Causality and Trustworthy AI: A Survey

www.surveyx.cn

Abstract

This survey explores the intersection of causality, causal representation learning, graph neural networks (GNNs), and trustworthy AI, highlighting the importance of understanding and modeling cause-and-effect relationships in AI systems. By integrating causal reasoning, AI models can transcend mere correlation, enhancing interpretability and decision-making across domains like healthcare and finance. Recent advancements in causal representation learning and GNNs have improved the robustness and transparency of AI, exemplified by frameworks that leverage causal insights to enhance model performance and interpretability. Despite these advancements, challenges remain in accurately modeling causal relationships, particularly in high-dimensional data environments. The survey underscores the necessity of systematic evaluation processes to advance causal learning research, emphasizing the critical role of causality in developing trustworthy and explainable AI systems. By addressing existing challenges and leveraging causal frameworks, AI systems can achieve greater transparency, interpretability, and ethical alignment, enhancing their applicability across diverse fields.

1 Introduction

1.1 Multidisciplinary Nature of Causality in AI

Causality in AI is a multidisciplinary endeavor, integrating insights from computer science, cognitive science, and statistics to elucidate complex cause-and-effect relationships within intelligent systems [1]. This integration is vital for creating AI systems that are not only effective and interpretable but also aligned with human reasoning processes. A solid grasp of causality is essential for human problem-solving and decision-making, underscoring the necessity of embedding causal reasoning into AI to improve its decision-making capabilities [2].

The growing emphasis on causality in the machine learning community highlights its potential to enhance model generalization and robustness, addressing knowledge gaps across various sectors [1]. For example, in recommendation systems, causal insights can transform accurate predictions into actionable, explainable decisions, thereby boosting user engagement and system reliability [3].

Frameworks like ECHO, which prioritize human-centric social interactions, further illustrate the necessity for enhanced reasoning capabilities in AI systems [4]. In navigation tasks within Embodied AI, a nuanced understanding of causality is critical for effectively modeling task-specific characteristics [5].

The interplay of causality and neural network repair exemplifies the interdisciplinary nature of AI, demonstrating how causal insights can bolster the reliability and accuracy of AI systems [6]. Additionally, the advancement of frameworks for modeling input-dependent, dynamic causal orders showcases the extensive reach of causality in AI, opening new avenues for research and application [7].

The integration of causality into AI not only propels the field forward but also enhances the reliability, interpretability, and alignment of AI systems with scientific principles and societal needs. This integration addresses the limitations of existing AI models, which often struggle with understanding

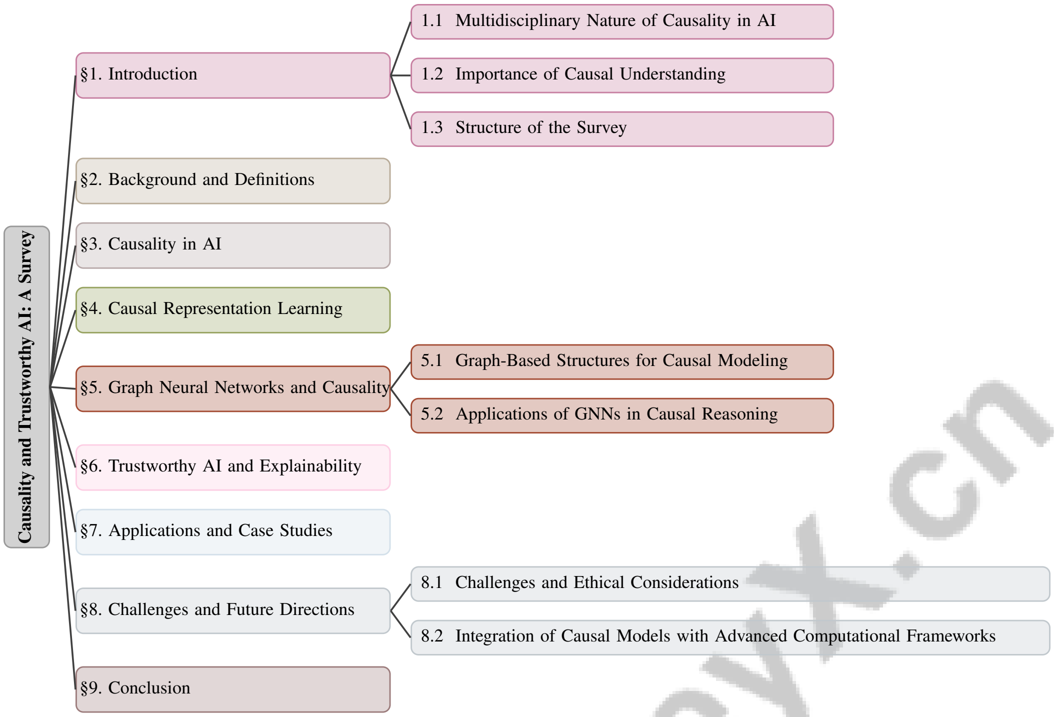

Causality and Trustworthy AI: A Survey
§1. Introduction
1.1 Multidisciplinary Nature of Causality in AI
1.2 Importance of Causal Understanding
1.3 Structure of the Survey
§2. Background and Definitions
§3. Causality in AI
§4. Causal Representation Learning
§5. Graph Neural Networks and Causality
5.1 Graph-Based Structures for Causal Modeling
5.2 Applications of GNNs in Causal Reasoning
§6. Trustworthy AI and Explainability
§7. Applications and Case Studies
§8. Challenges and Future Directions
8.1 Challenges and Ethical Considerations
8.2 Integration of Causal Models with Advanced Computational Frameworks
§9. Conclusion

Figure 1: chapter structure

cause-effect relationships, leading to poor generalization, biased outcomes, and interpretability challenges. By employing causal modeling and inference methods, researchers are crafting more trustworthy AI systems capable of providing clearer insights into their decision-making processes. Moreover, the focus on causality is paving the way for interpretable machine learning, facilitating a deeper understanding of model mechanisms beyond mere associations. The incorporation of causality into AI is anticipated to drive significant advancements, fostering scientific discovery and real-world applications across diverse domains [8, 9, 10].

1.2 Importance of Causal Understanding

Causal understanding is crucial for AI systems, enabling counterfactual reasoning and inference that address limitations in current models that overlook causal relationships [11]. This capability is vital for enhancing the interpretability of deep learning models, which are frequently perceived as opaque black boxes [12]. Distinguishing between correlation and causation remains a core challenge in scientific research, making its resolution essential for developing trustworthy AI systems [13].

In finance, establishing cause-and-effect relationships, rather than mere correlations, is vital for improving the trustworthiness and accuracy of predictions critical for decision-making. Similarly, in healthcare analytics, a robust understanding of causality is imperative for effective decision-making, ensuring AI systems yield reliable insights into patient care [14].

Causal understanding also significantly contributes to addressing fairness challenges, particularly in classification tasks where sensitive attributes may lead to biased outcomes [15]. This is essential for preventing unfair discrimination in algorithmic decision-making and ensuring AI systems do not perpetuate biases based on gender, race, or religion [16].

In the context of Graph Neural Networks (GNNs), understanding causality is pivotal for mitigating trustworthiness risks, including poor out-of-distribution generalizability, unfair representations, and lack of explainability [17]. Furthermore, in reinforcement learning frameworks, causality enhances sample efficiency and interpretability, thereby improving decision-making processes [18].

Integrating causal reasoning into AI systems facilitates systematic defect addressing, leading to advancements in reliability, fairness, and safety. This alignment with human reasoning and societal expectations ensures that AI systems accurately reflect relationships between variables and do not

AI-generated, for reference only.

perpetuate biases [19]. Understanding causality is essential for enhancing transparency, reliability, and ethical alignment in AI systems, especially in domains like biometric systems where interpretability challenges are pronounced [20].

The lack of interpretability in control policies learned by imitation learning agents, particularly those modeled as deep neural networks, further emphasizes the need for causal understanding [21]. Additionally, causality is critical for resolving misidentification challenges linked to invisible confounders [22]. In multi-agent systems, a solid grasp of causality enhances reasoning faithfulness and reduces inference errors in knowledge-based reasoning tasks [23].

Causality is also vital for evaluating causal reasoning abilities in large language models (LLMs) through structured benchmarks [24]. In human-computer interaction, understanding event causality inference through human-centric reasoning is crucial for effective communication [4]. Incorporating causality in machine learning models can enhance their robustness and generalization capabilities [1]. Moreover, it is essential for generating actionable counterfactual explanations that convert undesired outcomes into desired ones [25]. Understanding harm in autonomous systems is vital for ensuring accountability and addressing adverse outcomes [26]. In decision-making contexts involving moral responsibility and ethical norms, understanding causality is indispensable [27].

1.3 Structure of the Survey

This survey is systematically organized to provide a comprehensive understanding of the intersection between causality and artificial intelligence. The initial sections establish the groundwork by introducing the multidisciplinary nature of causality in AI, emphasizing its significance for developing reliable and interpretable systems. The Introduction highlights the importance of understanding cause-and-effect relationships in AI, followed by a discussion on multidisciplinary aspects and the necessity of causal understanding.

The Background and Definitions section defines key concepts such as causality, causal inference, causal representation learning, graph neural networks, trustworthy AI, and explainability. This foundational knowledge is crucial for appreciating subsequent discussions on the role of causality in AI.

The survey then explores the application of causality in AI, focusing on its role in enhancing decision-making and interpretability in machine learning models. An examination of challenges and techniques for applying causality in high-dimensional and complex data scenarios reveals significant insights into the practical difficulties faced by researchers, including the limitations of traditional correlation analyses, the complexities of multivariate regression models, and the need for robust methodologies that can accommodate various observed factors. This exploration encompasses recent advancements in causal inference methods, including non-parametric approaches applicable to financial and social media data, as well as innovative uses of text classifiers for causal analysis, highlighting the evolving landscape of causal interpretability in machine learning [28, 29, 2, 10].

Causal Representation Learning is examined in detail, focusing on innovative techniques and the tools and algorithms for causal discovery and inference. This section underscores cutting-edge advancements in learning causal representations and their implications for AI systems.

The integration of Graph Neural Networks with causal models is discussed, showcasing how graph-based structures can effectively model complex causal relationships. The incorporation of causal learning techniques into GNNs has led to significant advancements in their application for causal reasoning, as evidenced by studies demonstrating how these models can address trustworthiness issues such as susceptibility to distribution shifts and biases. These examples illustrate the practical utility of GNNs in understanding complex causal relationships within graph-structured data, thereby enhancing their reliability and interpretability across various real-world scenarios [12, 17].

The importance of trustworthiness and explainability in AI systems is analyzed, focusing on how causal reasoning can enhance explainability and provide actionable insights. This section emphasizes the ethical and practical implications of integrating causality into AI.

Applications and case studies illustrate practical implementations of causality, causal representation learning, and GNNs in addressing intricate challenges. The integration of causal learning techniques into GNNs enhances their trustworthiness by tackling issues such as susceptibility to distribution shifts and biases, while also improving their explainability. Recent research has shown that causal analysis

AI-generated, for reference only.

can significantly bolster GNN performance in various tasks, including classification and prediction, by revealing underlying causal mechanisms rather than relying solely on superficial correlations [17, 30]. These case studies demonstrate the tangible benefits and challenges of implementing causal models across diverse domains.

The survey identifies current challenges in integrating causal models with advanced computational frameworks, highlighting critical ethical considerations and proposing future research directions aimed at enhancing fairness and interpretability in machine learning algorithms. It emphasizes the necessity of causal-based fairness notions, the importance of addressing interpretability in machine learning, and the potential of causal inference to improve the robustness of natural language processing models, thereby providing a roadmap for researchers in these interconnected fields [31, 32, 33, 10]. The Conclusion reiterates the key points discussed, emphasizing the significance of causality in developing trustworthy and explainable AI systems.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Core Concepts of Causality and Causal Inference

Understanding causality and causal inference is fundamental for distinguishing correlation from causation, crucial for developing AI systems capable of making reliable decisions across domains. The Pearl Causal Hierarchy (PCH) provides a structured framework that includes associational, interventional, and counterfactual levels, enhancing AI models' ability to reason about interventions and counterfactuals, thereby improving interpretability and reliability [34]. This is particularly relevant in healthcare, where causal relationships must be established from observational data due to the impracticality of randomized controlled trials [27].

Causal inference methodologies aim to estimate intervention effects from observational data, often complicated by confounding variables. In multivariate time series, the challenge is to infer a sparse causality graph amidst dense connections while managing confounding effects [35]. Identifying causal relationships from multivariate functional data, which can be noisy and non-Gaussian, presents additional challenges [36]. In dynamic environments, causal reinforcement learning frameworks like Markov Decision Processes (MDPs) and Partially Observed Markov Decision Processes (POMDPs) enhance decision-making under uncertainty [5]. Integrating causal reasoning into multi-agent reinforcement learning addresses complexities arising from agent interactions, although accurately inferring a directed acyclic graph (DAG) from limited data remains a significant challenge.

In natural language processing, detecting causal relationships within text enhances understanding of event sequences and implications, crucial in fields like healthcare, business risk management, and finance. This task is challenging due to implicit causality in text, requiring advanced methods for identification and interpretation. Language models like Causal BERT have improved capabilities in recognizing causal relationships, enabling the extraction of causal diagrams and event chains from unstructured text [32, 37]. In video analysis, predicting human actions based on causal relationships requires advanced causal reasoning techniques. Moreover, biometric systems often function as black boxes, complicating the understanding of their decision-making processes.

The core concepts of causality and causal inference are vital for enhancing AI systems' ability to model, interpret, and predict complex phenomena. Addressing these challenges requires integrated approaches that embed causal reasoning within AI frameworks, improving their capacity to provide reliable and actionable insights. The limitations of single-world interventionism in causal analysis, particularly regarding counterfactual variables, underscore the need for robust methodologies. Proposed methodologies offer more interpretable frameworks for causal inference compared to traditional regression and machine learning methods. Understanding the complex structure of causality beyond conventional models is crucial, especially in scenarios with indefinite and input-dependent causal relationships. Additionally, defining harm through causal models emphasizes the importance of causal relationships in understanding adverse outcomes [27].

2.2 Interconnections and Relevance to AI

Integrating causality into AI frameworks is crucial for enhancing interpretability, robustness, and ethical alignment across diverse applications. The interplay between causality and AI is rooted in the necessity to distinguish causal relationships from statistical associations, which is critical for reliable decision-making [34]. Traditional machine learning approaches often overlook the causal structure of data, focusing on correlations that can lead to inaccurate predictions [35]. This oversight is problematic when causal inference is hindered by the absence of experimental data, resulting in challenges like the third variable problem and selection bias.

In healthcare, causal inference methodologies are essential for mitigating confounding and selection biases, which can mislead causal conclusions and affect patient treatment [38, 39, 14, 40, 41]. Similarly, in atmospheric sciences, quantitative causality analysis is crucial for understanding climate predictability and its implications across disciplines. Granger causality exemplifies the application of causal inference in understanding temporal dependencies across various domains.

The challenge of recovering nonlinear relations in large-scale networks, compounded by the curse of dimensionality, underscores the need for advanced causal discovery methods. In reinforcement learning, integrating causal knowledge addresses limitations of traditional models, such as the requirement for substantial interaction data. Non-identifiability in causal graphs, coupled with the combinatorial complexity of DAGs, poses significant barriers to effective causal inference across fields [42, 43, 44, 45].

The ECHO dataset, integrating visual and textual information to evaluate reasoning capabilities, highlights the interconnections between visual data and causal inference, demonstrating the potential for AI systems to leverage multimodal data for enhanced reasoning. Confounding variables affecting social media sentiment and stock prices complicate the determination of true causality, emphasizing the necessity for robust methodologies in causal analysis [28, 46, 2, 47, 3].

The interconnections between causality and AI are essential for advancing the field, facilitating the creation of AI systems that offer transparent and interpretable insights. By integrating causal modeling and inference methods, researchers can enhance the trustworthiness of AI models, addressing challenges such as generalization to unseen data, fairness, and interpretability. This approach transcends traditional black-box machine learning techniques, emphasizing the importance of understanding cause-effect relationships to improve model performance and coherence, leading to more reliable AI applications across various domains [48, 49, 9, 10]. Addressing the challenges associated with causal inference is essential for enhancing the reliability and generalizability of AI applications across diverse fields.

In recent years, the exploration of causality within artificial intelligence (AI) has garnered significant attention, particularly in the context of machine learning. As researchers delve deeper into this domain, understanding the hierarchical structure of key concepts becomes essential. Figure 2 illustrates this structure, encompassing causal inference in machine learning and its application in high-dimensional data. The figure highlights not only the importance and applications of causal inference but also the methodologies employed, alongside the challenges, advanced models, and sensitivity issues that arise in complex data environments. This comprehensive overview provides a visual representation that complements the theoretical discussions presented in this paper, thereby enhancing our understanding of the intricate relationships within the field of AI.

3 Causality in AI

3.1 Causal Inference and Machine Learning

Causal inference is essential for advancing AI models from simple correlation analysis to understanding the causal mechanisms underlying data generation. This understanding is crucial for developing robust AI systems capable of accurately modeling interventions and addressing complex causal queries, which enhances decision-making and interpretability [35]. By emphasizing causal relationships, AI models provide more reliable insights, as seen in frameworks generating causally consistent counterfactuals.

Incorporating causal reasoning into machine learning significantly improves interpretability and predictive fairness, especially in sensitive domains like healthcare and recommender systems. Causality-aware techniques elucidate causal relations among state and action variables in dynamic environments [5]. Moreover, methods that explore causal interactions in complex datasets, such as EEG, offer innovative approaches for analyzing connectivity and understanding brain networks.

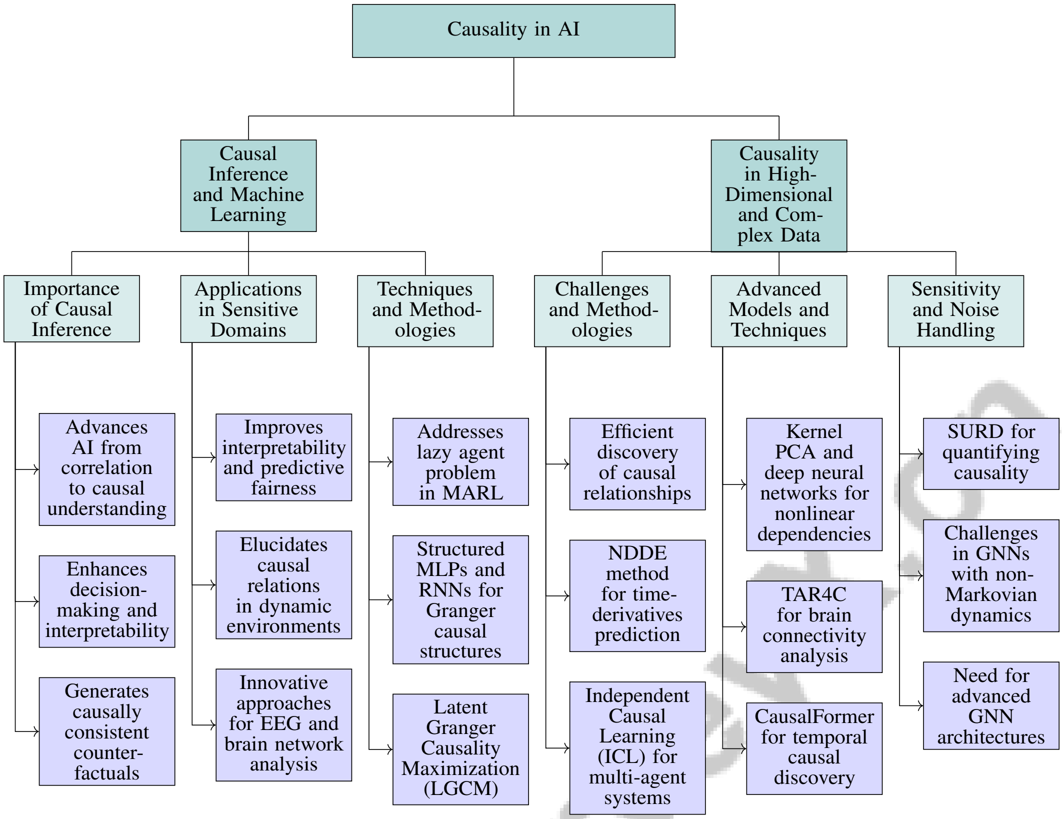

Causality in AI
Causal Inference and Machine Learning
Causality in High-Dimensional and Complex Data
Importance of Causal Inference
Applications in Sensitive Domains
Techniques and Methodologies
Challenges and Methodologies
Advanced Models and Techniques
Sensitivity and Noise Handling
Advances AI from correlation to causal understanding
Enhances decision-making and interpretability
Generates causally consistent counterfactuals
Improves interpretability and predictive fairness
Elucidates causal relations in dynamic environments
Innovative approaches for EEG and brain network analysis
Addresses lazy agent problem in MARL
Structured MLPs and RNNs for Granger causal structures
Latent Granger Causality Maximization (LGCM)
Efficient discovery of causal relationships
NDDE method for time-derivatives prediction
Independent Causal Learning (ICL) for multi-agent systems
Kernel PCA and deep neural networks for nonlinear dependencies
TAR4C for brain connectivity analysis
CausalFormer for temporal causal discovery
SURD for quantifying causality
Challenges in GNNs with non-Markovian dynamics
Need for advanced GNN architectures

Figure 2: This figure illustrates the hierarchical structure of key concepts in causality within AI, encompassing causal inference in machine learning and its application in high-dimensional data. It highlights the importance, applications, and methodologies in causal inference, as well as challenges, advanced models, and sensitivity issues in complex data environments.

In multi-agent systems, causal inference tackles issues like the lazy agent problem in Multi-Agent Reinforcement Learning (MARL) by using causal reasoning to enhance learning outcomes and agent interactions, fostering better cooperation strategies [50, 51, 18, 52]. Techniques utilizing structured multilayer perceptrons (MLPs) and recurrent neural networks (RNNs) with sparsity-inducing penalties to identify Granger causal structures exemplify the potential of causal inference in understanding temporal dependencies.

Advanced methodologies like Latent Granger Causality Maximization (LGCM) uncover latent sources in complex networks by maximizing Granger Causality between component pairs, enhancing the understanding of information flow and causal relationships. Sophisticated statistical methods that identify causal relationships while managing noise and testing criteria underscore the critical role of causal inference in bolstering AI systems' trustworthiness. These methods are vital for overcoming AI models' current limitations regarding generalization, fairness, and interpretability, often stemming from an inadequate grasp of cause-effect dynamics. By integrating causal modeling and inference techniques, researchers aim to create AI systems that not only perform better but also align with human understanding of reality, fostering reliability and ethical outcomes in AI applications [53, 9].

Causal inference enhances AI models by providing a systematic framework for understanding and modeling causal relationships, addressing critical limitations such as generalization to unseen data, biases, and interpretability challenges. By incorporating causal modeling techniques, AI systems improve decision-making capabilities and offer clearer insights across diverse fields, including natural language processing, machine learning, and social media analysis, ultimately promoting trustworthiness and robustness in AI applications [10, 2, 32, 9, 54]. Ongoing research focuses on accurately recovering causal graphs from observational data, particularly in the presence of latent confounders and selection bias, which is crucial for establishing reliable AI applications across various domains.

AI-generated, for reference only.

3.2 Causality in High-Dimensional and Complex Data

Applying causality in high-dimensional and complex data environments poses significant challenges, necessitating sophisticated methodologies. A primary challenge is efficiently discovering causal relationships in high-dimensional datasets, as the computational complexity of existing constraint-based algorithms can hinder progress. Traditional methods, such as Granger causality, often rely on assumptions of linearity and stationarity, which may not hold in many real-world scenarios, leading to potential misinterpretations [55].

Innovative approaches, such as the NDDE method, utilize multi-layer perceptrons to predict time-derivatives of variables based on their current and delayed values, capturing causal relationships not apparent in models considering only current states. Independent Causal Learning (ICL) enhances agents' understanding of individual actions' contributions to collective rewards in multi-agent systems, fostering cooperation and addressing challenges like the "lazy agent pathology." By leveraging causal relationships between observations and team rewards, ICL improves overall team performance in complex interactive environments, enhancing safety, interpretability, and robustness in decentralized setups [56, 50, 51, 18].

High-dimensional data complexities require advanced models that effectively identify and capture nonlinear dependencies. Recent methodologies leveraging techniques such as kernel principal component analysis and deep neural networks are crucial in fields like physics and climatology, where real-world observations often exhibit complex, nonlinear relationships. These models have demonstrated efficacy even in scenarios with limited temporal samples, addressing challenges posed by the curse of dimensionality in large-scale systems [57, 58, 59]. For instance, TAR4C provides a comprehensive analysis of brain connectivity by capturing nonlinear dynamics and phase transitions often missed by traditional methods.

CausalFormer, a transformer-based model designed for temporal causal discovery, effectively tackles the challenge of interpreting deep learning models in a causal context. Unlike traditional approaches focusing on local parameters, CausalFormer emphasizes understanding the global structure of the model, employing a causality-aware transformer that captures causal relationships through a multi-kernel causal convolution and a decomposition-based causality detector. This holistic approach enhances the identification of potential causal relations and facilitates constructing causal graphs, advancing the interpretability of deep learning models in temporal contexts [60, 61, 10].

As illustrated in Figure 3, a significant hurdle is the sensitivity of causal inference methods to noise and test criteria, as seen in trivariate Granger causality analyses. Innovative methodologies that project observational data into component spaces address challenges from nonlinear dependencies and hidden confounders. The Synergistic-Unique-Redundant Decomposition (SURD) quantifies causality by analyzing redundant, unique, and synergistic information contributions from past observations, applicable in computational and experimental contexts even with limited data [62, 10, 49, 42, 63].

In graph neural networks (GNNs), traditional models often struggle to capture non-Markovian dynamics from temporal edge ordering, leading to suboptimal performance in time-resolved tasks [64]. This limitation underscores the need for advanced GNN architectures accommodating temporal complexities in dynamic systems.

The application of causality in high-dimensional and complex data scenarios necessitates innovative techniques to navigate these intricacies. These initiatives enhance AI models' trustworthiness by integrating causal modeling and inference techniques, improving accuracy and reliability of insights while facilitating greater interpretability. This advancement is particularly significant in complex data environments, where understanding cause-effect relationships is essential for robust applications in fields such as Natural Language Processing and forecasting systems, contributing to the evolution of causal inference methodologies [10, 2, 32, 9, 54].

4 Causal Representation Learning

4.1 Innovative Techniques in Causal Representation Learning

Advancements in causal representation learning are pivotal for enhancing AI systems' ability to model and comprehend causal dynamics within complex environments. Reformulating matching design for time-series data has enabled causal inference without heavy reliance on parametric assumptions

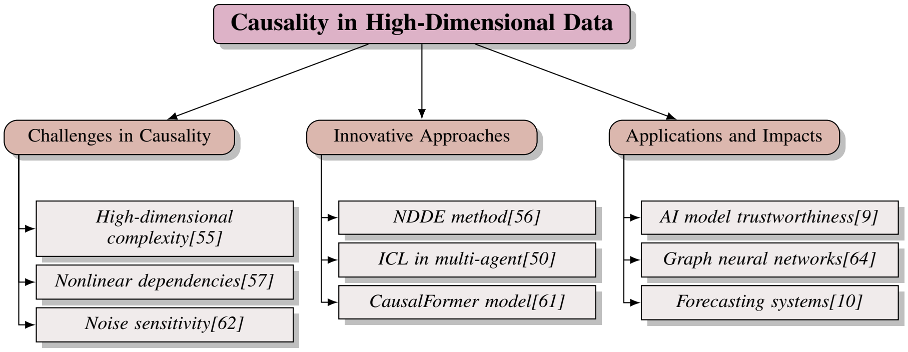


Figure 3: This figure illustrates the key challenges, innovative approaches, and applications of causality in high-dimensional and complex data environments. It highlights the intricacies of handling nonlinear dependencies and noise sensitivity, while showcasing advanced methodologies like NDDE and CausalFormer. The figure also emphasizes the impact of these causal techniques on AI model trustworthiness, graph neural networks, and forecasting systems.

Method Name	Causal Structures	Application Domains	Scalability and Efficiency
NCD[28]	Matching Design	Financial Analyses	Real-time Applications
CARE[6]	Fault Localization Approach	Neural Network Repair	Efficient And Effective
TAR4C[65]	Threshold Autoregressive Modeling	Brain Network Analysis	Computationally Intensive
CAT[5]	Causal Framework	Robotic Systems	Efficiently Update

Table 1: Overview of innovative methods in causal representation learning, highlighting their causal structures, application domains, and scalability. The table includes methods such as NCD, CARE, TAR4C, and CAT, demonstrating their diverse applications ranging from financial analyses to robotic systems, and their efficiency in various computational contexts.

or extensive model specifications, broadening its applicability across diverse scenarios [28]. The development of causally complete spaces expands the range of causal structures analyzed, enhancing the robustness and flexibility of causal inference techniques [7].

CARE highlights the causal contributions of neurons to system behavior, emphasizing the optimization of neural networks through targeted causal analysis [6]. In brain network analysis, TAR4C captures threshold dynamics of node interactions, offering a more accurate depiction of causal relationships [65]. This is supported by a theoretical framework utilizing Directed Transfer Function (DTF) and its variants for deriving confidence intervals and hypothesis testing [36].

A computational architecture for local belief propagation has been developed to enhance causal representation learning by allowing efficient updates without global adjustments, thus improving scalability and real-time application suitability [34]. Formalizing actual causality within action languages provides insights into ethical decision-making, demonstrating causality's role in evaluating AI systems' moral implications [27]. Moreover, the Causal Understanding Module in Causality-Aware Transformer (CAT) Networks introduces a novel approach to causal representation in robotic systems by focusing on direct causal relationships [5].

These techniques collectively advance AI by enhancing interpretability, reliability, and adaptability, addressing the limitations of traditional machine learning models that often overlook causal relationships. Prioritizing causal analysis not only improves the understanding of machine learning mechanisms but also fosters models capable of producing coherent outputs in complex real-world applications, such as related work generation [49, 10]. Leveraging these innovations enables AI to offer more precise, reliable, and ethically aligned insights, broadening their applicability across various fields. Table 1 provides a comprehensive summary of various innovative techniques in causal representation learning, illustrating their causal structures, application domains, and scalability.

4.2 Tools and Algorithms for Causal Discovery and Inference

The evolution of tools and algorithms for causal discovery and inference is crucial for strengthening AI systems' capacity to model causal relationships within complex datasets. The CLADDER dataset serves as a comprehensive resource for causal discovery, offering binary causal questions and answers to assess and enhance causal reasoning in AI models [24].

Figure 4 illustrates the categorization of tools and algorithms for causal discovery and inference, highlighting datasets, methods, and applications. As depicted, the CLADDER dataset is pivotal for causal reasoning evaluation, while methods such as Functional Bayesian Networks, Teleporter Theory, and RCL-OG showcase advancements in causal structure learning. Applications of these methodologies include legal judgment prediction and related work generation, underscoring the practical utility of these innovations.

Functional Bayesian Networks offer a sophisticated approach for inferring causal structures from multivariate functional data, utilizing a Bayesian framework that accommodates noise and non-Gaussianity [66]. The Teleporter Theory introduces a probabilistic causal model framework for complete graphical representation of counterfactuals, enhancing causal inference through counterfactual reasoning [67].

Reinforcement Causal Structure Learning (RCL-OG) innovatively integrates a reward mechanism within reinforcement learning to enhance the efficacy of causal structure search, surpassing traditional MCMC-based methods [68]. These tools and algorithms collectively represent significant progress in causal discovery and inference, providing robust frameworks for understanding and modeling causal relationships across various domains.

By employing innovative methodologies such as causal analysis and intervention, AI systems achieve improved accuracy and interpretability, enhancing robustness and generalizability across applications. These approaches enable models to differentiate between causal and non-causal information, addressing limitations in traditional machine learning techniques and fostering a deeper understanding of the mechanisms driving predictions [49, 69, 10].

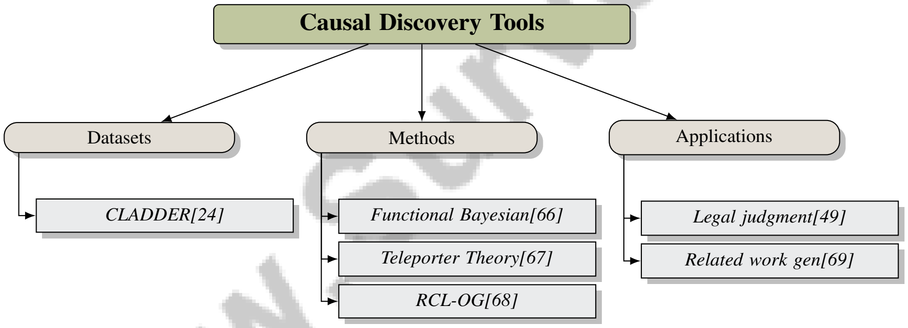


Figure 4: This figure illustrates the categorization of tools and algorithms for causal discovery and inference, highlighting datasets, methods, and applications. The CLADDER dataset is pivotal for causal reasoning evaluation, while methods such as Functional Bayesian Networks, Teleporter Theory, and RCL-OG showcase advancements in causal structure learning. Applications include legal judgment prediction and related work generation, underscoring the practical utility of these innovations.

5 Graph Neural Networks and Causality

The convergence of graph neural networks (GNNs) and causality is pivotal for elucidating causal modeling within graph-based structures. These frameworks adeptly represent and analyze causal relationships, capturing the complexities of interconnected entities. By utilizing graph-based methodologies, researchers can deepen their understanding of causal dynamics, leading to innovative approaches and applications in causal inference. The following subsection delves into various graph-based structures utilized for causal modeling, underscoring their importance in modern research.

5.1 Graph-Based Structures for Causal Modeling

Graph-based structures provide a robust framework for modeling intricate causal relationships, leveraging their capacity to represent entities and their interactions. Their strength lies in preserving causal information while minimizing non-causal factors, which is crucial for invariant representation learning. Graph Contrastive Invariant Learning (GCIL) exemplifies this by ensuring invariant representations are learned despite varying non-causal influences [70].

A comprehensive study introduces a framework embedding causal relationships within GNN architectures, significantly enhancing classification performance and interpretability [30]. The Relation-First Modeling Paradigm, demonstrated by RIRL, produces DAG-structured indices to explore causal relationships in graph-based models [71].

Integrating DAGs into generative models like DAG-WGAN illustrates scalability and accuracy in generating data that adheres to causal structures, outperforming traditional methods for both continuous and discrete datasets [72]. Furthermore, the Rationale Alignment Framework for GNN Explanations (RAF-GNN) optimizes GNN internal embeddings to produce accurate and consistent explanations by addressing inherent causal relationships [73].

As illustrated in Figure 5, which depicts the hierarchical structure of key advancements in graph-based causal modeling, the emphasis on invariant representation learning, causal frameworks in GNNs, and GNN explanation methods underscores the interconnected nature of these developments. This visual representation enhances our understanding of the advancements in the field and their implications for causal modeling.

Iterative Causal Discovery (ICD) enhances the stability and accuracy of recovering causal relationships by improving the statistical power of conditional independence tests with smaller conditioning sets [74]. This highlights the importance of precise causal discovery in graph-based models for reliable interpretations and predictions.

A theoretical framework combining network theory and causal inference distinguishes between undirected graphs, like Markov random fields, and DAGs, emphasizing DAGs' unique advantages in representing causal structures [75]. The De Bruijn Graph Neural Network (DBGNN) extends message passing to higher-order De Bruijn graphs, crucial for learning causal patterns in dynamic graphs, capturing temporal and structural complexities [64].

PAGE presents a parametric generative explainer that aligns predictions with explanations through a generative approach, utilizing an autoencoder and a discriminator, streamlining the generation of coherent causal explanations without perturbation-based techniques [76].

Advancements in graph-based structures for causal modeling reveal the significant potential of integrating graph theory and causal inference to enhance AI systems' interpretability and reliability. Focusing on causal relationships within graph data addresses critical trustworthiness issues in AI, such as susceptibility to distribution shifts, biases, and lack of explainability. This integration not only deepens the understanding of complex causal dynamics across various domains but also provides a framework for developing AI models that generalize better to unseen data, yield fairer outcomes, and offer insights into the mechanisms governing real-world phenomena [17, 9].

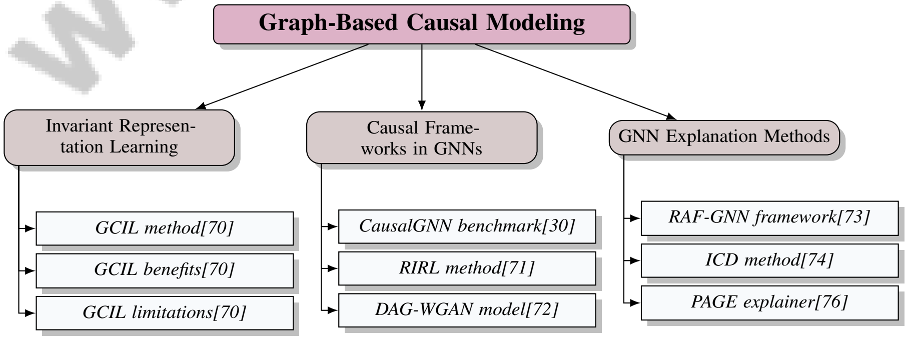


Figure 5: This figure illustrates the hierarchical structure of key advancements in graph-based causal modeling, emphasizing invariant representation learning, causal frameworks in GNNs, and GNN explanation methods.

5.2 Applications of GNNs in Causal Reasoning

Graph Neural Networks (GNNs) have emerged as formidable tools for causal reasoning, adeptly modeling complex relationships and dependencies within graph-structured data. The Casper framework, utilizing a Spatiotemporal Causal Attention (SCA) mechanism, uncovers sparse causal relationships among embeddings, thereby enhancing the interpretability and accuracy of causal models in spatiotemporal contexts [77]. This demonstrates GNNs' capability to manage temporal data intricacies, facilitating robust causal inferences.

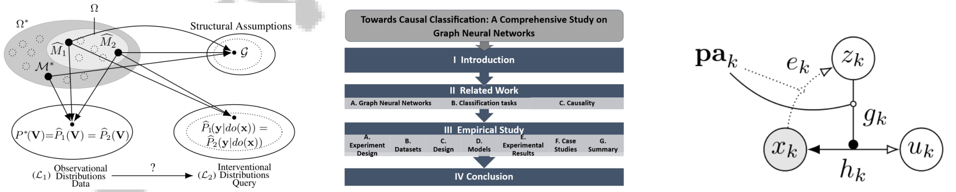
The synergy of Convolutional Neural Networks (CNNs) with GNNs showcases these models' ability to capture complex patterns in visual data, revealing causal relationships that traditional statistical methods may overlook [78]. This highlights the versatility of graph-based models in enhancing causal understanding across diverse data modalities.

Benchmark studies on nine representative GNN models, including standard models like GCN, GAT, GIN, and GraphSAGE, alongside causal variants such as GCN-CAL and GAT-CAL, in causal classification tasks provide a comprehensive overview of embedding causal insights into GNN architectures, improving interpretability and effectiveness in classification scenarios [30].

The De Bruijn Graph Neural Network (DBGNN) exemplifies GNN applications in dynamic graph contexts, where the temporal ordering of edges informs causal topology. Experimental evaluations on six time series datasets show that DBGNN significantly enhances node classification performance by leveraging causal topology effectively [64]. This advancement underscores the importance of considering temporal dynamics in causal reasoning tasks.

The PAGE framework introduces a parametric generative explainer that aligns predictions with explanations through a generative approach, evaluated on synthetic and real-world datasets. This method streamlines the generation of coherent and accurate causal explanations, outperforming baseline methods like GNNExplainer and PGExplainer [76]. By eliminating the need for perturbation-based techniques, PAGE enhances the efficiency and accuracy of causal reasoning in GNN applications.

The examples provided illustrate the extensive applications of GNNs in causal reasoning, emphasizing their capacity to enhance understanding of intricate causal dynamics across multiple fields. These applications range from analyzing spatiotemporal data to navigating dynamic graph contexts, addressing critical issues such as trustworthiness and explainability through causal learning techniques. Recent studies demonstrate how Causality-Inspired GNNs (CIGNNs) can mitigate risks associated with superficial correlations, while research on causal classification highlights their effectiveness in improving predictive accuracy across diverse datasets. Additionally, advancements in automating causal explanation analysis on social media underscore GNNs' potential to uncover underlying causal beliefs, emphasizing their significance in data-centric research [17, 2, 30].



(a) Structural Assumptions in Machine Learning[79]

(b) Towards Causal Classification: A Comprehensive Study on Graph Neural Networks[30]

(c) A Graphical Model with a Directed Acyclic Graph (DAG) Structure[11]

Figure 6: Examples of Applications of GNNs in Causal Reasoning

As depicted in Figure 6, GNNs have emerged as powerful tools in causal reasoning, offering novel insights and methodologies for understanding causality within complex systems. The application of GNNs in this domain is exemplified by three distinct studies. The first study, "Structural Assumptions in Machine Learning," elucidates the intricate relationship between structural assumptions and the use of observational and intervention data within machine learning frameworks, highlighting the foundational role of structural assumptions in causal inference. The second study, "Towards Causal Classification: A Comprehensive Study on Graph Neural Networks," systematically dissects how GNNs can be employed for causal classification tasks, providing a roadmap for leveraging these models to uncover causal relationships. Lastly, "A Graphical Model with a Directed Acyclic

Graph (DAG) Structure" illustrates a model where nodes and directed edges represent variables and their dependencies, encapsulating the essence of causal structures as understood through GNNs. Collectively, these examples demonstrate the versatility and potential of GNNs in advancing causal reasoning, paving the way for more robust analyses across various domains [79, 30, 11].

6 Trustworthy AI and Explainability

The convergence of trustworthy AI and explainability is increasingly critical as the demand for transparent AI systems grows. Central to this is causal reasoning, which elucidates AI models' underlying causal mechanisms, enhancing transparency and aligning with ethical standards and societal values.

6.1 Explainability through Causal Reasoning

Causal reasoning substantially improves AI systems' explainability by offering a structured method to comprehend the causal mechanisms of complex phenomena. It aligns with human reasoning, generating interpretable insights that aid decision-making [34]. Unlike traditional statistics, which focus on correlations, causal reasoning explores interventions and counterfactuals, providing a deeper understanding of causal structures that enhance machine learning algorithms' interpretability and reliability [2].

Methods like CARE exemplify causal reasoning in AI, targeting neural network defects to improve fairness and safety [6]. Additionally, formalizing harm in autonomous systems clarifies causal relationships, enhancing explainability [26], especially in high-stakes environments where understanding causal impacts is crucial.

In computational ethics, causality integration enhances complex ethical dilemma handling, improving AI systems' explainability [27]. This ensures AI models provide transparent, interpretable, and ethically aligned insights.

Causal reasoning is crucial for advancing AI explainability, enabling models to offer transparent, interpretable, and ethically aligned insights. By incorporating causal frameworks, AI systems enhance decision-making across fields like healthcare, finance, and legal judgment prediction, improving interpretability and generalization of outcomes by addressing cause-effect relationships, fostering trustworthiness, and reducing biases [37, 69, 10, 52, 9].

6.2 Causal Inference in Actionable Insights

Causal inference is foundational for deriving actionable insights in AI systems, offering a structured approach to understanding and manipulating causal relationships in complex datasets. This enhances AI models' interpretability and predictive power, allowing for accurate, actionable insights. Integrating causal inference with predictive modeling has led to robust results, as demonstrated by frameworks merging these methodologies [46].

A proposed framework using information fields and topological separation offers a robust causal inference method in complex systems, particularly involving cycles or spurious edges [44]. This framework improves the ability to discern true causal relationships, yielding actionable insights crucial for informed decision-making.

The CoGS framework generates interpretable counterfactuals respecting causal dependencies, offering actionable insights by elucidating causal pathways to specific outcomes [25]. This is valuable in domains where understanding intervention impacts is critical for optimizing outcomes.

Challenges remain, as trivariate Granger causality analysis can produce inaccuracies due to test criteria and noise [80]. Addressing these is vital for improving causal inferences' reliability and ensuring derived insights are actionable and accurate.

In social media and financial markets, nonparametric causality detection methods reveal that social media sentiment significantly impacts stock market prices, demonstrating causality rather than mere correlation [28]. This highlights causal inference's potential to uncover actionable insights in dynamic environments where traditional correlational analyses may fall short.

These advancements underscore causal inference's critical role in providing actionable insights across diverse AI applications. By enhancing AI models' interpretability and robustness, causal inference improves decision-making processes and increases AI systems' effectiveness and reliability. It addresses traditional AI models' limitations, which often struggle with generalization, fairness, and transparency due to a lack of understanding of cause-effect relationships. Integrating causal modeling techniques allows AI systems to account for hidden biases and missing information, leading to more informed predictions and insights. This approach fosters trust in AI applications and encourages further research into causality-driven solutions for developing trustworthy AI across various domains, including natural language processing and forecasting [32, 10, 9, 54].

7 Applications and Case Studies

7.1 Gene Expression Case Study

Causal inference methods have substantially advanced the analysis of gene expression, offering critical insights into gene regulation's complex biological processes. This case study leverages real-world datasets from cancer bioinformatics to illustrate how causal inference techniques elucidate gene expression patterns and their implications for gene dysregulation in carcinogenesis [32, 81, 82]. The Nonparanormal Causal Estimation (NCE) method, applied to Arabidopsis Thaliana gene expression data, demonstrates its superiority over traditional causal inference approaches, enhancing our understanding of biological mechanisms [83]. Further experiments with a breast cancer gene expression dataset, comprising 50 samples and 1500 variables, reveal causal interactions among genes, providing insights into regulatory networks affecting cancer progression and treatment [84]. Additionally, synthetic datasets generated from random directed acyclic graphs (DAGs) assess causal discovery in varied environments, emphasizing the need for adaptable causal inference methods to accurately identify causal pathways [85]. These studies collectively highlight the essential role of causal inference in gene expression analysis, empowering researchers to decode intricate causal relationships influencing gene regulation, thereby enhancing the robustness and interpretability of gene expression models [32, 40].

7.2 DISTANA's Performance in Spatio-Temporal Dynamics

The DISTANA architecture exemplifies the use of distributed graph neural networks (GNNs) for modeling spatiotemporal dynamics, effectively inferring, predicting, and denoising causal relationships within spatially distributed data [86]. By capturing complex dependencies across spatial and temporal dimensions, DISTANA enhances prediction accuracy and noise filtering. Its distributed approach, employing advanced causal explanation analysis techniques, allows for automatic identification of causal explanations in social media data, revealing insights into prevalent beliefs and their psychological impacts [87, 2]. This enhances understanding of social dynamics and health-related outcomes. DISTANA's efficacy is particularly notable in high-dimensional data and dynamic environments, where accurately inferring causal relationships is crucial for decision-making. By integrating causal reasoning, DISTANA not only improves prediction accuracy for complex spatiotemporal dynamics but also elucidates underlying causal mechanisms [37, 48, 2, 32, 86]. This makes it effective at denoising data streams and producing reliable predictions over extended timeframes, suitable for applications in brain imaging, supply chain dynamics, and environmental data analysis. The application of DISTANA in spatiotemporal dynamics highlights the potential of distributed GNNs to advance causal inference in complex environments, improving the interpretability and practical utility of AI models across various fields [43, 10, 49, 2, 9].

7.3 Synthetic Datasets and Causality 4 Climate Challenge

The use of synthetic datasets in the Causality 4 Climate Challenge marks a significant advancement in addressing climate-related causal challenges. Table 2 provides a detailed summary of the benchmarks used in the Causality 4 Climate Challenge, highlighting their relevance and application in advancing causal inference methodologies in climate science. These datasets provide a controlled experimental framework for testing and validating causal inference methodologies, allowing researchers to explore intricate climate phenomena without real-world data limitations like confounding variables and non-stationarity [94, 95, 96, 8, 32]. They enable the integration of qualitative domain knowledge with quantitative analyses, enhancing the robustness of models used in Earth and environmental

Benchmark	Size	Domain	Task Format	Metric
CF-Feasibility[88]	199,522	Income Classification	Counterfactual Explanation Generation	Validity, Feasibility Score
CELLO[89]	14,094	Causal Reasoning	Multi-choice	Accuracy, F1-score
CausalGNN[30]	7,000	Graph Classification	Multi-class Classification	Accuracy, F1-score
CLEAR[90]	2,808	Causal Inference	Causal Reasoning	Accuracy
SURD[62]	1,000,000	Causality Analysis	Causal Inference	Accuracy, Causality Leak
DCCM[91]	1,000,000	Causality Analysis	Causality Detection	F1-score, Sensitivity
FinSen[92]	160,000	Financial Market Analysis	Volatility Prediction	Expected Calibration Error, Focal Calibration Loss
IID[93]	40,612	Biomedical	Molecular Prediction	Accuracy, F1-score

Table 2: This table presents a comprehensive overview of various benchmarks utilized in the Causality 4 Climate Challenge. It details the size, domain, task format, and evaluation metrics for each benchmark, providing essential insights into their applicability in climate-related causal research.

sciences. In climate science, synthetic datasets play a crucial role in enhancing models and deepening understanding of complex interactions among climatic variables. Advanced techniques such as causality-informed deep learning address biases in climate projections, particularly those related to subgrid-scale processes like clouds and convection [95, 96, 8, 94, 58]. These datasets aid in refining climate simulations and pave the way for accurate predictions and insights into climate dynamics, contributing to reliable assessments of climate change impacts. The Causality 4 Climate Challenge employs synthetic datasets to enhance the development and evaluation of causal discovery algorithms tailored to climate data's complex characteristics, improving the accuracy and reliability of climate models and projections [94, 8]. By simulating realistic climate scenarios, these datasets validate the robustness and accuracy of causal inference methods, ensuring they effectively handle climate data complexities and uncertainties. Integrating synthetic datasets in climate-related causal research provides a powerful tool for advancing understanding of climate dynamics, enhancing predictive frameworks' accuracy and reliability, and fostering informed decision-making in climate policy and management [97, 2, 69].

8 Challenges and Future Directions

8.1 Challenges and Ethical Considerations

The integration of causality into AI systems presents significant challenges and ethical considerations, particularly in modeling complex causal relationships accurately. A critical issue is the dependency on precise causal graphs, where inaccuracies can undermine data quality and model reliability. This challenge is compounded by the complexities of multivariate systems and the non-white noise nature of error processes, often neglected in current research, leading to potential misinterpretations [35]. The necessity for explicit assumptions, such as ignorability and unconfoundedness, complicates the application of causal models, with potential bias if key variables remain unmeasured.

Methodologies also face hurdles like selection bias and limited generalizability across populations due to varying data collection processes, raising ethical concerns. The reliance on strong assumptions regarding unobserved variables complicates causal inference, particularly with hidden confounders. Furthermore, the computational complexity of analyzing emerging causal spaces limits effectiveness with larger datasets, presenting scalability challenges. These constraints arise from the need to maintain logical relationships among attributes while generating feasible counterfactual examples, as highlighted in recent studies on interpretable machine learning [98, 88, 10].

Ethical considerations stress the need for interpretability and transparency, with many studies lacking mechanisms to ensure these aspects, leading to biases in practical applications. The limitations of causal inference, including assumptions of the causal Markov condition and faithfulness, may not apply in real-world contexts, significantly hindering the generalizability of findings. This is particularly relevant in healthcare, where observational studies often encounter biases that can lead to erroneous causal conclusions. Moreover, reliance on spurious correlations in predictive models can undermine robustness. Addressing these issues necessitates integrating causal structures into models, despite inherent complexities. Recent advancements in causal modeling offer promising avenues for enhancing AI systems' reliability and interpretability, facilitating more informed decision-making in complex domains [83, 99, 14, 9]. Additionally, methods like PAGE face challenges due to their reliance on specific interpreters, limiting applicability in certain contexts.

Balancing performance and resource utilization is another ethical consideration, as advanced methods often require more resources than baseline approaches [5]. In scenarios with highly nonlinear dynamics or where foundational assumptions do not hold, methods may struggle, impacting the reliability of causal inference. Furthermore, ethical considerations surrounding the definition of harm highlight the complexities of applying these definitions across diverse scenarios.

Addressing these challenges and ethical considerations is crucial for enhancing AI models' accuracy and reliability while ensuring a robust understanding of causality that aligns with societal expectations. Leveraging causal modeling to improve interpretability and fairness will foster trust in AI systems. By aligning causal attributions with social behavior, AI models can excel technically while adhering to ethical standards, contributing positively to society [100, 101, 102, 9, 31]. Continuous advancements in causal inference methodologies are essential for enhancing reliability, scalability, and ethical alignment in AI applications, ensuring they provide trustworthy and interpretable insights across various domains.

8.2 Integration of Causal Models with Advanced Computational Frameworks

Integrating causal models with advanced computational frameworks is crucial for enhancing the robustness and scalability of causal inference methodologies across diverse domains. Future research should focus on refining covariance estimation techniques, particularly in high-dimensional data settings, to address challenges posed by confounding variables effectively [103]. These advancements will improve causal models' accuracy and reliability, enabling them to manage more complex causal structures effectively [46].

Incorporating iterative approximation algorithms within hypernetwork frameworks offers a promising avenue for managing computational complexity, with potential applications across various scientific fields. This approach may be further enhanced by exploring alternative methods for integrating causal inference with network analysis, which could involve examining causal relationships' dynamics over time and incorporating latent variables [75]. Additionally, enhancing the causal convolution process with constraints to improve delay precision while maintaining overall causal discovery performance remains a critical area for future research [61].

In multi-agent systems, future research should focus on enhancing causal imitation learning and improving knowledge sharing among agents, particularly in safety-critical applications. Developing decentralized reasoning methods that leverage causal models can significantly enhance these systems' interpretability and reliability [50]. Furthermore, improving biometric systems' interpretability through causal models and exploring counterfactual explanations are essential areas for investigation [20].

In healthcare, advancing methods for causal discovery from data and enhancing model interpretability in AI applications are critical for improving causal inference methodologies [14]. Additionally, exploring the benefits of utilizing causal relations in data generation tasks and refining causal discovery methods could significantly enhance models like Causal-TGAN [104].

Future research should also aim to extend Structural Hawkes Processes (SHPs) to more general point processes with diverse intensity functions, thereby enhancing causal discovery capabilities [105]. Moreover, refining causal inference techniques and exploring new applications in climate science are essential for advancing the field [8]. By leveraging these innovations, AI systems can achieve more accurate, reliable, and ethically aligned insights, enhancing their applicability and impact across diverse fields.

9 Conclusion

Causality plays a crucial role in enhancing the trustworthiness and explainability of AI systems across various sectors. The distinction between correlation and causation provides a critical framework for reliable decision-making, especially in sensitive areas like healthcare and finance, where ethical considerations are significant. Integrating causal reasoning into AI models enhances interpretability and predictive fairness, addressing challenges in domains where understanding causal mechanisms is essential for effective decision-making.

Advancements in causal representation learning, graph neural networks, and causal inference methodologies hold promise for improving AI systems' robustness and transparency. For instance, the application of causal reasoning in Embodied AI has demonstrated the importance of causality in enhancing navigation capabilities, underscoring its necessity for AI performance improvements. Moreover, rigorous statistical tools for neural connectivity analysis highlight causality's significance in neural network studies.

Despite these advancements, challenges persist in accurately modeling causal relationships, particularly in high-dimensional and complex data environments. To advance causal learning research and ensure reliable application, systematic, objective, and transparent evaluation processes are vital. Continued research and innovation in causal inference methodologies are essential for overcoming existing challenges and fully leveraging causality's potential in AI.

References

[1] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022.

[2] Youngseo Son, Nipun Bayas, and H. Andrew Schwartz. Causal explanation analysis on social media, 2018.

[3] Emanuele Cavenaghi, Alessio Zanga, Fabio Stella, and Markus Zanker. The importance of causality in decision making: A perspective on recommender systems, 2024.

[4] Yuxi Xie, Guanzhen Li, and Min-Yen Kan. Echo: A visio-linguistic dataset for event causality inference via human-centric reasoning, 2023.

[5] Ruoyu Wang, Yao Liu, Yuanjiang Cao, and Lina Yao. Causality-aware transformer networks for robotic navigation, 2024.

[6] Bing Sun, Jun Sun, Long H Pham, and Jie Shi. Causality-based neural network repair. In *Proceedings of the 44th International Conference on Software Engineering*, pages 338–349, 2022.

[7] Stefano Gogioso and Nicola Pinzani. The combinatorics of causality, 2023.

[8] X. San Liang, Dake Chen, and Renhe Zhang. Quantitative causality, causality-guided scientific discovery, and causal machine learning, 2024.

[9] Niloy Ganguly, Dren Fazlija, Maryam Badar, Marco Fisichella, Sandipan Sikdar, Johanna Schrader, Jonas Wallat, Koustav Rudra, Manolis Koubarakis, Gourab K. Patro, Wadhah Zai El Amri, and Wolfgang Nejdl. A review of the role of causality in developing trustworthy ai systems, 2023.

[10] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. Causality learning: A new perspective for interpretable machine learning, 2021.

[11] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

[12] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective, 2019.

[13] Cao Zhihao and Qu Hongchun. Review on causality detection based on empirical dynamic modeling, 2023.

[14] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. Causal inference in medicine and in health policy, a summary, 2022.

[15] Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder, 2020.

[16] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality, 2019.

[17] Wenzhao Jiang, Hao Liu, and Hui Xiong. When graph neural network meets causality: Opportunities, methodologies and an outlook, 2024.

[18] Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning, 2023.

[19] Bijan Mazaheri, Siddharth Jain, Matthew Cook, and Jehoshua Bruck. Omitted labels in causality: A study of paradoxes, 2024.

[20] Pedro C. Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F. Sequeira, Arun Ross, and Jaime S. Cardoso. Causality-inspired taxonomy for explainable artificial intelligence, 2024.

[21] Tianxiang Zhao, Wenchao Yu, Suhang Wang, Lu Wang, Xiang Zhang, Yuncong Chen, Yanchi Liu, Wei Cheng, and Haifeng Chen. Interpretable imitation learning with dynamic causal relations, 2024.

[22] Jinling Yan, Shao-Wu Zhang, Chihao Zhang, Weitian Huang, Jifan Shi, and Luonan Chen. Dynamical causality under invisible confounders, 2024.

[23] Ziyi Tang, Ruilin Wang, Weixing Chen, Yongsen Zheng, Zechuan Chen, Yang Liu, Keze Wang, Tianshui Chen, and Liang Lin. Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms, 2025.

[24] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024.

[25] Sopam Dasgupta, Joaquín Arias, Elmer Salazar, and Gopal Gupta. Cogs: Model agnostic causality constrained counterfactual explanations using goal-directed asp, 2024.

[26] Sander Beckers, Hana Chockler, and Joseph Y. Halpern. A causal analysis of harm, 2023.

[27] Camilo Sarmiento, Gauvain Bourgne, Katsumi Inoue, Daniele Cavalli, and Jean-Gabriel Ganascia. Action languages based actual causality for computational ethics: a sound and complete implementation in asp, 2023.

[28] Fani Tsapeli, Mirco Musolesi, and Peter Tino. Non-parametric causality detection: An application to social media and financial data, 2017.

[29] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference, 2018.

[30] Simi Job, Xiaohui Tao, Taotao Cai, Lin Li, Haoran Xie, and Jianming Yong. Towards causal classification: A comprehensive study on graph neural networks, 2024.

[31] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality? *arXiv preprint arXiv:2206.10591*, 2022.

[32] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond, 2022.

[33] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions, 2022.

[34] Judea Pearl. Fusion, propagation, and structuring in belief networks. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 139–188. 2022.

[35] Michael Eichler. Comment on: Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance, 2012.

[36] Luiz A. Baccalá, Daniel Y. Takahashi, and Koichi Sameshima. Consolidating a link centered neural connectivity framework with directed transfer function asymptotics, 2015.

[37] Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Shubhashis Sengupta, and Andrew E. Fano. Causal bert : Language models for causality detection between events expressed in text, 2021.

[38] Tom Michoel and Jitao David Zhang. Causal inference in drug discovery and development, 2022.

[39] Marcus Klasson, Kun Zhang, Bo C. Bertilson, Cheng Zhang, and Hedvig Kjellström. Causality refined diagnostic prediction, 2017.

[40] Erica EM Moodie and David A Stephens. Causal inference: critical developments, past and future, 2022.

[41] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[42] Gaël Gendron, Michael Witbrock, and Gillian Dobbie. A survey of methods, challenges and perspectives in causality. *arXiv preprint arXiv:2302.00293*, 2023.

[43] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery, 2021.

[44] Benjamin Heymann, Michel de Lara, and Jean-Philippe Chancelier. Causal inference theory with information dependency models, 2021.

[45] Zhuangyan Fang, Yue Liu, Zhi Geng, Shengyu Zhu, and Yangbo He. A local method for identifying causal relations under markov equivalence, 2022.

[46] Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

[47] Boris Lorbeer and Axel Küpper. Robust causal analysis of linear cyclic systems with hidden confounders, 2024.

[48] Xiao Xie, Moqi He, and Yingcai Wu. Causalflow: Visual analytics of causality in event sequences, 2020.

[49] Jiachang Liu, Qi Zhang, Chongyang Shi, Usman Naseem, Shoujin Wang, and Ivor Tsang. Causal intervention for abstractive related work generation, 2023.

[50] St John Grimbly, Jonathan Shock, and Arnu Pretorius. Causal multi-agent reinforcement learning: Review and open problems, 2021.

[51] Rafael Pina, Varuna De Silva, and Corentin Artaud. Learning independently from causality in multi-agent environments, 2023.

[52] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019.

[53] Bernhard Schölkopf and Julius von Kügelgen. From statistical to causal learning, 2022.

[54] Zhixuan Chu, Hui Ding, Guang Zeng, Shiyu Wang, and Yiming Li. Causal interventional prediction system for robust and explainable effect forecasting, 2024.

[55] Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9(1):289–319, 2022.

[56] Ruichu Cai, Siyang Huang, Jie Qiao, Wei Chen, Yan Zeng, Keli Zhang, Fuchun Sun, Yang Yu, and Zhifeng Hao. Learning by doing: An online causal reinforcement learning framework with causal-aware policy, 2024.

[57] Sachin Kasture. Discovering dependencies in complex physical systems using neural networks, 2021.

[58] M. Ali Vosoughi and Axel Wismuller. Leveraging pre-images to discover nonlinear relationships in multivariate environments, 2021.

[59] Zijian Zhang, Vinay Setty, Yumeng Wang, and Avishek Anand. Disco: Discovering overfittings as causal rules for text classification models, 2024.

[60] Jinfan Hu, Jinjin Gu, Shiyao Yu, Fanghua Yu, Zheyuan Li, Zhiyuan You, Chaochao Lu, and Chao Dong. Interpreting low-level vision models with causal effect maps, 2024.

[61] Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Causalformer: An interpretable transformer for temporal causal discovery, 2024.

[62] Álvaro Martínez-Sánchez, Gonzalo Arranz, and Adrián Lozano-Durán. Decomposing causality into its synergistic, unique, and redundant components, 2024.

[63] Xiaoxuan Li, Yao Liu, Ruoyu Wang, and Lina Yao. Regularized multi-llms collaboration for enhanced score-based causal discovery, 2024.

[64] Lisi Qarkaxhija, Vincenzo Perri, and Ingo Scholtes. De bruijn goes neural: Causality-aware graph neural networks for time series data on dynamic graphs, 2022.

[65] Sipan Aslan and Hernando Ombao. Nonlinear causality in brain networks: With application to motor imagery vs execution, 2024.

[66] Fangting Zhou, Kejun He, Kunbo Wang, Yanxun Xu, and Yang Ni. Functional bayesian networks for discovering causality from multivariate functional data, 2022.

[67] Jiangmeng Li, Bin Qin, Qirui Ji, Yi Li, Wenwen Qiang, Jianwen Cao, and Fanjiang Xu. Teleporter theory: A general and simple approach for modeling cross-world counterfactual causality, 2024.

[68] Dezhi Yang, Guoxian Yu, Jun Wang, Zhengtian Wu, and Maozu Guo. Reinforcement causal structure learning on order graph, 2022.

[69] Haotian Chen, Lingwei Zhang, Yiran Liu, Fanchao Chen, and Yang Yu. Knowledge is power: Understanding causality makes legal judgment prediction models more generalizable and robust, 2023.

[70] Yanhu Mo, Xiao Wang, Shaohua Fan, and Chuan Shi. Graph contrastive invariant learning from the causal perspective, 2024.

[71] Jia Li and Xiang Li. Relation-first modeling paradigm for causal representation learning toward the development of agi, 2024.

[72] Hristo Petkov, Colin Hanley, and Feng Dong. Causality learning with wasserstein generative adversarial networks, 2022.

[73] Tianxiang Zhao, Dongsheng Luo, Xiang Zhang, and Suhang Wang. Faithful and consistent graph neural network explanations with rationale alignment, 2023.

[74] Raanan Y. Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias, 2022.

[75] Fabian Dablander and Max Hinne. Node centrality measures are a poor substitute for causal inference. *Scientific reports*, 9(1):6846, 2019.

[76] Yang Qiu, Wei Liu, Jun Wang, and Ruixuan Li. Page: Parametric generative explainer for graph neural network, 2024.

[77] Baoyu Jing, Dawei Zhou, Kan Ren, and Carl Yang. Causality-aware spatiotemporal graph neural networks for spatiotemporal time series imputation, 2024.

[78] Karamjit Singh, Garima Gupta, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Deep convolutional neural networks for pairwise causality, 2017.

[79] Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

[80] Leo Carlos-Sandberg and Christopher D. Clack. The sensitivity of trivariate granger causality to test criteria and data errors, 2019.

[81] Raha Moraffah, Paras Sheth, Mansooreh Karami, Anchit Bhattacharya, Qianru Wang, Anique Tahir, Adrienne Raglin, and Huan Liu. Causal inference for time series analysis: Problems, methods and evaluation. *Knowledge and Information Systems*, 63:3041–3085, 2021.

[82] Jean Pierre Gomez. Nonparametric causal discovery with applications to cancer bioinformatics, 2024.

[83] Seyed Mahdi Mahmoudi and Ernst Wit. Estimating causal effects from nonparanormal observational data, 2016.

[84] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, and Shu Hu. Parallelpc: an r package for efficient constraint based causal exploration, 2015.

[85] Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis, 2022.

[86] Matthias Karlbauer, Sebastian Otte, Hendrik P. A. Lensch, Thomas Scholten, Volker Wulfmeyer, and Martin V. Butz. Inferring, predicting, and denoising causal wave dynamics, 2020.

[87] James R. Clough, Jamie Gollings, Tamar V. Loach, and Tim S. Evans. Transitive reduction of citation networks, 2014.

[88] Kleopatra Markou, Dimitrios Tomaras, Vana Kalogeraki, and Dimitrios Gunopulos. A framework for feasible counterfactual exploration incorporating causality, sparsity and density, 2024.

[89] Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. Cello: Causal evaluation of large vision-language models, 2024.

[90] Sirui Chen, Mengying Xu, Kun Wang, Xingyu Zeng, Rui Zhao, Shengjie Zhao, and Chaochao Lu. Clear: Can language models really understand causal graphs?, 2024.

[91] Angeliki Papana, Elsa Siggiridou, and Dimitris Kugiumtzis. Detecting direct causality in multivariate time series: A comparative study, 2020.

[92] Wenhao Liang, Zhengyang Li, and Weitong Chen. Enhancing financial market predictions: Causality-driven feature selection, 2024.

[93] Jiqing Wu, Inti Zlobec, Maxime Lafarge, Yukun He, and Viktor H. Koelzer. Towards iid representation learning and its application on biomedical data, 2022.

[94] Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, and Veronika Eyring. Causally-informed deep learning to improve climate models and projections, 2024.

[95] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

[96] Yan Lyu, Sunhao Dai, Peng Wu, Quanyu Dai, Yuhao Deng, Wenjie Hu, Zhenhua Dong, Jun Xu, Shengyu Zhu, and Xiao-Hua Zhou. A semi-synthetic dataset generation framework for causal inference in recommender systems, 2022.

[97] Ilias Tsoumas, Vasileios Sitokonstantinou, Georgios Giannarakis, Evagelia Lampiri, Christos Athanassiou, Gustau Camps-Valls, Charalampos Kontoes, and Ioannis Athanasiadis. Causality and explainability for trustworthy integrated pest management, 2023.

[98] Matthew J. Vowels. Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables, 2022.

[99] Khurram Javed, Martha White, and Yoshua Bengio. Learning causal models online, 2020.

[100] Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning, 2024.

[101] Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution, 2021.

[102] Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality?, 2022.

[103] Jacek P. Dmochowski. Learning latent causal relationships in multiple time series, 2022.

[104] Bingyang Wen, Luis Oliveros Colon, K. P. Subbalakshmi, and R. Chandramouli. Causal-tgan: Generating tabular data using causal generative adversarial networks, 2021.

[105] Jie Qiao, Ruichu Cai, Siyu Wu, Yu Xiang, Keli Zhang, and Zhifeng Hao. Structural hawkes processes for learning causal structure from discrete-time event sequences, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

AI-generated, for reference only.