The role of causality in explainable artificial intelligence

Gianluca Carloni[a,b,c], Andrea Berti[a,b], Sara Colantonio[b]

[a]*Department of Information Engineering, University of Pisa, Via Caruso 16, Pisa, 56122, Italy*
[b]*Institute of Information Science and Technologies (ISTI) - National Research Council (CNR), Via Moruzzi 1, Pisa, 56124, Italy*
[c]*Corresponding author ORCiD: 0000-0002-5774-361X*

Abstract

Causality and eXplainable Artificial Intelligence (XAI) have developed as separate fields in computer science, even though the underlying concepts of causation and explanation share common ancient roots. This is further enforced by the lack of review works jointly covering these two fields. In this paper, we investigate the literature to try to understand how and to what extent causality and XAI are intertwined. More precisely, we seek to uncover what kinds of relationships exist between the two concepts and how one can benefit from them, for instance, in building trust in AI systems. As a result, three main perspectives are identified. In the first one, the lack of causality is seen as one of the major limitations of current AI and XAI approaches, and the "optimal" form of explanations is investigated. The second is a pragmatic perspective and considers XAI as a tool to foster scientific exploration for causal inquiry, via the identification of pursue-worthy experimental manipulations. Finally, the third perspective supports the idea that causality is propaedeutic to XAI in three possible manners: exploiting concepts borrowed from causality to support or improve XAI, utilizing counterfactuals for explainability, and considering accessing a causal model as explaining itself. To complement our analysis, we also provide relevant software solutions used to automate causal tasks. We believe our work provides a unified view of the two fields of causality and XAI by highlighting potential domain bridges and uncovering possible limitations.

Keywords: causality, explainable artificial intelligence, causal discovery, counterfactuals, structural causal models

Article Category: ADVANCED REVIEW.
Conflict of Interest: We declare that we have no conflict of interest.

1. Graphical/Visual Abstract and Caption

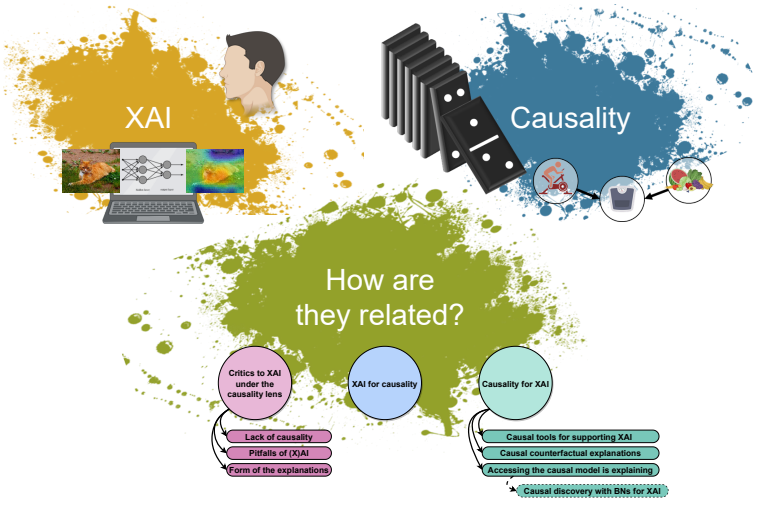

Figure 1: Graphical Abstract: Reviewing the literature to uncover how eXplainable Artificial Intelligence (XAI) and causality are related - the three main perspectives.

2. Introduction

Causation and explanation are not new concepts, since they have always drawn humans' attention. They are, indeed, highly intertwined since the ancient Greeks and throughout the philosophy of science (Sec. 3.1). Unfortunately, it seems that these concepts have had a diverse evolution in the field of Artificial Intelligence (AI). Regarding explanations, the eXplainable AI (XAI) research field has been formalized in the past few years to overcome the limitations of conventional black-box machine learning (ML) and deep learning (DL) models (Sec. 3.2). Regarding the field of causality (Sec. 3.3), some seminal works have been investigating its integration within ML and DL systems (Schölkopf et al., 2021; Berrevoets et al., 2023). What seems to emerge from the current literature is that there is no clear vision of whether there is a dependent relationship between the two fields.

In this review, we investigate the interdisciplinary literature regarding causality and XAI from both theoretical and methodological viewpoints to try to gain a clearer understanding of this question. Our results show three main perspectives can be identified. The first way to relate the two fields is to move some critics to XAI under a causal lens, to serve as a watch out. In this regard, a non-negligible subset of publications recognizes causality as a missing component of current XAI research to achieve robust and explainable systems. Other works highlight how the field of XAI (and AI by extension) suffers from certain innate issues, making the problem itself ill-posed. In a similar light, a further branch of works investigates different forms and desiderata of the XAI-produced explanations and their link with the causal theory. The second perspective tries to relate XAI and causality in a pragmatic way and sees the former as a means to get to the latter. Such works believe XAI has the potential to foster scientific exploration for causal inquiry. Indeed, by means of approaches able to identify pursue-worthy experimental manipulations, XAI may help scientists generate hypotheses about possible causal relationships to be tested. The third perspective turns the previous one around, claiming that causality is propaedeutic to XAI. Causal tools and metrics are exploited to implement XAI, and specific XAI approaches are brought back to their formal causal definition to improve generalization capabilities. Among the distinctive ideas of this perspective, getting access to the causal model of a system is a way to intrinsically explain the system itself.

We argue that the third of the perspectives is the one to be preferred to correctly combine the two areas of causality and XAI to advance the research toward reliable systems that are truly useful to humans. Overall, the novelty of our work lies in bridging the XAI-causation gap rigorously, highlighting areas of future development, and exposing limitations.

3. Rationale and Objective

3.1. Ancient roots

The study of causation and explanation can be traced back to the ancient Greek philosophers. Aristotle, for instance, introduced causality as the foundation of explanation and argued that there must be a necessary and sufficient reason for every event (Hankinson, 1998).

As early as the 18th century, the empiricist David Hume formalized causation in terms of sufficient and necessary conditions: an event c causes an event e if and only if there are event-types C and E such that C is necessary and sufficient for E. He, however, remained skeptical about humans' ability to explain and truly know any event. Indeed, he argued that we cannot perceive any necessary connection between cause and effect, but only events occurring in regular succession based on habit (Hume, 2003).

From the 1950s onward, some others also investigated scientific explanations. Initially, the "standard model" of explanation was deductive, following the Deductive-Nomological (DN) model by Hempel and Oppenheim (1948). An outcome was implied logically from universal laws plus initial conditions via deductive inference (e.g., explaining the volume of gas via the ideal gas law and some observations such as pressure). Regarding Hempel's viewpoint on causality, causal explanations are special cases of DN explanations, but not all laws and explanations are causal.

Later, Salmon (1984) developed a model in which good scientific explanations must be statistically relevant to the outcome to get explained. He argued that, in attempting to explain probabilistic phenomena, we seek not merely a high probability but screen for causal influence by removing system components to find ones that alter the probability. Salmon found causality ubiquitous in scientific explanation and was convinced that the time had come to put the "cause" back into "because". Although remaining vague as to how to attain it, he invited scientists to reconsider the role of causal relations as potentially fundamental constituents of adequate explanations.

3.2. The need for XAI

Given the rapidly increasing interest in data mining for knowledge discovery, AI is becoming pervasive in our lives, and understanding and trusting its decisions has become imperative. This is further enforced, for instance, by the current guidelines for trustworthy AI by the European Commission[1]. Indeed, opacity in such decisions can lead to reluctance when adopting AI in a product, a decision process, or research. This, therefore, can result in missed opportunities in the use of AI to its fullest potential. To prevent this scenario, the research field of XAI aims to provide humans with explanations to understand the reasoning behind an AI system and its decision-making process. In other words, the goal of XAI is to enable end-users to understand the underlying explanatory factors of why an AI decision is taken. The term XAI was first introduced in Van Lent et al. (2004), but its popularity has spread across the literature only after the DARPA's XAI program (Gunning and Aha, 2019), reaching a certain degree of maturity to date (Guidotti et al., 2018; Du et al., 2019; Carvalho et al., 2019; Rudin, 2019; Arrieta et al., 2020; Molnar, 2020).

XAI systems have been prioritized in different fields, such as healthcare, finance, education, and legal. In healthcare, XAI has been utilized for medical image analysis, acute critical illness prediction, intraoperative decision support systems, drug discovery, and treatment recommendations (Van der Velden et al., 2022; Lauritsen et al., 2020; Gordon et al., 2019; Jiménez-Luna et al., 2020). Regarding finance, popular applications of XAI are credit risk management and prediction, loan underwriting automation, and investment advice (Bussmann et al., 2021; Moscato et al., 2021; Sachan et al., 2020; Yang et al., 2021). In education, XAI has been applied in automatic essay scoring systems, educational data mining, and adaptive learning systems (Kumar and Boulanger, 2020; Alonso and Casalino, 2019; Khosravi et al., 2022), while digital forensics for law enforcement context represents an example in the legal domain (Hall et al., 2022).

Regardless of the application field, XAI is driven by the idea of making the reasoning process of AI transparent and, therefore, AI models more intelligible to humans. Accordingly, when it comes to explaining the logic of an inferential system or a learning algorithm, four aspects can be identified as the main driving motivations for XAI (Adadi and Berrada, 2018): (i) explain to justify (i.e., provide justifications for particular decisions to make sure they are not unfairly yielded by bias), (ii) explain to control (i.e., understand the system behavior for debugging vulnerabilities and potential flaws), (iii) explain to improve (i.e, understand the system behavior for enhancing its accuracy and efficiency), and (iv) explain to discover (i.e., learn from machines their knowledge on relationships and patterns).

3.3. A causal perspective

Even though the wide literature on causality spans different interpretations, such as the causal potential theory (Xu, 2018) and Wiener-Granger causality (Granger, 1969), the one by computer scientist Judea Pearl is popularly associated with AI. Pearl identifies some major obstacles still undermining the ability of AI systems in reasoning in a way akin to humans, to be overcome by equipping machines with causal modeling tools (Pearl, 2019). Among those obstacles, is the lack of robustness of AI systems in recognizing or responding to new situations without being specifically programmed (i.e., adaptability), as well as their inability to grasp cause-effect relationships. Instead, those abilities are innate features of human beings, who can communicate with, learn from, and instruct each other since all their brains reason in terms of cause-effect relationships (Pearl, 2018).

Pearl argues humans organize their knowledge of the world according to three distinct levels of cognitive ability, which he embodies in distinct rungs of the *Ladder of Causation* (Pearl and Mackenzie, 2018). As Tab. 1 shows, the first rung is *Association* and involves passive observation of data. Reasoning on this level could not distinguish the cause from the effect and, although this might come as a surprise to some, Pearl argues that it is where conventional AI approaches to classification or regression stand today. The second rung is *Intervention* and involves not just viewing what exists, but also changing it. However, reasoning on this rung cannot reveal what will happen in an imaginary world where some observed facts are bluntly negated. To this end, we need to climb to the third rung, i.e. *Counterfactuals* (CF). It involves imagination since to answer counterfactual queries one needs to go back in time and change history. For instance, we may wonder whether it was, indeed, *turning the heating system on* that caused a *warm apartment* or, rather, for instance, the outdoor weather.

Note that, in a somewhat confusing way, the term "counterfactual" may be encountered also in the XAI literature, where it applies to any instance with an alternative outcome. There, a *counterfactual explanation* (CFE)

[1] `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

Table 1: The Ladder of Causation by Pearl and Mackenzie (2018).

Level (rung)	Cognitive ability (activity)	Typical questions	Examples
Association	Seeing, observing (i.e., recognizing recurrent patterns in an environment)	“What if I see . . . ?”	"What is the probability that an apartment is warm if I see the heating system being on?"
Intervention	Doing (i.e., predicting the effect(s) of multiple intentional actions on the environment and choosing the best to produce a desired outcome)	"What if I do . . . ?"	"What is the probability that the apartment will get warm if I turn on the heating system?".
Counterfactuals	Imagining, reasoning in retrospection, and understanding	"What if I had done . . . ?"	"What would have happened to the indoor comfort of the apartment if I had kept the heating system off?".

refers to the smallest change in an input that changes the prediction of an ML classifier (Wachter et al., 2017; Mothilal et al., 2020). This concept is quite distinct from the causal meaning of the term. In this regard, as a piece of clarification, we utilize *CFE* and *CF* to address, respectively, the XAI method and the causality concept.

In general, building models that represent causal relationships among variables from observations may be challenging without relying on assumptions that are hard to verify in practice, such as the absence of unmeasured confounding between the variables (Robins and Wasserman, 1999; Greenland and Mansournia, 2015). Nevertheless, Pearl’s work was revolutionary in that it transformed causality from a notion clouded in mystery into a concept with logical foundations and defined semantics. The formalization of causality in mathematical terms within an axiomatic framework allowed the development of automatic computational systems for causal modeling. We refer the reader to Appendix A for some notations and terminology regarding Pearl’s causality (and related concepts).

3.4. Objective

This review investigates the role(s) of causality in the world of XAI today or, broadly, the relationship between causality and XAI. Throughout the paper, we aimed to refer to an interdisciplinary audience, which reflects the use of an accurate (yet not overly zealous) register, leaving the more technical parts (e.g., mathematical notations and supplementary details) to Appendix A and Appendix B.

Three main pieces of information led us to believe that those could be complementary fields, and, thus, motivated us to start our investigation. First, the concepts of causality and explanation have been jointly investigated since ancient times (Sec. 3.1). Second, even though they were born separately in the field of AI (*explanation* as XAI and *causation* as Pearl’s causality theory), they share a common goal. Indeed, both fields feature human-centricity in AI systems and aim to ensure true usefulness to humans, be it by explaining in a human-comprehensible way what an AI system did, or by designing the system in such a way that it reasons like humans (Sec. 3.2 — 3.3). Third, another "canary in the coal mine" for us was the presence of the same "counterfactual" term in both fields (Sec. 3.3).

To the best of our knowledge, Chou et al. (2022) are the only ones investigating a somewhat similar question, albeit with a narrower scope. They systematically review current counterfactual model-agnostic approaches (i.e., CFEs) studying how they could promote *causability*. Causability is a relatively new term representing "the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use." (Holzinger et al., 2019). Since causability differs from causality, this is the first (and major) difference with our study, which covers the wide notion of causality itself. Our work also departs from Chou et al. (2022) in that they solely investigate CFE methods, while, in our analysis, we consider the whole corpus of XAI literature, which also includes (but is not limited to) CFEs.

4. Methods

This review aims at exploring the literature surrounding the relationship between causality and XAI, from both theoretical and methodological viewpoints. We conducted our work by adopting a structured process that involved the following: (i) specifying the eligibility criteria; (ii) detailing the information sources; (iii) illustrating the search strategy on specified databases; (iv) describing the selection process; (v) conducting a high-level analysis on the cohort of selected studies; (vi) extracting relevant data and information from studies; and (vii) synthesizing results.

We carried out our search on four popular bibliographic databases, *Scopus*[2], *IEEE Xplore digital library* (s. IEEE)[3], *Web of Science* (s. WoS)[4], and *ACM Guide to Computing Literature* (s. ACM)[5], utilizing the following query:

(CAUSAL*) **AND** (EXPLA*) **AND** ("XAI" OR "EXPLAINABLE ARTIFICIAL INTELLIGENCE" OR "EXPLAINABLE AI") **AND** ("MACHINE LEARNING" OR "AI" OR "ARTIFICIAL INTELLIGENCE" OR "DEEP LEARNING")

Elements within brackets had to be present within at least one of the title, abstract, or keywords of the manuscript. Terms ending with the wildcard "$*$" matched all the terms with the specified common prefix. Among the obtained publications, we ensured that only peer-reviewed papers from conference proceedings and journals were included. Upon completion of the process of identification, screening, eligibility, and inclusion of articles, 51 publications formed the basis of our review. We describe the technical details of the whole study collection process in Appendix B.

In our study, we first performed a high-level analysis of the final cohort of records regarding keywords co-occurrence, then, we extracted information from the publications to answer our research question, and, finally, we collected any cited software solutions in a structured way.

4.1. Keywords' co-occurrence analysis

Regarding the high-level analysis of the final cohort of records, we constructed a bibliometric network of articles' keywords co-occurrence, by utilizing the Java-based application *VOS Viewer*[6]. Bibliometric networks are methods to visualize, in the form of graphs, the collective interconnection of specific terms or authors within a corpus of written text. In our setting, we applied such networks to study the paired presence of articles' keywords within a corpus of scientific manuscripts.

4.2. Research question analysis

For each of the papers that were included in the review, we identified the most relevant aspects on a conceptual level. According to the research question, we searched for any theoretical viewpoints and comments on the possible ways in which causality and XAI may relate, including formalization frameworks and insights from AI, cognitive, and philosophical perspectives.

Based on the collected information, we performed a topic clustering procedure to organize the literature in related concepts and gain a global view of the field. Selecting cluster topics for a multidisciplinary field as that of causality in the broad field of XAI proved challenging. Topics that are too general would result in an excessively vague and superficial division of papers and therefore be of little use in answering the research question. On the other hand, topics that are too specific would create many quasi-empty clusters, resulting in an improper division, which lacks abstraction capabilities and prevents an overall view of the field. Therefore, we iteratively refined the clusters during a trial-and-error process.

4.3. Software tools collection

During the analysis of the full-text manuscripts, we kept track, in a structured collection, of any cited software solutions (e.g., tools, libraries, packages), whenever they were used to automate causal tasks. Specifically, for each one, we analyzed: (i) the URL of the corresponding web-page; (ii) whether the software was commercial or with an open-source license, according to the Open Source Initiative[7]; (iii) the name of the company for cases of commercial software; (iv) the eventual release publication that launched the software; (v) whether the frontend consisted in a command line interface (CLI) or a graphical user interface (GUI); and, finally, (vi) the main field of application and purpose.

[2] https://www.scopus.com/
[3] https://ieeexplore.ieee.org/
[4] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[5] https://dl.acm.org/browse
[6] https://www.vosviewer.com/
[7] https://opensource.org

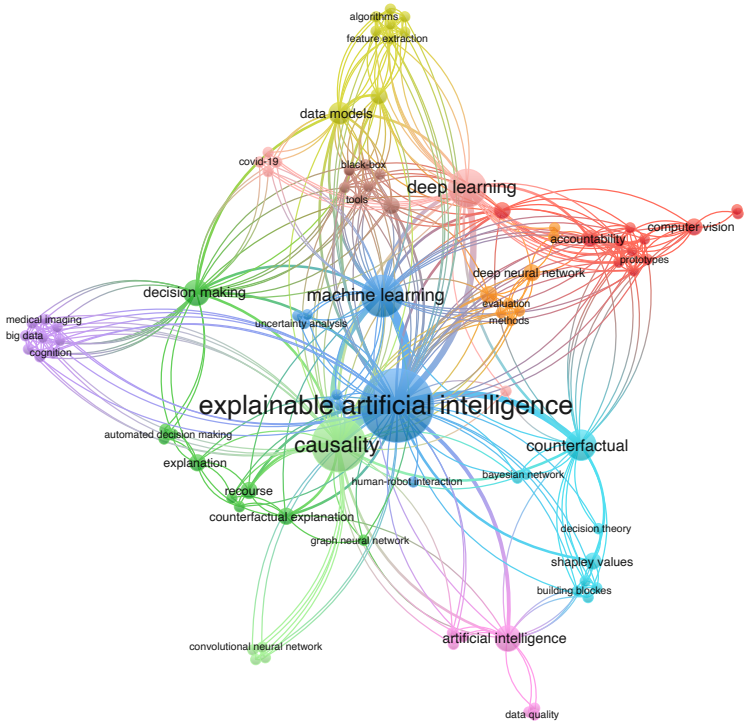

Figure 2: Bibliometric network of papers' keywords for the cohort of publications included in the review.

5. Results to the keywords' co-occurrence analysis

As a result of the high-level analysis of the final cohort of records, we obtained the bibliometric network shown in Fig. 2. The items (i.e., nodes) of the network represent terms (specifically, articles' keywords); the link (i.e., edge) between two items represents a co-occurrence relation between two keywords; the strength of a link indicates the number of articles in which two keywords occur together; and, finally, the importance of an item is given by the number of links of that keyword with other keywords and by the total strength of the links of that keyword with other keywords. Accordingly, more important keywords are represented by bigger circles in the network visualization, and more prominent links are represented by larger edges between keywords.

This visualization provides insight into how and to what extent the literature relates different research concepts, and it helped us to appreciate the multidisciplinary nature of our research question. Moreover, it is possible to marginalize the scope of specific keywords by identifying the terms to which they relate, as shown in Figs. 3a-b for the keywords *causality* and *counterfactual*, respectively. The relevance and wide scope of the first are justified by the structure of our query, where it was an obligatory search term. Regarding the latter, its scope and relevance represent the central role of the term in both the research fields of causality and XAI.

6. Results to the research question analysis

This review allowed us to understand how the theory of causality could intertwine with the XAI literature and, specifically, which methodologies and theoretical frameworks could be adopted to approach the bridge between these two fields. We conceived three main topic clusters of studies, which are presented together with their possible sub-clusters in Fig. 4. Specifically, they embody the following perspectives:

- *critics to XAI under the causality lens*;
- *XAI for causality*;
- *causality for XAI*.

This procedure led us to identify which of the three possible perspectives is the preferable one in order to correctly combine the two areas of causality and XAI. We discuss them in Sec. 6.1 — 6.3.

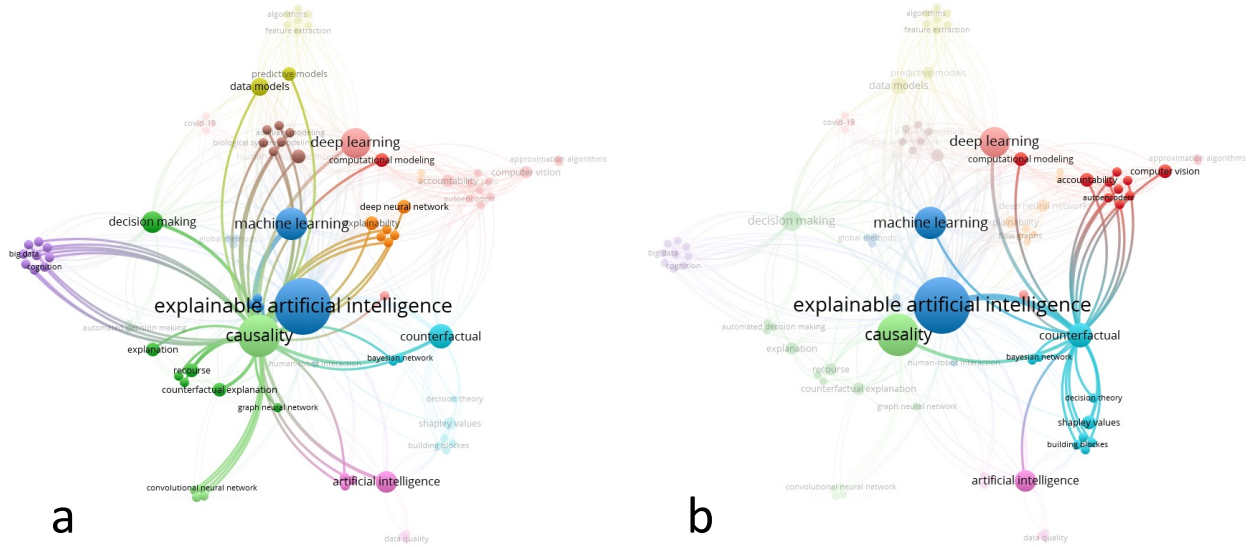

Figure 3: The isolated connections from Fig. 2 for the terms *causality* (a) and *counterfactual* (b).

6.1. Critics to XAI under the causality lens

This first perspective utilizes a causal viewpoint to identify some issues in current XAI. The focus of such papers is either: (i) to point out the inability of XAI to consider causality, (ii) to highlight the profound limitations of current (X)AI both on a methodological and a conceptual level, or (iii) to investigate the forms of the produced explanations.

6.1.1. Lack of causality

A fundamental aspect that hinders the value of classical AI models' inference and explainability methods is the lack of a foundation in the theory of causality. Indeed, classical ML and DL predictive models are based on the correlation found among training data instead of true causation. This might be of particular concern in specific fields, such as epidemiology, that have always been grounded in the theory of causation (Broadbent and Grote, 2022). Moreover, this lack of causality makes models more easily affected by adversarial attacks and less valuable for decision-making (Molnar et al., 2020). Since the parameters and predictions of classical data-driven AI models cannot be interpreted causally, they should not be used to draw causal conclusions.

As Naser (2021) points out, meeting specific performance metrics does not necessarily mean that an AI/ML model captures the physics behind a phenomenon. In other words, there is no guarantee that the found correlations map to causal relations between input data and final decisions. For this reason, determining whether such models reflect the true causal structure is crucial (Ryo et al., 2021). This inability of today's ML/DL to grasp causal links reflects also on XAI, constituting a major broad challenge to the ability of AI systems to provide sound explanations.

Hamon et al. (2022) stress how this poses serious challenges to the possibility of satisfactory, fair, and transparent explanations. Regarding the soundness of the generated explanations, Watson et al. (2022) demonstrate that they are volatile to changes in model training that are perpendicular to the classification task and model structure. This raises further questions about trust in DL models which just rely on spurious correlations that are made visible via explanation methods. Since causal explanations cannot be provided for AI yet, explanatory methods are fundamentally limited for the time being.

6.1.2. Pitfalls of (X)AI

In addition to the weaknesses due to the lack of causality, some works highlight how the fields of AI and XAI may suffer from some innate issues. On a methodological level, Molnar et al. (2022) present a number of pitfalls of

local and global model-agnostic interpretation techniques, such as in case of poor model generalization, interactions between features, or unjustified causal interpretations. At a deeper level, some researchers advocate some concerns about XAI based on its very nature. For instance, Landgrebe (2022) argues that the human inability to interpret the behavior of deep models in a more objective manner still restricts XAI methods to provide merely a partial, subjective interpretation. Undeniably, deep neural networks solve their classification in a manner that differs completely from the way humans interpret text, language, sounds, and images. For instance, convolutional neural networks (CNNs) use features of the input space to perform their classifications, which are different from those humans use. Not only is it true, but what's more, we do not understand how humans themselves classify texts or images or conduct conversations. Indeed, as of now, human or physical behavior can only be emulated by creating approximations, but approximations cannot be understood any more than complex systems can be.

Under similar considerations, Leventi-Peetz et al. (2022) study the scope and sense of explainability in AI systems. In their view, it is impossible or unwise to follow the intention of making every ML system explainable. Indeed, even domain experts cannot always provide explanations for their decisions and, furthermore, on AI systems much higher demands are made than on humans when they have to make decisions.

6.1.3. Form of the explanations

These works explore different forms, qualities, and desiderata of the explanations produced by XAI methods and their link with causality. Depending on the application domain, less accurate yet simpler explanations may be preferable to convey a proper understanding of an AI decision. For instance, in Natural Language Generation,

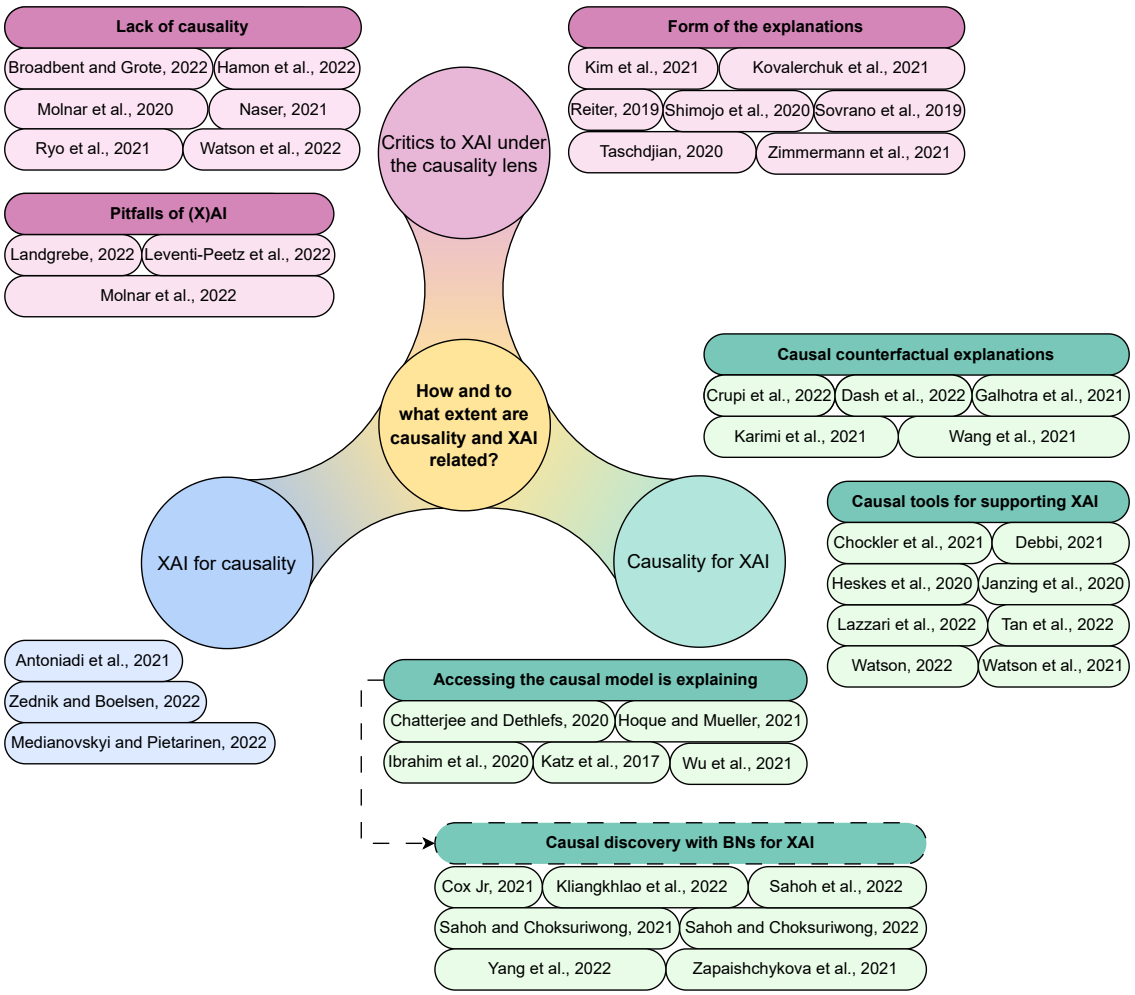

Figure 4: The included studies are classified according to the three main perspectives on how causality and XAI may be related: *Critics to XAI under the causality lens*, *XAI for causality*, and *Causality for XAI*. Next to each of them, are the possible sub-clusters.

a narrative explanation where facts are linked with causal relations is probably a better explanation for narrative-inclined individuals, even though it may not be the most accurate way to describe how the model works (Reiter, 2019). Similarly, in image classification via CNNs, a simpler visualization (e.g., natural dataset examples) may lead to an equal causal understanding of unit activation instead of using complex activation maximization approaches (Zimmermann et al., 2021).

Shimojo et al. (2020) examine what a good explanation is by drawing on psychological evidence regarding two explanatory virtues: (i) the number of causes enforced in an explanation[8], and (ii) the number of effects invoked by cause(s) in an explanation[9]. The authors report that, in a user study, the two virtues had independent effects, with a higher impact for the first one. Similarly, Kim et al. (2021) discuss several desiderata of XAI systems, among which, they should adjust explanations based on the knowledge of the explainee, to match their background knowledge and expectations. This is further stated by Kovalerchuk et al. (2021), who define as "quasi-explanations" those explanations using terms that are foreign to a certain application domain (e.g., medicine, finance, law), such as distances, weights, and hidden layers, and that consequently do make sense only for the data scientists. Kim et al. (2021) further states that explanations are considered to be *causal* when they arise from the construction of causal models, serving as the basis for recreating a causal inference chain to (i.e., a “recipe” for reconstructing) a prediction. According to the authors, intelligent systems must be able to provide causal explanations for their actions or decisions when they are critical or difficult to understand. When a causal explanation answers a "why" question, it can be referred to as a *scientific* explanation. In general, answers to questions such as “How does a personal computer work?” are not considered to be scientific explanations. Such answers are still part of a scientific discipline, but they are descriptive rather than explanatory.

Some other works argue that useful explanations are not only causal explanations but many types of non-causal explanations (e.g., semantic, contrastive, justificatory) may help (Sovrano et al., 2019). A pilot user study from Taschdjian (2020) supports this idea revealing that participants preferred causal explanations over the others only when presented in chart form, whilst they resulted as the least favorite choice when in text form.

6.2. XAI for causality

Only three papers openly support a pragmatic line of thinking according to which XAI is a basis for causal inquiry. Indeed, such works recognize certain limits of current XAI methods but approach the discussion pragmatically.

Zednik and Boelsen (2022) discuss the role of post-hoc analytic techniques from XAI in scientific exploration. The authors show that XAI techniques, such as CFEs, can serve as a tool for identifying potentially pursue-worthy experimental manipulations within a causal framework and, therefore, for recognizing causal relationships to investigate. In this regard, the authors remark on an asymmetry between the role of CFEs in *industry* and in *science*. The following two hypothetical scenarios clarify this idea:

- *industry*: a bank decides whether to accept or reject a loan application based on an AI agent. A CFE for a rejection case has revealed that doubling the client’s income would have led to the acceptance of the loan. Here, the AI agent is not trying to model reality, but it is reality itself. Indeed, a change in the client’s income would actually change the application outcome, meaning that CFEs are *perfect* guides to causal inference.
- *science*: an AI agent determines the probability of type-2 diabetes based on patients’ features. A CFE for a high-probability case has revealed that losing weight would decrease that probability. Here, the AI agent is trying to model the biological reality of the problem, but still, it remains an approximation. Indeed, it is still possible that losing weight does not actually reduce the probability of type-2 diabetes. That is to say that a change in the model’s behavior does not actually change the way the world works, but at best constitutes a changed representation of how the world could possibly work. In this light, CFEs are *imperfect* guides to causal inference.

All in all, it is just because the relevant ML models might not perfectly adhere to reality that the generated XAI explanations only foster scientific *exploration* rather than scientific *explanation*. At most, products of XAI may be thought of as starting points to study potentially causal relationships that have yet to be confirmed.

[8] This is sometimes referred to as *simplicity* and is conforming with the *Occam’s razor* principle, according to which, an event should not be explained by more causes than necessary (Jefferys and Berger, 1992).

[9] This is sometimes referred to as *scope*. Explanations with a broader scope (i.e., correctly predict more events) make humans feel more certain than explanations with a narrower one (Johnson et al., 2014).

Similarly, Medianovskyi and Pietarinen (2022) consider the outputs of the current XAI methods, such as CFEs, to be far from conclusive explanations. Rather, they are initial sketches of possible explanations and invitations to explore further. Those sketches must go through validation processes and experimental procedures before satisfactorily answering the "why" questions, long sought after by XAI.

According to the review by Antoniadi et al. (2021), XAI can help to shed some light onto causality. Indeed, since causation involves correlation, an explainable ML model could validate the results provided by causality inference techniques. Additionally, XAI can provide a first intuition of (i.e., generate hypotheses about) possible causal relationships that scientists could then test (Arrieta et al., 2020; Lipton, 2018).

6.3. Causality for XAI

This third perspective is driven by the idea that causality is propaedeutic to XAI. Indeed, these works either: (i) exploit causality-based concepts to support XAI, (ii) restore the causal foundation of CFEs, or (iii) argue that accessing the causal model of a system is intrinsically explaining the system itself.

6.3.1. Causal tools for supporting XAI

Such papers interpret the role of causality in XAI in the sense that some causal concepts, such as structural causal model (SCM) and *do*-operator (Appendix A.3) and causal metrics, may bring useful tools for explainability and for finding the causes of AI predictions. Regarding the use of **Structural causal models** to foster XAI, Reimers et al. (2020) reduce DL to a basic level and frame the constitutional structure of a CNN model into an SCM. In this setting, the random variables represent, for instance, the network's weights and the final prediction, while the functions linking the variables are the *training function* (from labeled images to the network's weights), and the *inference function* (from unlabeled images and weights to the prediction). By doing so, the authors aim to establish whether a feature is relevant to a CNN prediction by leveraging causal inference and Reichenbach's Common Cause Principle[10].

Lazzari et al. (2022), in order to predict employee turnover, utilize the concept of SCM to revisit and equip the Partial Dependence Plot (PDP)[11] method with causal inference properties. Their SCM-based PDP can now go beyond correlation-based analyses and reason about causal interventions, allowing one to test causal claims around factors. This, in turn, provides an intuitive visual tool for interpreting the results and achieving the explainability of automatic decisions.

Regarding ***do*-operator**, some authors employ this concept to bring the theory of Shapley values a step further. A fundamental component of Shapley values is to evaluate the reference distribution of dropped (i.e., 'out-of-coalition') features, which has implications on how Shapley values are estimated since this helps define the value function. Based on this distribution, the following variants of Shapley values exist (Watson, 2022; Heskes et al., 2020): *marginal* Shapley values (they ignore relations among features and are used to discover the model's decision boundary), *conditional* Shapley values (they consider feature dependencies and condition by observation), and *interventional* Shapley values. The latter was introduced by Janzing et al. (2020) who replaced conventional *conditioning by observation* with *conditioning by intervention* (*do*-operator).

Extending this concept, Heskes et al. (2020) introduce *causal* Shapley values by explicitly considering the causal relationships between the data in the real world to enhance the explanations. Using the interventional distribution is optimal when, with access to the underlying SCM, one seeks explanations for causal data-generating processes. These methods are required when seeking to use XAI for discovery and/or planning, as they seem to provide sensible, human-like explanations that incorporate causal relationships in the real world.

Finally, some other works borrow **metrics from the causal theory** to aid XAI, and, specifically, Probability of Necessity (PN) and Probability of Sufficiency (PS) from Glymour et al. (2016) and the metric of *responsibility* from Chockler and Halpern (2004). Regarding PN and PS, two works investigate their implications for XAI. Indeed, such probabilities, often addressed as "probabilities of causation", play a major role in all "attribution" questions. Watson et al. (2021) formalize the relationship between existing XAI methods and the probabilities of causation. For instance, they highlight the role of PN and PS in feature attribution methods and CFEs. Regarding the former, the authors reformulate the theory of Shapley values in their framework and show how the value function (i.e.,

[10] According to Reichenbach (1956), if two variables A and B are dependent, then there exists a variable C that causes A and B. In particular, C can be identical to A or B meaning that A causes B or B causes A.

[11] A visual tool introduced by Friedman (2001), commonly used for model-agnostic XAI, that shows the marginal effect of one feature on the predicted outcome of a system.

the payoff associated with a feature subset) precisely corresponds to the PS of a factor. Regarding the latter, the authors rewrite the CFE optimization problem with an objective based on the PS of the factor with respect to the opposite of the outcome. Moreover, Tan et al. (2022) borrow PN and PS and adapt them to evaluate the necessity and sufficiency of the explanations extracted for a graph neural network (GNN). This makes it possible to conduct a quantitative evaluation of GNN explanations even without ground-truth explanations for real-world graph datasets.

On the other hand, regarding the metric of responsibility, Chockler et al. (2021) propose DC-CAUSAL, a greedy, compositional, perturbation-based approach to computing explanations for image classification. It leverages causal reasoning in its feature masking phase with the goal of finding causes in input images by causally ranking parts of the input image (i.e., superpixels) according to their responsibility for the classification. In addition to responsibility, Debbi (2021) borrows from Chockler and Halpern (2004) the concept of blame to compute visual explanations for CNN decisions. The author abstracts the CNN model into a causal model by virtue of similarity in a hierarchical structure, and filters are considered as actual causes for a decision. First, each filter is assigned a degree of responsibility (i.e., weight) as a measure of its importance to the related class. Then, the responsibilities of these filters are projected back to compute the blame for each region in the input image. The regions with highest blame are returned then as the most important explanations.

PN is the probability that the garden would not have got wet had the sprinkler not been activated ($Y_0 = 0$), given that, in fact, the garden did get wet ($Y = 1$) and the sprinkler was activated ($X = 1$). Mathematically, this becomes: $PN = P(Y_0 = 0|X = 1, Y = 1)$. In other words, this probability quantifies to what extent activating the sprinkler is necessary to get the garden wet, and consequently if other factors (e.g., rain) may have caused the wet garden.

PS is the probability that the garden would have got wet had the sprinkler been activated ($Y_1 = 1$), given that the sprinkler had not in fact been activated ($X = 0$), and the garden did not get wet ($Y = 0$). Mathematically, this becomes: $PS = P(Y_1 = 1|X = 0, Y = 0)$. In other words, this probability quantifies to what extent activating the sprinkler is sufficient to wet the garden, and consequently, if there may exist scenarios (e.g., hardware malfunctioning) where activating the sprinkler does not wet the garden.

Responsibility is a quantification of causality, attributing to each actual cause its degree of responsibility $\frac{1}{1+k}$, which is based on the size k of the smallest contingency feature set required to obtain a change in the prediction (i.e., creating a counterfactual dependence). The degree of responsibility is always between 0, for variables that have no causal influence on the outcome ($k \to \infty$), and 1, for counterfactual causes ($k = 1$). Responsibility extends the actual causality framework of Halpern and Pearl (2005) .

6.3.2. Causal counterfactual explanations

As noted in Sec. 3.3, the *counterfactual* concept seems to belong both to the XAI literature and to the causality literature. Some authors remark on how CFEs and CF are two separate concepts (Crupi et al., 2022) and, strictly speaking, some would not even call the former *counterfactuals*, precisely to contrast the causal perspective (Dash et al., 2022). Interestingly, however, these two seemingly separate concepts may be bridged in what we could name structural causal explanations. Indeed, the papers in this sub-cluster present methods for generating CF based on their formal causal definition, restoring the causal underpinning to CFEs by using the concept of SCM and Pearl's CF three-step "recipe" (Appendix A.3).

In their quest to explain an image classifier's output and its fairness using counterfactual reasoning, Dash et al. (2022) propose IMAGECFGEN, a system that combines knowledge from an SCM over image attributes and uses an inference mechanism in a generative adversarial network-like framework to generate counterfactual images. The proposed architecture directly maps to Pearl's three steps: (i) for *abduction*, an encoder infers the latent vector of an input image coupled with its attributes; (ii) for *action*, a subset of desired attributes is changed and, accordingly, the values of their descendants in the SCM are updated; (iii) for *prediction*, a generator takes the latent vector together with the modified set of attributes and produces a counterfactual image. A subset of work focuses on a specific aim of the XAI research tightly bound with counterfactual reasoning, i.e., *recourse*. Recourse can be seen as the act of recommending a set of feasible actions to assist an individual to achieve a desired outcome. Karimi et al. (2021) argue that the conventional, non-causal CFEs are unable to convey a relevant recourse to the end-user of AI algorithms since they help merely understand rather than act (i.e., inform an individual to where they need to get, but not how to get there). Shifting from explanation to *minimal intervention*,

the authors leverage causal reasoning (i.e., tools of SCMs and structural interventions) to incorporate knowledge of the causal relationships governing the world in which actions will be performed. This way, the authors are able to compute what they refer to as *structural CF* by performing the *abduction-action-prediction* steps and provide *algorithmic recourse*. Galhotra et al. (2021) introduce LEWIS, a principled causality-based approach for explaining black-box decision-making systems. They propose to achieve *counterfactual recourse* by solving an optimization problem that searches for minimal interventions on a pre-specified set of actionable variables that have a high probability of producing the algorithm's desired future outcome. Notably, the authors propose a GUI that implements LEWIS, of which they show a demo in Wang et al. (2021). Crupi et al. (2022) also contribute to the recourse objective by proposing CEILS, a new post-hoc method to generate causality-grounded CFEs and recommendations. It involves the creation of an SCM in the latent space, the generation of causality-grounded CFEs, and their translation to the original feature space.

6.3.3. Accessing the causal model is explaining

Part of the work relates to the common thought that accessing the causal model of a system intrinsically explains the system itself. Under this view, two fundamental observations are supported:

- when a model is built on a causal structure, it is inherently an interpretable model;
- making the inner workings of a causal model directly observable, such as through a directed acyclic graph (DAG) (Appendix A.1), makes the model inherently interpretable.

Much of the causality theory focuses on explaining observed events, that is, inferring causes from effects. According to its retrospective attribution, causality lies at the heart of explanation-based social constructs such as explainability and, therefore, causal reasoning is an important component of XAI (Wu et al., 2021).

Ibrahim et al. (2020) try to fill the lack in the causality literature of automatic and explicit operationalizations to enable explanations. The authors propose an extensible, open-source, interactive tool (Actual Causality Canvas) able to implement three main activities of causality (causal modeling, context setting, and reasoning) in a unifying framework. According to the authors, what Canvas can provide, through answers to causal queries, largely overlaps with the ultimate goal of XAI, which is providing the end-user with explanations of why particular factors occurred. Hoque and Mueller (2021) propose Outcome Explorer, an interactive framework guided by causality, that allows expert and non-expert users to select a dataset, choose a causal discovery (CD) algorithm for structure discovery (Appendix A.2), generate (and eventually refine) a causal diagram, and interpret it by setting values to the input features to observe the changes in the outcome. Katz et al. (2017) propose an XAI system that encodes the causal relationships between actions, intentions, and goals from an autonomous system and explains them to a human end-user with a cause-effect reasoning mechanism (i.e., causal chains). Chatterjee and Dethlefs (2020) exploit the representational power of CNNs with attention, to discover causal relationships across multiple features from observed time-series and historical error logs. The authors believe causal reasoning can enhance the reliability of decision support systems making them more transparent and interpretable.

A subset of publications sees CD as the most appropriate way of operationalizing the idea that accessing the causal model of a system intrinsically explains the system itself. In this regard, all of them utilize **Bayesian networks (BNs)** (Appendix A.2) as the methodological tool. Since establishing unique directions for edges based on passive evidence alone may be challenging, knowledge-based constraints can help orient arrows to reflect causal interpretations (Cox Jr, 2021). In line with this, some works perform CD with BNs in a mixed approach: on the one hand, they leverage knowledge from domain-experts to outline the causal structure of the system (i.e., finding nodes and related edges); on the other hand, they fit the model parameters on observed, real-world data.

Sahoh and Choksuriwong (2022) propose a new system to support emergency management (e.g., terrorist events) based on the Deep Event Understanding perspective, introduced in an earlier work of theirs (Sahoh and Choksuriwong, 2021). Deep Event Understanding aims to model expert knowledge based on the human learning process and offers explanation abilities that mimic human reasoning. Their model utilizes BNs based on social sensors as an observational resource (i.e., text data from Twitter), with prior knowledge from experts to infer and interpret new information. Their approach helps in recognition of an emergency event and in the uncovering of its possible causes, contributing to the explanation of "why" questions for decision-making.

Sahoh et al. (2022) propose discovering cause-effect ML models for indoor thermal comfort in Internet of Things (IoT) applications. They employ five different CD algorithms and show how these may converge to the ground-truth SCM of the problem variables obtained from domain experts. Kliangkhlao et al. (2022) introduce

a BN model for agricultural supply chain applications, initially constructed from causal assumptions from expert qualitative knowledge, which conventional ML cannot reasonably conceive. Therefore, a data-driven approach using observational evidence is employed to encode these causal assumptions into quantitative knowledge (i.e., parameter fitting). The authors report their system constitutes a framework that is able to provide reasonable explanations of events for decision-makers.

In Zapaishchykova et al. (2021) the authors leverage the respective strengths of DL for feature extraction and BNs for causal inference, achieving an automatic and interpretable system for grading pelvic fractures from CT images. The BN model is constructed upon variables extracted with the neural network, together with a variable from the clinical practice (i.e., patient age). By doing so, the authors believe that the framework provides a transparent inference pipeline supplying fracture location and type, by establishing causal relationships between trauma classification and fracture presence.

Yang et al. (2022) propose a new process monitoring scheme based on BNs to explain (diagnose) a detected fault and promote decision-making. Their system allows the identification of the root cause (i.e., labeling the abnormal variables) so that the result of the analysis can be linked to the repairing action, reducing the investigation time. Among one of their use cases, the authors fit a BN model on observed, real-world data for manufacturing fault events. During this CD process, they employ a blacklist obtained from domain experts to exclude causally-unfeasible relationships.

7. Results of software tools collection

We hereby present a summary of the main data mining software tools collected within the cohort of papers. Table 2 comprises tools for performing CD with BNs (i.e., PySMILE[12], CausalNex[13], bnlearn[14], CompareCausalNetworks[15], CaMML[16], Python Causal Discovery Toolbox[17], and Tetrad[18]), creating and analysing SCMs (i.e., IBM® SPSS® Amos[19], lavaan[20], and semopy[21]), and editing and analyzing DAGs (i.e., DAGitty[22]). We believe this list of software solutions may be of interest to AI practitioners in helping them save valuable time when choosing the right tool to automate causal tasks.

The most popular choice is an open-source license type, and this reflects the great interest in sharing code and information across the AI research community. The first benefit of that is flexibility. Researchers often need to access the source code of software implementations to eventually customize its functionalities according to a desired (yet not implemented) purpose. This would be highly unfeasible with closed and commercial software. Another advantage of having open-source implementations is software security. According to Linus's law, "given enough eyeballs, all bugs are shallow" (Raymond, 1999). That is, when all the source code for a project is made open to professionals worldwide, it is more likely that security checks could discover eventual flaws.

Furthermore, Table 2 shows that the CLI is the preferred frontend interface across such solutions. This aspect also reflects the AI research community viewpoint. Opting for CLI over the GUI brings some advantages, such as faster and more efficient computing, easier handling of repetitive tasks, lighter memory usage, and availability of the history of commands. On the other hand, using CLI involves a steeper learning curve associated with memorizing commands and complex arguments, together with the need for correct syntax. This may explain why GUI is preferred in cases where the end-user does not have a programming background. Typical examples of that include physicians in healthcare facilities or product managers in finance companies, who prefer, in general, a more user-friendly product.

[12] https://www.bayesfusion.com/smile/
[13] https://causalnex.readthedocs.io/en/latest
[14] https://www.bnlearn.com
[15] https://cran.r-project.org/web/packages/CompareCausalNetworks/
[16] https://bayesian-intelligence.com/software/
[17] https://fentechsolutions.github.io/CausalDiscoveryToolbox/html/index.html
[18] https://htmlpreview.github.io/?https:///github.com/cmu-phil/tetrad/blob/development/docs/manual/index.html
[19] https://www.ibm.com/products/structural-equation-modeling-sem
[20] https://cran.r-project.org/web/packages/lavaan/index.html
[21] https://semopy.com/
[22] http://www.dagitty.net/

Table 2: Software tools within the cohort of papers useful to automate causal tasks. BSD: Berkeley Software Distribution, CD: causal discovery, CLI: common line interface, GPL: General Public License, GUI: graphical user interface.

Name	**License type**	**Release paper**	**Frontend interface**	**Main purpose**
bnlearn	Open-source (GPL)	Scutari (2010)	CLI (R)	BNs for CD
CaMML by Bayesian Intelligence Pty Ltd	Open-source (BSD)	n.a.	CLI (Bash) and GUI	BNs for CD
CausalNex by QuantumBlack, AI by McKinsey	Open-source (Apache 2.0)	n.a.	CLI (Python)	BNs for CD
CompareCausalNetworks	Open-source (GPL)	Heinze-Deml et al. (2018)	CLI (R)	BNs for CD
DAGgity	Open-source (GPL)	Textor et al. (2016)	CLI (R) and GUI	Create and analyze causal diagrams
IBM SPSS Amos by IBM Corp.	Commercial	n.a.	GUI	Create and analyze SCMs
lavaan	Open-source (GPL)	Rosseel (2012)	CLI (R)	Create and analyze SCMs
PySMILE by BayesFusion LLC	Commercial	n.a.	CLI (Python)	BNs for CD
Python Causal Discovery Toolbox by Fentech	Open-source (MIT)	*Kalainathan et al. (2020)*	CLI (Python)	BNs for CD
semopy	Open-source (MIT)	Igolkina and Meshcheryakov (2020) Meshcheryakov et al. (2021)	CLI (Python)	Create and analyze SCMs
Tetrad	Open-source (GPL)	Ramsey et al. (2018)	GUI	BNs for CD

8. Conclusion

The concepts of causation and explanation have always been part of human nature, from influencing the philosophy of science to impacting the data mining process for knowledge discovery of today's AI. In this study, we investigated the relationship between causality and XAI, by exploring the literature from both theoretical and methodological viewpoints, to reveal whether a dependent relationship between the two research fields exists. We provided a unified view of the two fields by highlighting which methodologies could be adopted to approach the bridge between these two fields and uncovering possible limitations. As a result of the analysis, we found and formalized three main perspectives.

The *Critics to XAI under the causality lens* perspective analyses how the lack of causality is one of the major limitations of current (X)AI approaches as well as the "optimal" forms to provide explanations. Regarding the former, traditional AI systems are only able to detect correlation instead of true causation, which affects the robustness of models against adversarial attacks and of the produced explanations. This is of concern since pure associations are not enough to accurately describe causal effects. Regarding the latter, optimal explanations may be characterized by being expressed according to the explainee's knowledge and domain terminology and being able to explain many effects with few causes. However, it is debated whether causal explanations (i.e., causal inference chains to a prediction) are the only useful ones in the XAI landscape. This first perspective states the problem and serves as a watch out.

The *XAI for causality* perspective openly claims that XAI may be a basis for further causal inquiry. Despite the recognized limits of XAI explanations, they may be pragmatically thought of as starting points to generate hypotheses about possible causal relationships that scientists could then confirm. That is, XAI can only foster scientific exploration, rather than scientific explanation. Although underrepresented in the final cohort, this perspective suggests a really thoughtful idea in our opinion.

The *Causality for XAI* perspective supports the idea that causality is propaedeutic to XAI. This is realized in three manners. First, some causal concepts (i.e., SCM and *do*-operator) are leveraged to revisit existing XAI methods to empower them with causal inference properties. Second, the formal causal definition of CF (Sec. 3.3) is invoked to generate causal-CFEs using the SCM tool, which may also enable recourse. Third, and lastly, it is argued that, when a model is built on a causal structure, it is inherently an interpretable model. In a related way, making the inner workings of a causal model directly observable (e.g., through a DAG) makes the model inherently interpretable.

Among the three main perspectives, we believe *Causality for XAI* to be the most promising one. Naturally, it comes with limitations. Much work in causal modeling is based on specific and (by far) non-unique causal views of the problems at hand. Interventions and CF make sense as long as the specified causal graph makes sense, which may hinder the generalization of their results. Overall, their causal claims depend on strong and often non-testable assumptions about the underlying data-generating process. On the other hand, however, this may be in line with what already happens in our life, and we should not request from AI more than we request from human beings. Another weak point is the interpretability of a causal model with hundreds of variables. In this scenario, a DAG would encode too much information and the complexity of the underlying SCM would rise exponentially with the number of modeled variables. This, however, is common to other simpler and more traditional approaches such as Decision Trees with hundreds of nodes.

We acknowledge three main limitations that may have led us to miss publications that could have potentially been included in the review: (i) the exclusion of non-peer-reviewed e-prints, (ii) the usage of only four databases, and (iii) not having extracted any references from the collected papers to enrich our search. The latter was motivated by the fact that, this being an unexplored field, the papers we collected were sufficient and significant enough to produce a first scenario. Obviously, as with any human-made assignment, the search process for relevant material may have been affected by the cognitive bias of the authors, who have brought their knowledge and assumptions in the study.

We believe our results could be useful to a wide spectrum of readers, from upper-level undergraduate students to research managers in the industry, and have implications for practice, policy, and future research. Indeed, having a clear view of how the two concepts of causality and XAI are related can benefit both areas individually, as well as the joint research field. Considering our conceptual framework, future publications may be framed in a precise and rigorous way and have the potential to expand (or generate new flavors of) one of the identified perspectives.

All in all, our work disclosed how causality and XAI may be related in a profound way. In our opinion, the *Causality for XAI* perspective has great potential to produce significant scientific results and we expect the field to flourish the most soon.

9. Funding Information

This work was partially funded by: the European Union's Horizon 2020 research and innovation programme under grant agreement No 952159 (ProCAncer-I), and the Regional Projects PAR FAS Tuscany - PRAMA and NAVIGATOR. The funders had no role in the design of the study, collection, analysis and interpretation of data, or writing of the manuscript.

Appendix A. Background notions

Appendix A.1. Directed acyclic graphs

From graph theory, a *graph* consists of a set V of vertices (i.e., variables) and a set E of edges (i.e., relationships) that connect some pairs of vertices. A graph is *directed* when all the edges are directed (i.e., marked by a single arrowhead). In a directed graph, an edge goes from a *parent* node to a *child* node. A *path* in a directed graph is a sequence of edges such that the ending node of each edge is the starting node of the next edge in the sequence (e.g., nodes A, B, D in Fig. A.5). A *cycle* is a path in which the starting node of its first edge equals the ending node of its last edge (e.g., nodes C, E, F in Fig. A.5a), and this represents mutual causation or feedback processes. When a directed graph does not include directed cycles, it is called a *directed acyclic graph* (DAG), and much of the discussion of causality and qualitative modeling is occupied by it (Pearl, 2009).

Appendix A.2. Bayesian networks

A Bayesian network (BN) is a probabilistic graphical model that consists of two parts, a qualitative one based on a DAG, representing a set of variables and their dependencies, and a quantitative one based on local probability distributions for specifying the probabilistic relationships (Pearl, 1985). Let $\mathbf{X} = [X_1, X_2, \ldots, X_m]$ be a data matrix with n samples and m variables. In the DAG $G = (V, E)$ of a BN, each node $V_k \in V$ represents the random variable X_k in $\mathbf{X}$, $k \in \{1, 2, \ldots, m\}$, and each edge $e \in E$ describes the conditional dependency between pairs of variables. The absence of an edge implies the existence of conditional independence.

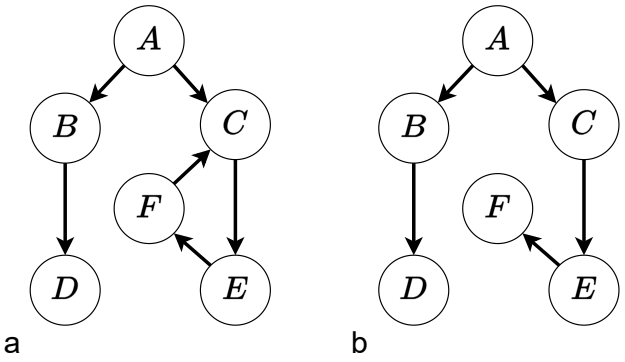

Figure A.5: Examples of directed graphs: (a) directed cyclic graph, (b) directed acyclic graph (DAG).

The structure of the DAG can be constructed either manually, with expert knowledge of the underlying domain (knowledge representation), or automatically learned from a large dataset. In this regard, causal discovery (CD) denotes a broad set of methods aiming at retrieving the topology of the causal structure governing the data-generating process, using the data generated by this process. CD algorithms are commonly divided into two families: *constraint*-based and *score*-based.

Constraint-based methods begin with fully-connected edges between random variables and leverage conditional independence tests to identify a set of edge constraints for the graph. By deleting relations if there is no statistical significance between variables, they narrow down the candidate graphs that explain the data and then try to determine the direction of the found relationships. Popular examples include the PC algorithm (Spirtes and Glymour, 1991), assuming no latent confounders (i.e., variables that are not directly observed but interact with the observables), and the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2000), whose results are asymptotically correct even in the presence of (possibly unknown) confounders. Although constraint-based methods can handle various types of data distributions and causal relations, they do not necessarily provide complete causal information, since they output a set of causal structures satisfying the same conditional independence.

On the other hand, *score*-based methods iteratively generate candidate graphs, assign them a relevance score to evaluate how well each one explains the data (i.e., "model fit"), and select the best one. Since enumerating (and scoring) every possible graph among the given variables is computationally expensive, these algorithms apply greedy heuristics to restrict the number of candidates. Among them, Greedy Equivalence Search (GES) (Chickering, 2002) is a well-known two-phase procedure that directly searches over the space of equivalence classes. Starting with an empty graph, at each step, it adds currently needed edges (if that increases fit), and then eliminates unnecessary edges in a pattern.

Regarding the quantitative part of which a BN consists, the local probability distributions can be either *marginal*, for nodes without parents (root nodes), or *conditional*, for nodes with parents. In the latter case, the dependencies are quantified by Conditional Probability Tables (CPTs) for each node given its parents in the graph. These quantities can be estimated from data in a process known as Parameter Estimation, two popular examples of which are the Maximum Likelihood approach and the Bayesian approach.

Once the DAG and CPTs are determined, a BN is fully specified and compactly represents the Joint Probability Distribution (JPD). An example of a fully specified BN is shown in Fig. A.6. According to the *Markov condition*, each node is conditionally independent of its non-descendants, given its parents. As a result, the JPD can be expressed in a product form:

$$p(X_1, X_2, \ldots, X_m) = \prod_{k=1}^{m} p(X_k | \mathbb{X}_{pa(k)}) \tag{A.1}$$

Where $\mathbb{X}_{pa(k)}$ is the set of parent nodes of X_k and $p(X_k|\mathbb{X}_{pa(k)})$ is the conditional probability of X_k given $\mathbb{X}_{pa(k)}$. Thus, such a BN can be used for predictions and inference, that is, computing the posterior probabilities of any subset of variables given evidence about any other subset.

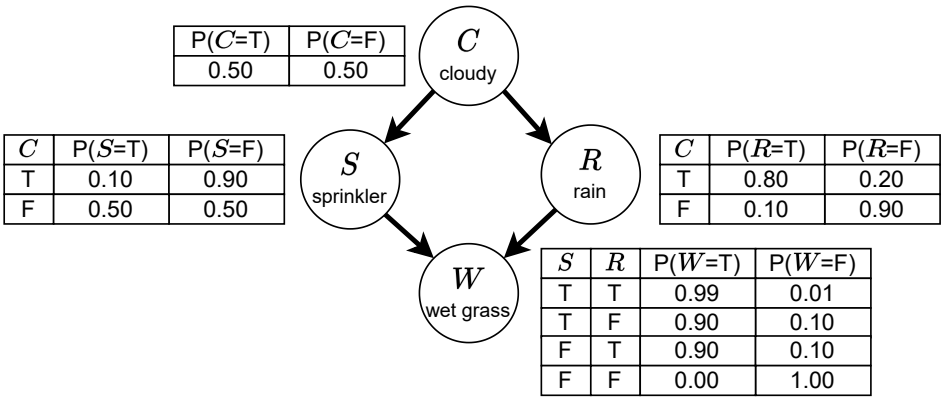

Figure A.6: Example of a fully specified BN which models the probability of observing wet grass. In this (simplified) real-world scenario, grass can be wet either by turning on a sprinkler or by rainfall, and both can be influenced by the presence of clouds in the sky.

Appendix A.3. Structural Causal Models

Consider the set **X** of variables associated with the vertices of a DAG. When each of them appears on the left-hand side (i.e., the dependent variable) of an equation of the type:

$$X_k = f_k(\mathbb{X}_{pa(k)}, U_k), \quad k = 1, \ldots, m \tag{A.2}$$

that represents an autonomous mechanism, then the model is called a *structural causal model* (SCM) (Pearl, 2009; Schölkopf et al., 2021). In this equation, f_k represents a deterministic function depending on the X_k's parents in the graph (i.e., $\mathbb{X}_{pa(k)}$), and on U_k, which represents the exogenous variables (i.e., errors or noises due to omitted factors). These noises are assumed to be jointly independent, and hence ensure that each structural equation can represent a general conditional distribution $p(X_k|\mathbb{X}_{pa(k)})$, Recursively applying Eq. A.2, when the distributions of $U = \{U_1, \ldots, U_m\}$ are specified, allows the computation of the entailed observational joint distribution $p(X_1, X_2, \ldots, X_m)$, which, in turn, can be canonically factorized into Eq. A.1. The advantages of using the SCM language include modeling unobserved variables (i.e., latent variables and counfounders), easily formalizing interventions, and computing CF. Interventions and CF are defined through a mathematical concept called *do*-operator, which simulates physical interventions by modifying a subset of structural equations (e.g., replacing them with a constant), while keeping the rest of the model unchanged. Specifically, to compute the probability of CF, Pearl proposes a three-step procedure. Given a known SCM M over the set **X** of variables, let $x_{factual} = [X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m]$ be the evidence. To compute the probability of a counterfactual instance $x_{counterfactual}$, one needs to:

1. *abduction*: infer the values of exogenous variables in U for $x_{factual}$, i.e., calculate $P(U|x_{factual})$;
2. *action*: intervene on $X = x_{factual}$ by replacing (some of) the equations by the equations $X = x_{counterfactual}$, where $x_{counterfactual} = [X_1 = x'_1, X_2 = x'_2, \ldots, X_m = x'_m]$, and thus obtain a new SCM M';
3. *prediction*: use M' to compute the probability of $P(x_{counterfactual}|x_{factual})$.

Appendix B. Study selection process

Although we did not apply any temporal constraint to the search, we adopted some exclusion criteria in the process. We excluded works that were not written in English, articles from electronic preprint archives (e.g., ArXiv[23]), book chapters, and theses. In addition, we excluded too-short papers and/or papers of poor quality that hindered our ability to extract data meaningfully. We also deemed off-topic those papers that considered causality in the common and everyday sense of the term, not based on theoretical definitions. Indeed, they frequently present few occurrences of the causal domain terms, which were often either poorly contextualized or only present in the abstract/keywords of the article.

[23] `https://arxiv.org`

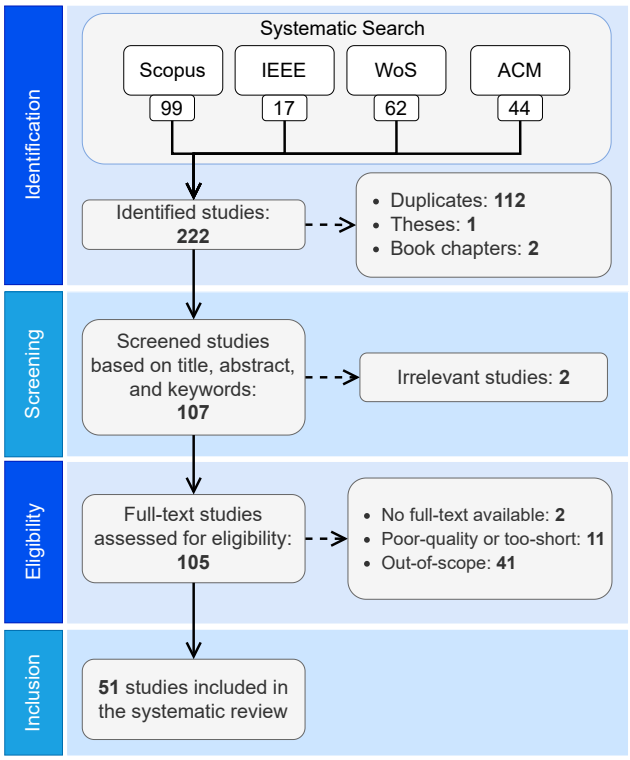

Figure B.7: Flowchart of the study collection process, from identification, through screening, to eligibility and inclusion.

Regarding information sources, we selected Scopus, IEEE, Wos, and ACM because they cover a comprehensive range of AI works and provide powerful interfaces for retrieving the required data with limited restrictions. Conversely, we excluded Google Scholar[24], SpringerLink[25], and Nature[26] since they do not allow to formulate the query string with the same level of detail as the selected databases do, and, on the other hand, we excluded PubMed[27], since it provides this capability, but its coverage is restricted solely to the medical field.

As for the search strategy on the specified databases, the use of the wildcard made word-matching easier. For instance, **causal**$*$ matched terms like *causal* and *causality*, while **expla**$*$ matched terms such as *explanation(s)*, *explainable*, *explainability*, *explaining*, and *explained*.

On July 14, 2022, we utilized the research query on the four databases for the first time. We collected the retrieved publications and started analyzing them. Then, on September 5, 2022, we repeated the search in the same settings. This allowed us to refine our cohort of papers with new works that have been published in the meanwhile, therefore enriching our analyses. In general, although we utilized the same research query across the four databases (Sec. 4), the actual query string was edited according to the specific syntax of each of them. In this regard, those strings are shown in Tab. B.3.

Fig. B.7 shows the process of identification, screening, eligibility, and inclusion of articles in our work.

From the search, we obtained the following number of records from the four databases: 99 (Scopus), 17 (IEEE), 62 (WoS), and 44 (ACM). As a result, we collected a total of 222 publications. Upon extraction of query results from the databases, we operated the identification phase. For the retrieved records, we extracted the BibTeX files and uploaded them into a popular reference manager application by Elsevier, namely Mendeley[28], desktop version 1.19.8. We then utilized its *Check for Duplicates* feature to perform duplicate removal. Then, we removed one thesis and two book chapters, according to the defined exclusion criteria. After these steps, the joint

[24] https://scholar.google.com/
[25] https://link.springer.com/
[26] https://www.nature.com/siteindex
[27] https://pubmed.ncbi.nlm.nih.gov/
[28] https://www.mendeley.com/

Table B.3: Query strings used for each database. AB, ABS: abstract; AK, KEY: keywords; TI: title.

Database	Query string
Scopus	`TITLE-ABS-KEY(causal*) AND TITLE-ABS-KEY(expla*) AND TITLE-ABS-KEY(xai OR "explainable artificial intelligence" OR "explainable ai") AND TITLE-ABS-KEY("machine learning" OR ai OR "artificial intelligence" OR "deep learning")`
Web of Science	`(TI=causal* OR AB=causal* OR AK=causal*) AND (TI=expla* OR AB=expla* OR AK=expla*) AND (TI=(xai OR "explainable artificial intelligence" OR "explainable ai") OR AB=(xai OR "explainable artificial intelligence" OR "explainable ai") OR AK=(xai OR "explainable artificial intelligence" OR "explainable ai")) AND (TI=("machine learning" OR ai OR "artificial intelligence" OR "deep learning") OR AB=("machine learning" OR ai OR "artificial intelligence" OR "deep learning") OR AK=("machine learning" OR ai OR "artificial intelligence" OR "deep learning"))`
IEEE Xplore	`("Document Title":causal* OR "Abstract":causal* OR "Author Keywords":causal*) AND ("Document Title":expla* OR "Abstract":expla* OR "Author Keywords":expla*) AND ("Document Title":xai OR "Document Title":"explainable artificial intelligence" OR "Document Title":"explainable ai" OR "Abstract":xai OR "Abstract":"explainable artificial intelligence" OR "Abstract":"explainable ai" OR "Author Keywords":xai OR "Author Keywords":"explainable artificial intelligence" OR "Author Keywords":"explainable ai") AND ("Document Title":"machine learning" OR "Document Title":ai OR "Document Title":"artificial intelligence" OR "Document Title":"deep learning" OR "Abstract":"machine learning" OR "Abstract":ai OR "Abstract":"artificial intelligence" OR "Abstract":"deep learning" OR "Author Keywords":"machine learning" OR "Author Keywords":ai OR "Author Keywords":"artificial intelligence" OR "Author Keywords":"deep learning")`
ACM	`(Title:causal* OR Abstract:causal* OR Keyword:causal*) AND (Title:expla* OR Abstract:expla* OR Keyword:expla*) AND (Title:xai OR Title:"explainable artificial intelligence" OR Title:"explainable ai" OR Abstract:xai OR Abstract:"explainable artificial intelligence" OR Abstract:"explainable ai" OR Keyword:xai OR Keyword:"explainable artificial intelligence" OR Keyword:"explainable ai") AND (Title:"machine learning" OR Title:ai OR Title:"artificial intelligence" OR Title:"deep learning" OR Abstract:"machine learning" OR Abstract:ai OR Abstract:"artificial intelligence" OR Abstract:"deep learning" OR Keyword:"machine learning" OR Keyword:ai OR Keyword:"artificial intelligence" OR Keyword: "deep learning")`

output was 107 publications.

During the screening phase, we examined independently the resulting works by title, abstract, and keywords to verify and ensure that proper results were retrieved by the query. Whenever both authors deemed a paper irrelevant, it was discarded from the cohort. Specifically, two publications were hereby discarded. Instead, publications for which the authors agreed on the inclusion, together with those on which they disagreed, passed to the next phase.

Next, in the eligibility phase, we first checked for the availability of full-text manuscripts for the records in the cohort. We excluded two studies as we could not access their full text. We then jointly analyzed the available full-text publications to remove papers that were clearly out of scope, together with poor-quality or too-short papers. As a result, we identified 11 poor-quality or too-short papers and 41 out-of-scope works. Lastly, once we reached a common decision for each of the publications, we collected the final cohort of studies to be included in the review.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE access 6, 52138–52160.

Alonso, J.M., Casalino, G., 2019. Explainable artificial intelligence for human-centric data analysis in virtual learning environments, in: Higher Education Learning Methodologies and Technologies Online: First International Workshop, HELMeTO 2019, Novedrate, CO, Italy, June 6-7, 2019, Revised Selected Papers 1, Springer. pp. 125–138.

Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C., 2021. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Applied Sciences 11, 5088.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58, 82–115.

Berrevoets, J., Kacprzyk, K., Qian, Z., van der Schaar, M., 2023. Causal deep learning. arXiv preprint arXiv:2303.02186 .

Broadbent, A., Grote, T., 2022. Can robots do epidemiology? machine learning, causal inference, and predicting the outcomes of public health interventions. Philosophy & Technology 35, 1–22.

Bussmann, N., Giudici, P., Marinelli, D., Papenbrock, J., 2021. Explainable machine learning in credit risk management. Computational Economics 57, 203–216.

Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8, 832.

Chatterjee, J., Dethlefs, N., 2020. Temporal causal inference in wind turbine scada data using deep learning for explainable ai, in: Journal of Physics: Conference Series, IOP Publishing. p. 022022.

Chickering, D.M., 2002. Optimal structure identification with greedy search. Journal of machine learning research 3, 507–554.

Chockler, H., Halpern, J.Y., 2004. Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research 22, 93–115.

Chockler, H., Kroening, D., Sun, Y., 2021. Explanations for occluded images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1234–1243.

Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J., 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. Information Fusion 81, 59–83.

Cox Jr, L.A., 2021. Information structures for causally explainable decisions. Entropy 23, 601.

Crupi, R., González, B.S.M., Castelnovo, A., Regoli, D., 2022. Leveraging causal relations to provide counterfactual explanations and feasible recommendations to end users., in: ICAART (2), pp. 24–32.

Dash, S., Balasubramanian, V.N., Sharma, A., 2022. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 915–924.

Debbi, H., 2021. Causal explanation of convolutional neural networks, in: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21, Springer. pp. 633–649.

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Communications of the ACM 63, 68–77.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

Galhotra, S., Pradhan, R., Salimi, B., 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals, in: Proceedings of the 2021 International Conference on Management of Data, pp. 577–590.

Glymour, M., Pearl, J., Jewell, N.P., 2016. Causal inference in statistics: A primer. John Wiley & Sons.

Gordon, L., Grantcharov, T., Rudzicz, F., 2019. Explainable artificial intelligence for safe intraoperative decision support. JAMA surgery 154, 1064–1065.

Granger, C.W., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: journal of the Econometric Society , 424–438.

Greenland, S., Mansournia, M.A., 2015. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. European journal of epidemiology 30, 1101–1110.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51, 1–42.

Gunning, D., Aha, D., 2019. Darpa's explainable artificial intelligence (xai) program. AI magazine 40, 44–58.

Hall, S.W., Sakzad, A., Choo, K.K.R., 2022. Explainable artificial intelligence for digital forensics. Wiley Interdisciplinary Reviews: Forensic Science 4, e1434.

Halpern, J.Y., Pearl, J., 2005. Causes and explanations: A structural-model approach. part i: Causes. The British journal for the philosophy of science .

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., De Hert, P., 2022. Bridging the gap between ai and explainability in the gdpr: towards trustworthiness-by-design in automated decision-making. IEEE Computational Intelligence Magazine 17, 72–85.

Hankinson, R.J., 1998. Cause and explanation in ancient Greek thought. Clarendon Press.

Heinze-Deml, C., Maathuis, M.H., Meinshausen, N., 2018. Causal structure learning. Annual Review of Statistics and Its Application 5, 371–391.

Hempel, C.G., Oppenheim, P., 1948. Studies in the logic of explanation. Philosophy of science 15, 135–175.

Heskes, T., Sijben, E., Bucur, I.G., Claassen, T., 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. Advances in neural information processing systems 33, 4778–4789.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H., 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9, e1312.

Hoque, M.N., Mueller, K., 2021. Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making. IEEE Transactions on Visualization and Computer Graphics .

Hume, D., 2003. A treatise of human nature. Courier Corporation.

Ibrahim, A., Klesel, T., Zibaei, E., Kacianka, S., Pretschner, A., 2020. Actual causality canvas: a general framework for explanation-based socio-technical constructs, in: ECAI 2020. IOS Press, pp. 2978–2985.

Igolkina, A.A., Meshcheryakov, G., 2020. semopy: A python package for structural equation modeling. Structural Equation Modeling: A Multidisciplinary Journal 0, 1–12. URL: https://doi.org/10.1080/10705511.2019.1704289, doi:10.1080/10705511.2019.1704289, arXiv:https://doi.org/10.1080/10705511.2019.1704289.

Janzing, D., Minorics, L., Blöbaum, P., 2020. Feature relevance quantification in explainable ai: A causal problem, in: International Conference on artificial intelligence and statistics, PMLR. pp. 2907–2916.

Jefferys, W.H., Berger, J.O., 1992. Ockham's razor and bayesian analysis. American scientist 80, 64–72.

Jiménez-Luna, J., Grisoni, F., Schneider, G., 2020. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2, 573–584.

Johnson, S., Johnston, A., Toig, A., Keil, F., 2014. Explanatory scope informs causal strength inferences, in: Proceedings of the Annual Meeting of the Cognitive Science Society.

Kalainathan, D., Goudet, O., Dutta, R., 2020. Causal discovery toolbox: Uncovering causal relationships in python. J. Mach. Learn. Res. 21, 1–5.

Karimi, A.H., Schölkopf, B., Valera, I., 2021. Algorithmic recourse: from counterfactual explanations to interventions, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 353–362.

Katz, G.E., Dullnig, D., Davis, G.P., Gentili, R.J., Reggia, J.A., 2017. Autonomous causally-driven explanation of actions, in: 2017 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE. pp. 772–778.

Khosravi, H., Shum, S.B., Chen, G., Conati, C., Tsai, Y.S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., Gašević, D., 2022. Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence 3, 100074.

Kim, M.Y., Atakishiyev, S., Babiker, H.K.B., Farruque, N., Goebel, R., Zaïane, O.R., Motallebi, M.H., Rabelo, J., Syed, T., Yao, H., et al., 2021. A multi-component framework for the analysis and design of explainable artificial intelligence. Machine Learning and Knowledge Extraction 3, 900–921.

Kliangkhlao, M., Limsiroratana, S., Sahoh, B., 2022. The design and development of a causal bayesian networks model for the explanation of agricultural supply chains. IEEE Access 10, 86813–86823.

Kovalerchuk, B., Ahmad, M.A., Teredesai, A., 2021. Survey of explainable machine learning with visual and granular methods beyond quasi-explanations. Interpretable artificial intelligence: A perspective of granular computing , 217–267.

Kumar, V., Boulanger, D., 2020. Explainable automated essay scoring: Deep learning really has pedagogical value, in: Frontiers in education, Frontiers Media SA. p. 572367.

Landgrebe, J., 2022. Certifiable ai. Applied Sciences 12, 1050.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B., 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nature communications 11, 3852.

Lazzari, M., Alvarez, J.M., Ruggieri, S., 2022. Predicting and explaining employee turnover intention. International Journal of Data Science and Analytics , 1–14.

Leventi-Peetz, A.M., Östreich, T., Lennartz, W., Weber, K., 2022. Scope and sense of explainability for ai-systems, in: Proceedings of SAI Intelligent Systems Conference, Springer. pp. 291–308.

Lipton, Z.C., 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 31–57.

Medianovskyi, K., Pietarinen, A.V., 2022. On explainable ai and abductive inference. Philosophies 7, 35.

Meshcheryakov, G., Igolkina, A.A., Samsonova, M.G., 2021. semopy 2: A structural equation modeling package with random effects in python. arXiv preprint arXiv:2106.01140 .

Molnar, C., 2020. Interpretable machine learning. Lulu. com.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 417–431.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2022. General pitfalls of model-agnostic interpretation methods for machine learning models, in: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, Springer. pp. 39–68.

Moscato, V., Picariello, A., Sperlí, G., 2021. A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications 165, 113986.

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 607–617.

Naser, M., 2021. An engineer's guide to explainable artificial intelligence and interpretable machine learning: Navigating causality, forced goodness, and the false perception of inference. Automation in Construction 129, 103821.

Pearl, J., 1985. Bayesian netwcrks: A model cf self-activated memory for evidential reasoning, in: Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA, pp. 15–17.

Pearl, J., 2009. Causality. Cambridge university press.

Pearl, J., 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pp. 3–3.

Pearl, J., 2019. The seven tools of causal inference, with reflections on machine learning. Communications of the ACM 62, 54–60.

Pearl, J., Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic books.

Ramsey, J.D., Zhang, K., Glymour, M., Romero, R.S., Huang, B., Ebert-Uphoff, I., Samarasinghe, S., Barnes, E.A., Glymour, C., 2018. Tetrad—a toolbox for causal discovery, in: 8th International Workshop on Climate Informatics.

Raymond, E., 1999. The cathedral and the bazaar. Knowledge, Technology & Policy 12, 23–49.

Reichenbach, H., 1956. The direction of time. volume 65. Univ of California Press.

Reimers, C., Runge, J., Denzler, J., 2020. Determining the relevance of features for deep neural networks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16, Springer. pp. 330–346.

Reiter, E.B., 2019. Natural language generation challenges for explainable ai, in: 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence.

Robins, J.M., Wasserman, L., 1999. On the impossibility of inferring causation from association without background knowledge. Computation, causation, and discovery 1999, 305–21.

Rosseel, Y., 2012. lavaan: An r package for structural equation modeling. Journal of statistical software 48, 1–36.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence 1, 206–215.

Ryo, M., Angelov, B., Mammola, S., Kass, J.M., Benito, B.M., Hartig, F., 2021. Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. Ecography 44, 199–205.

Sachan, S., Yang, J.B., Xu, D.L., Benavides, D.E., Li, Y., 2020. An explainable ai decision-support-system to automate loan underwriting. Expert Systems with Applications 144, 113100.

Sahoh, B., Choksuriwong, A., 2021. Beyond deep event prediction: Deep event understanding based on explainable artificial intelligence. Interpretable Artificial Intelligence: A Perspective of Granular Computing , 91–117.

Sahoh, B., Choksuriwong, A., 2022. A proof-of-concept and feasibility analysis of using social sensors in the context of causal machine learning-based emergency management. Journal of Ambient Intelligence and Humanized Computing 13, 3747–3763.

Sahoh, B., Kaewrat, C., Yeranee, K., Kittiphattanabawon, N., Kliangkhlao, M., 2022. Causal ai-powered event interpretation: A cause-and-effect discovery for indoor thermal comfort measurements. IEEE Internet of Things Journal 9, 23188–23200.

Salmon, W.C., 1984. Scientific explanation and the causal structure of the world. Princeton University Press.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y., 2021. Toward causal representation learning. Proceedings of the IEEE 109, 612–634.

Scutari, M., 2010. Learning bayesian networks with the bnlearn R package. Journal of Statistical Software 35, 1–22. doi:10.18637/jss.v035.i03.

Shimojo, A., Miwa, K., Terai, H., 2020. How does explanatory virtue determine probability estimation?—empirical discussion on effect of instruction. Frontiers in Psychology 11, 575746.

Sovrano, F., Vitali, F., Palmirani, M., 2019. The difference between explainable and explaining: Requirements and challenges under the gdpr., in: XAILA@ JURIX.

Spirtes, P., Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Social science computer review 9, 62–72.

Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D., 2000. Causation, prediction, and search. MIT press.

Tan, J., Geng, S., Fu, Z., Ge, Y., Xu, S., Li, Y., Zhang, Y., 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning, in: Proceedings of the ACM Web Conference 2022, pp. 1018–1027.

Taschdjian, Z., 2020. Why did the robot cross the road?, in: International Conference on Human-Computer Interaction, Springer. pp. 527–537.

Textor, J., Van der Zander, B., Gilthorpe, M.S., Liśkiewicz, M., Ellison, G.T., 2016. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. International journal of epidemiology 45, 1887–1894.

Van Lent, M., Fisher, W., Mancuso, M., 2004. An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the national conference on artificial intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. pp. 900–907.

Van der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., Viergever, M.A., 2022. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. Medical Image Analysis , 102470.

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. 31, 841.

Wang, P.Y., Galhotra, S., Pradhan, R., Salimi, B., 2021. Demonstration of generating explanations for black-box algorithms using lewis. Proceedings of the VLDB Endowment 14, 2787–2790.

Watson, D., 2022. Rational shapley values, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1083–1094.

Watson, D.S., Gultchin, L., Taly, A., Floridi, L., 2021. Local explanations via necessity and sufficiency: Unifying theory and practice, in: Uncertainty in Artificial Intelligence, PMLR. pp. 1382–1392.

Watson, M., Hasan, B.A.S., Al Moubayed, N., 2022. Agree to disagree: When deep learning models with identical architectures produce distinct explanations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 875–884.

Wu, X., Li, J., Qian, Q., Liu, Y., Guo, Y., 2021. Methods and applications of causal reasoning in medical field, in: 2021 7th International Conference on Big Data and Information Analytics (BigDIA), IEEE. pp. 79–86.

Xu, L., 2018. Machine learning and causal analyses for modeling financial and economic data, in: Applied Informatics, Springer. p. 11.

Yang, W.T., Reis, M.S., Borodin, V., Juge, M., Roussy, A., 2022. An interpretable unsupervised bayesian network model for fault detection and diagnosis. Control Engineering Practice 127, 105304.

Yang, Y., Zhuang, Y., Pan, Y., 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. Frontiers of Information Technology & Electronic Engineering 22, 1551–1558.

Zapaishchykova, A., Dreizin, D., Li, Z., Wu, J.Y., Faghihroohi, S., Unberath, M., 2021. An interpretable approach to automated severity scoring in pelvic trauma, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer. pp. 424–433.

Zednik, C., Boelsen, H., 2022. Scientific exploration and explainable artificial intelligence. Minds and Machines 32, 219–239.

Zimmermann, R.S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T., Brendel, W., 2021. How well do feature visualizations support causal understanding of cnn activations? Advances in Neural Information Processing Systems 34, 11730–11744.