model-agnostic nature, they primarily rely on statistical associations and therefore lack causal faithfulness [3],[5].

Causal reasoning provides a principled framework for moving beyond correlation by explicitly modeling cause–effect relationships through Structural Causal Models (SCMs) and intervention-based analysis [6]. By enabling reasoning about how changes in one variable influence outcomes, causal approaches offer more robust and faithful explanations, particularly under distribution shifts and confounding effects [7], [10]. In NLP applications, causal methods have been employed to disentangle factors such as sentiment, topic, and author attributes, thereby improving interpretability and robustness [7]. However, causal explanation methods often depend on strong assumptions about the underlying data-generating process and require access to causal structures or interventional data, which limits their practical applicability.

Counterfactual explanations provide an alternative, human-centered explanation paradigm by answering contrastive “what-if” questions through minimal input modifications that change model predictions [8]. Due to their intuitive nature, counterfactuals have been widely adopted in applications such as sentiment analysis, fairness evaluation, and bias detection [9], [11]. Nonetheless, many counterfactual generation methods optimize proximity or sparsity objectives without enforcing causal constraints, leading to implausible or causally invalid explanations [11]. This limitation has motivated recent efforts to integrate causal reasoning into counterfactual explanation frameworks, highlighting the complementary roles of causality and counterfactual reasoning in explainable AI [10], [12].

3.CAUSAL EXPLANATION METHODS

Causal explanation methods aim to provide faithful interpretations of model predictions by explicitly modeling cause–effect relationships rather than relying on observed correlations. Most causal explanation approaches are grounded in the Structural Causal Model (SCM) framework, which represents causal dependencies among variables using directed graphs and structural equations, enabling reasoning through interventions and counterfactual queries [6]. In contrast to traditional post-hoc explanations, causal methods seek to answer *why* a prediction occurred by identifying the variables that directly influence the outcome under hypothetical interventions.

One prominent line of work focuses on graph-based causal modeling, where domain knowledge or learned causal structures are used to explain predictions through intervention analysis. By simulating interventions on input variables, these methods can identify causal drivers of model behaviour and assess robustness under distributional changes [6], [13]. Such approaches have been shown to reduce the impact of confounding variables and improve explanation faithfulness, particularly in settings where spurious correlations are prevalent.

Another category of causal explanation methods involves causal representation learning, which aims to disentangle latent factors corresponding to causal variables within learned feature representations. By enforcing independence or invariance constraints across environments, these methods attempt to isolate causal features that remain stable under interventions [7], [14]. In NLP tasks, causal representation learning has been applied to separate sentiment-related information from confounding attributes such as topic or author style, leading to more reliable explanations [7], [10].

Despite their conceptual advantages, causal explanation methods face several practical challenges. Learning accurate causal structures from observational data is inherently difficult and often requires strong assumptions or external supervision [6], [15]. Moreover, causal explanations may be less intuitive for end users compared to contrastive explanations, limiting their usability in human-centered decision-support systems. These limitations motivate complementary approaches that emphasize interpretability and accessibility, such as counterfactual explanations, which are discussed in the next section.

4.COUNTERFACTUAL EXPLANATION METHODS

Counterfactual explanation methods aim to explain model predictions by identifying minimal and meaningful changes to input features that would result in a different model outcome. Formally, a counterfactual explanation answers the question: "What needs to change for the prediction to be different?" while keeping other factors constant [8]. Due to their contrastive and intuitive nature, counterfactual explanations have gained popularity in explainable AI, particularly in decision-support systems and human-centered applications.

Most counterfactual methods are formulated as optimization problems, where the objective is to minimize the distance between the original instance and a modified instance subject to a change in the model's prediction [8], [16]. Common optimization constraints include proximity, sparsity, and plausibility, which aim to ensure that generated counterfactuals are interpretable and actionable. These methods are typically model-agnostic and can be applied to a wide range of classifiers; however, they often depend heavily on heuristic distance metrics that may not align with real-world semantics.

In the context of natural language processing, counterfactual explanations are commonly generated through token-level or phrase-level substitutions, controlled text generation, or counterfactually augmented data [9], [11]. Such approaches have been applied to tasks including sentiment analysis, bias detection, and fairness evaluation, where small linguistic changes can significantly alter model predictions [9]. While effective in highlighting decision-sensitive features, textual counterfactuals may introduce grammatical inconsistencies or semantic drift, reducing explanation plausibility.

A key limitation of many counterfactual explanation methods is their lack of explicit causal grounding. Without modeling causal dependencies among features, generated counterfactuals may violate real-world constraints or represent changes that are not causally feasible [11]. Recent studies have emphasized the need to integrate causal reasoning into counterfactual generation to ensure validity and robustness, particularly in complex domains such as NLP [10], [17]. These observations highlight the complementary relationship between counterfactual and causal explanations and motivate hybrid approaches that combine interpretability with causal faithfulness.

5.COMPARATIVE ANALYSIS AND OPEN CHALLENGES

Causal and counterfactual explanations represent two complementary paradigms in explainable artificial intelligence, each addressing different aspects of model interpretability. Causal explanations aim to uncover underlying cause–effect mechanisms governing model behaviour, thereby offering faithful and robust interpretations grounded in intervention-based reasoning [6], [7]. In contrast, counterfactual explanations emphasize contrastive reasoning by identifying minimal changes to inputs that alter model predictions, making them particularly intuitive and actionable for end users [8], [16].

While causal explanations provide stronger guarantees of faithfulness, they often require explicit causal models, strong assumptions, or interventional data, which limits their scalability and applicability in real-world machine learning systems [6], [15]. Counterfactual explanations, on the other hand, are generally model-agnostic and easier to generate but may suffer from implausibility and lack of causal validity when causal constraints are not enforced [11], [17]. This trade-off highlights the need for systematic comparison and motivates hybrid approaches that combine causal reasoning with counterfactual interpretability.

Dimension	Causal Explanations	Counterfactual Explanations
Primary objective	Identify cause–effect relationships	Identify minimal decision-changing changes
Theoretical grounding	Strong (causal inference theory)	Moderate (optimization-based)
Faithfulness	High	Medium
Human interpretability	Medium	High
Causal validity	Explicitly enforced	Often implicit or absent
Scalability	Limited	High
NLP suitability	Moderate	High

Table 1: Comparison of Causal and Counterfactual Explanation

Despite significant progress, several open challenges remain. First, lack of causal grounding in many counterfactual methods leads to explanations that are not feasible or meaningful in real-world settings [11].

Second, evaluation of explanation quality remains an open problem, as existing metrics often fail to capture causal validity, plausibility, and human interpretability simultaneously [18]. Third, both causal and counterfactual methods struggle with scalability and deployment in large-scale models, including modern deep learning and language models. Finally, the absence of standardized benchmarks hinders fair comparison across explanation approaches.

Addressing these challenges requires the development of causally informed counterfactual frameworks, improved evaluation protocols, and domain-specific causal modeling strategies, particularly for complex applications such as natural language processing.

6.CONCLUSION

This paper presented a concise review of causal and counterfactual explanation methods in explainable artificial intelligence, highlighting their theoretical foundations, practical strengths, and inherent limitations. While causal explanations offer strong faithfulness by explicitly modeling cause–effect relationships, their applicability is often constrained by the need for causal knowledge and strong assumptions. In contrast, counterfactual explanations provide intuitive and actionable insights but frequently lack causal validity, leading to implausible or misleading interpretations. Through a comparative analysis, this review emphasized the complementary nature of causal and counterfactual explanations and identified key open challenges related to causal grounding, evaluation, and scalability. Overall, the findings underscore the importance of integrating causal reasoning into counterfactual frameworks to advance reliable and human-centered explainable AI systems.

7. REFERENCES

[1]. A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[2]. D. Gunning and D. Aha, “DARPA’s explainable artificial intelligence (XAI) program,” *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
[3]. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
[4]. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proc. NeurIPS*, 2017, pp. 4765–4774.
[5]. Z. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, no. 10, pp. 36–43, 2018.
[6]. J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[7]. V. Veitch, D. Sridhar, and D. Blei, “Adapting text embeddings for causal inference,” in *Proc. UAI*, 2020, pp. 919–928.
[8]. S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box,” *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
[9]. T. Kaushik, D. Hovy, and Z. Lipton, “Learning the difference that makes a difference with counterfactually-augmented data,” in *Proc. ICLR*, 2020.
[10]. A. Feder *et al.*, “Causal inference in natural language processing: Estimation, prediction, interpretation and beyond,” *Transactions of the ACL*, vol. 9, pp. 1138–1158, 2021.
[11]. R. Pryzant *et al.*, “Automatically neutralizing subjective bias in text,” in *Proc. AAAI*, 2020, pp. 480–489.
[12]. C. Molnar, *Interpretable Machine Learning*, 2nd ed. Munich, Germany: Leanpub, 2022.
[13]. E. Bareinboim and J. Pearl, “Causal inference and the data-fusion problem,” *Proc. Nat. Acad. Sci.*, vol. 113, no. 27, pp. 7345–7352, 2016.
[14]. A. Schölkopf *et al.*, “Toward causal representation learning,” *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
[15]. P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.
[16]. R. Karimi, G. Barthe, B. Balle, and I. Valera, “Model-agnostic counterfactual explanations for consequential decisions,” in *Proc. AISTATS*, 2020, pp. 895–905.
[17]. D. Kusner, J. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Proc. NeurIPS*, 2017, pp. 4066–4076.
[18]. F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.