Chapter 1
Causality:
The Next Step in Artificial Intelligence

Luis Cavique
https://orcid.org/0000-0002-5590-1493
Universidade Aberta, Portugal

ABSTRACT

Judea Pearl's ladder of causation framework has dramatically influenced the understanding of causality in computer science. Despite artificial intelligence (AI) advancements, grasping causal relationships remains challenging, emphasizing the causal revolution's significance in improving AI's understanding of cause and effect. The work presents a novel taxonomy of causal inference methods, clarifying diverse approaches for inferring causality from data. It highlights the implications of causality in responsible AI and explainable AI (xAI), addressing bias in AI systems. The chapter points out causality as the next step in AI for creating new questions, developing causal tools, and clarifying opaque models with xAI approaches. The work clarifies causal models' significance and implications in various AI subareas.

1. INTRODUCTION

The causal revolution heralded by Judea Pearl [Pearl 2000], [Pearl, Mackenzie 2018], [Pearl 2019] caught the attention of disciplines such as causal discovery and causal inference. Causal discovery aims to infer a causal structure based on observable data. In other words, given a dataset, find the causal model usually represented by a direct acyclic graph. Causal inference comprises a set of tools that allow data analysts to measure cause-and-effect relationships. In a complex world, causal inference helps establish the causes and effects of the actions studied, for example, the impact of minimum wage increases on employment or the influence of legislation on the number of enrolled students.

In the current era of Big Data, the evaluation of data-driven models favors more explanatory models than just predictive ones. The difference between correlational and causation is at the heart of the controversy over prediction and explanation. These two tasks must be distinguished for Artificial Intelligence (AI), giving rise to new disciplines such as explainable AI (xAI) [Belle, Papantonis 2020].

DOI: 10.4018/978-1-6684-9591-9.ch001

Objectives

This work has two objectives. Firstly, clarify the concept of causality, exploring the cause-and-effect relationship between variables. Then, the research aims to highlight how grasping causality impacts the current advancements in AI.

Contributions

Beyond the contribution of an integrated view of causal models with xAI, two incremental contributions are proposed. The first is related to the new taxonomy of causal inference approaches. The second regards identifying the current and emergent groups of techniques in xAI.

Organization

The structure of the rest of the paper can be summarized as follows. Section 2 covers some background information. Section 3 introduces the subject of causality and relevant definitions. Causal discovery is presented in Section 4. Causal inference is developed in Section 5. In Section 6, responsible AI and explainable AI are presented as a consequence of causality. Finally, in Section 7, we draw some conclusions.

2. BACKGROUND INFORMATION

This work aims to present a comprehensible data science maturity model that includes the well-known business intelligence and analytics areas, the new practices in business experimentation [Thomke 2020], and Pearl's latest ideas on causality [Pearl 2019]. The proposed pipeline can be scratched as BI → BA → BE. The proposed maturity model named IABE is the Intelligence, Analytics, and Business Experimentation acronym [Cavique et al. 2023].

In this section, three crucial components of Data Science have been detailed: Business Intelligence (BI), Business Analytics (BA), and Business Experimentation (BE), which includes causality concepts.

Data Science is the current term for the science that analyzes data, combining statistics with machine learning/data mining and database technologies to respond to Big Data's challenges.

The modern Data Science developed in the 2010s corresponds to the merges of several areas during that time [Davenport 2014]:

- in the 1960s, Machine Learning, ML,
- in the 1970s, Decision Support Systems, DSS,
- in the 1980s, Executive Information Systems, EIS,
- in the 1990s, Online Analytical Processing, OLAP,
- in the 2000s, Business Intelligence and Analytics, BI&A,
- in the 2010s, Big Data and IoT,
- and finally, in the 2020s, the rise of BE and causality.

2.1. Business Intelligence and Analytics

The work of Sharda, Delen, and Turban [2017] follows part of the Gartner Analytic Ascendancy Model (GAAM), dividing Business Analytics into three types: descriptive, predictive, and prescriptive. Unlike GAAM, the diagnostic type is not mentioned in this taxonomy. Moreover, the prescriptive type includes the areas of optimization, simulation, decision modeling, and expert systems. This type aims to find the best possible decisions and answer questions like "what should I do?" or "what is the best option?".

A classic taxonomy in Management Science [Eppen et al. 1998] split the equivalent BA types into two dimensions: reasoning and stochastic views. In reasoning, dimension models can be considered more deductive or inferential. Moreover, the stochastic dimension models are determinist or probabilistic.

Deductive modeling is a symbolic model in which variables, parameters, and algorithmic relationships are assumed from prior knowledge. This approach tends to be top-down and involves few data items. Deductive modeling can also be called decision-making and is equivalent to the prescriptive type of Sharda et al. [2017]. In this work, we use the term decision making instead of prescriptive. On the other hand, inferential modeling is a symbolic model in which variables, parameters, and algorithmic relationships are estimated by data analysis. This approach tends to be bottom-up and involves hundreds of data items, and inferential modeling is strongly related to data analysis. Table 1 shows a table with two entries with four types of models, given the possible pair of dimensions (reasoning, stochastic).

Table 1. Types of analytic models

	Deterministic modeling	**Probabilistic modeling**
Deductive modeling (decision making)	optimization	decision analysis, simulation
Inferential modeling (data analysis)	database query	statistical analysis, forecasting

The data-driven areas of Artificial Intelligence have changed names over time, from machine learning to knowledge discovery, data mining, and machine learning for the second time. The same happens in Operations Research (OR); after using the name Management Science, the Institute for Operations Research and the Management Sciences (INFORMS) promotes the new title of OR and Analytics research.

2.2. Business Experimentation

Fisher [1966] introduced the experimental design, which uses randomized trials with an experiment and a control sample, also known as A/B testing. Since then, this method has become a standard in science.

Although the origins of the earliest experimentalists are associated with health care, the social and economic laboratories became prominent in the 1930s. A relevant development in the structure of social networks came from an experiment by the American psychologist Stanley Milgram [1967]. Milgram's experiment consisted of sending letters from people in Nebraska in the Midwest to be received in Boston, on the East Coast, where people were instructed to pass the letters by hand to someone they knew. An average of six people passed on the letters that reached the destination. Milgram concluded that the

experiment showed that, on average, Americans are no more than six steps away from each other. This experiment led to the six-degree concept of separation and the notion of the small world in the analysis of social networks.

The rise of the internet opened new opportunities for digital companies in the global market. The design of interfaces led to the knowledge of rapid prototyping and a new experimentation culture. B2C companies accustomed to rapid prototyping experiences provide new insights instead of circular conversations based only on opinion [Bland, Osterwalder 2020].

The scientific method has been used in large technological companies such as Google, Facebook, and Netflix in randomized trials with millions of participants. Companies look at experiments to understand better user behavior instead of wasting millions of dollars yearly on advertising campaigns.

In search of excellence in customer experience, developing new products, or trying new models, a new discipline arises – business experimentation [Luca, Bazerman 2020], [Thomke 2020]. The experimentation follows the classic deductive process, which includes three phases: firstly, generate testable hypotheses, then run disciplined experiments, and finally, learn meaningful insights.

Several myths about business experimentation tend to disappear in the most recent approaches. We choose three examples from Thomke [2020]:

i) The need for many transactions to run experiments in brick-and-mortar companies can be simplified using small samples.
ii) The Big Data from GAFAM (Google, Amazon, Facebook, Apple, Microsoft) is unnecessary to find cause and effect relationships, as small samples can infer causality.
iii) Companies that carried out customer experiments without prior consent are currently more restricted following the publication of the General Data Protection Regulation (GDPR). On the other hand, fully informing customers can lead to problems of emotional contagion. It is necessary to find a balance between giving up the innovation provided by experiments and permission to carry out tests.

3. CAUSALITY

In this section, we first introduce the concept of causality notation and the significant contributions made by Judea Pearl [2019]. Subsequently, we explore the dichotomy between causal discovery and causal inference.

3.1. Definitions

Analytical models discover or describe exciting patterns and make predictions based on good fits of historical data. In evaluating data-driven models' recent movements are presented in favor of explanatory models. The difference between correlational analysis and causality is at the heart of the controversy over prediction and explanation. In data science, two tasks must be distinguished: prediction and explanation.

In prediction, two variables are used: the independent variable X and the dependent variable Y. The original data is divided into the training and testing data sets to find the Y=f(X) function, where X is a covariate, and Y is the outcome.

A new variable type should be included: the intervention/treatment T. In this task, outcome Y of treatment T is the subject of the study. For this purpose, test and control datasets are used for treatment accomplished T=1 and not accomplished T=0. In analogy with Y=f(X), the explanatory function uses three variables, Y=f (T, X).

Judea Pearl contributed to a new view of causality from the point of view of computer scientists, summarized by the ladder of causation [Pearl, Mackenzie 2018]. He argues that artificial intelligence still does not master causal-and-effect relationships.

Pearl [2019] proposes three levels of causality: association, intervention, and counterfactual. The association has no causal consequences, contrary to the intervention and counterfactuals. Association corresponds to the predictive approach Y=f(X), usual in Machine Learning. The intervention is exemplified by the A/B testing, where treatment T appears in the equation Y=f(T, X). Finally, counterfactual (or unavailable data) involves imaginary worlds and specific approaches that compare treatments.

Since the counterfactual level includes questions with 'what if?' and 'why?', we add a new rung on the ladder, which makes the reason for the working title "Book of Why" [Pearl, Mackenzie 2018]. The different levels are associated with specific questions, as follows:

- Association: What is the relation between X and Y?
- Intervention: What if they do treatment T?
- Counterfactuals: What if they had acted differently? (T=1 instead of T=0)

Similarly, the work of Hernán et al. [2019] advocates three tasks in data science: description, prediction, and causal inference, which are applied to a training program in data science. The students learn to differentiate the three tasks and then generate and analyze data for each task. The students also learn to ask scientific questions for each task.

The Direct Acyclic Graphs, DAG, are a handy tool in causal representation; they describe the causal assumptions of each study [Pearl, Glymour 2016]. The nodes correspond to the variables (treatment T, covariates X, and outcome Y), and the arrows are the eventual association between the nodes.

Moving up the second rung of Pearl's causality ladder, the intervention corresponds to an experiment trial with randomly chosen test and control groups. Intervention answers questions like 'what is the effect of T on Y?' or 'what if I do treatment T?'.

In the study of causality, the data description and the DAG must be presented before the modeling phase. The assumptions represented in the causal diagrams are drawn before the conclusions [Hernán 2017].

One of the biggest causality problems in observational studies is spurious or confounding relationships. If we associate the relation T→Y, a third variable X, can be a mediator between T and Y (T→X→Y) or a covariate that influences the two variables (T←X→Y), depending on the direction of the relation. The Pearl's back-door path is any path from T to Y that starts with an arrow pointing to T, *i.e.,* X→T→Y. If we lock the back door X→T, the variables T and Y will not be confounded.

In many situations, a linear regression model can be used to estimate the treatment effect. This is often done in the context of an observational study or a randomized controlled trial (RCT). The treatment effect is the outcome difference between the treatment and control groups. In the context of a simple linear regression model, the model might look something like this:

$$Y = a + b.T + c.X + e$$

where *e* represents the error and slope *b* measures the causal effect [Angrist, Pischke 2015]. For the special case of the bivariate regression:

$$Y = a + b.T + e$$

The conditional expectation Y, given the treatment T, takes two values, as follows:

$$E(Y|T=0) = a$$

$$E(Y|T=1) = a + b$$

and then $b = E(Y|T=1) - E(Y|T=0)$ is the difference in expected Y with treatment T.

Using this notation $E(Y|T) = E(Y|T=0) + [E(Y|T=1) - E(Y|T=0)]$. $T = a + b$. T. So, E(Y|T) is a linear function of T, with slope *b* and intercept *a*. The regression slope measures the difference in expected Y with treatment T switched on and off.

The fundamental problem of causal inference states that it is impossible to observe both potential outcomes of an individual treated and not treated [Holland 1986]. Since we only observe the outcome, the non-observable or unrealized outcome is called the counterfactual.

Counterfactual analyses have become popular since the philosophical developments in the 1970s. The best-known counterfactual analysis of causation is the theory of David Lewis [1973]. Counterfactual truths are fictional since they occur in a different world. A counterfactual world is Spatio-temporal disconnected from our world, and there is no interaction between worlds, so we cannot check its existence empirically.

The core idea behind the counterfactual theory of causation is causal dependence. A hypothetical scenario can define causal dependence: given that T and Y are different events, Y causally depends on T if and only if the counterfactual "if T were not to occur, Y would not occur" is a true sentence. The same reasoning is addressed in law, considering that jurists have searched for a direct test of the defendant's guilt called 'but-for causation' or objective cause for centuries. This case is also known as a *sine qua non* or necessary condition.

3.2. Causal Discovery and Causal Inference

As statistical learning and probabilistic reasoning, causal discovery (or causal learning) and causal reasoning (or causal reasoning) are two related concepts in the field of causality. Observations and outcomes support statistics and probabilities. On the other hand, causal discovery and inference include the treatment variable.

Causal discovery refers to identifying causal relationships, or dependencies between variables, from observational data without prior knowledge or assumptions about the underlying causal structure. It aims to uncover the underlying causal connections among variables based on statistical patterns and dependencies observed in the data.

Causal inference, on the other hand, involves using existing knowledge or assumptions about causal relationships to conclude the causal effects of interventions or treatments. It focuses on estimating the causal effect of the treatment variable T on outcome Y by considering the causal structure that is already known or assumed.

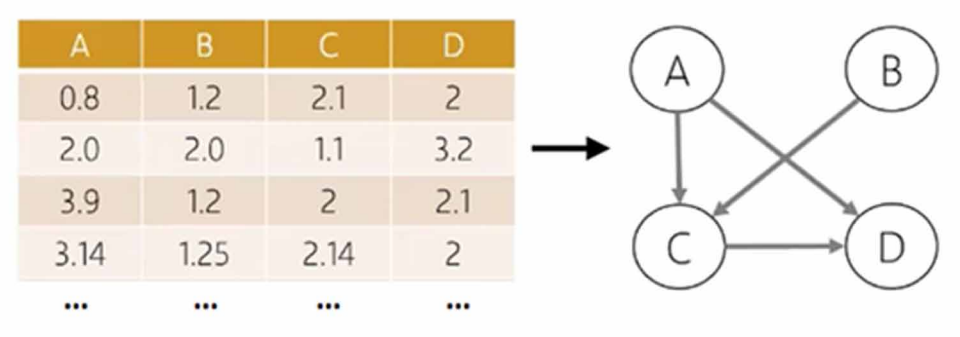
Causal discovery and causal inference work sequentially. Causal discovery techniques can help identify potential causal relationships between variables, which can be used as inputs to causal inference models. Causal inference, on the other hand, often relies on additional assumptions or experimental data to estimate causal effects more accurately. In the following sections, this topic is discussed.

4. CAUSAL DISCOVERY

Despite Pearl's levels of causality [Pearl 2019], there are two fundamental disciplines in the area of causality: causal discovery and causal inference.

In causal discovery (or causal structure learning), given a dataset, the objective is to draw a causal diagram or a DAG (Directed Acyclic Graph), where "X→ Y" can be read as "X causes Y", as shown in Figure 1. The causal inference methods are used to evaluate the causal effects after establishing the relation of the covariates, the interventions, and the outcomes.

Figure 1. Dataset and directed acyclic graph (DAG)



For DAG degeneration, there are three approaches: (i) functional causal models, (ii) score-based methods, and (iii) constraint-based methods [Zanga et al. 2022], [Molak 2023].

(i) Functional Causal Models

Functional Causal Models (FCMs) assume that the relationships between variables can be represented by structural equations involving the direct causes of each variable. LiNGAM (Linear Non-Gaussian Acyclic Models) focuses explicitly on the data's linear relationships between variables and non-Gaussianity. LiNGAM utilizes non-Gaussianity properties, such as the Independent Component Analysis (ICA), to identify the causal ordering of variables. By iteratively estimating the structural coefficients, LiNGAM aims to discover the underlying causal structure of the linear FCM.

(ii) Score-based methods

Score-based methods have also been used to discover causal structures. Score-based methods include the Greedy Equivalence Search (GES), Fast GES, and K2 algorithms. The Greedy Equivalence Search (GES) algorithm starts with an empty graph and iteratively adds and removes directed edges to maximize the improvement of the scoring function. The Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) are examples of scoring functions. Fast GES is an improved and parallelized version of GES. K2 performs a constructive heuristic search for the parents of each node. Max-Min Hill Climbing finds the skeleton of the Bayesian network followed by a heuristic to orient the direction of the edges.

(iii) Constraint-based methods

The Inductive Cause and PC algorithms are among the most common constraint-based algorithms. IC (Inductive Causation) returns the equivalent class of the DAG based on the estimated probability distribution of random variables and an underlying DAG structure. PC (Peter-Clark, named in honor of the authors Peter Spirtes and Clark Glymour) is produced by iteratively checking the conditional independence conditions of adjacent nodes given a partially directed acyclic graph. In causal discovery, the PC algorithm has more bibliographic references.

4.1 PC Algorithm

The PC algorithm [Spirtes et al. 2000] is a well-known algorithm used in causal discovery. It is used for learning the structure of a causal graph from observational data.

The PC algorithm is a constraint-based approach that combines statistical independence tests with graph-based algorithms. The algorithm follows a two-step process:

Step 1: *Determine the Graph Skeleton*

- Given a complete undirected graph, the algorithm identifies the skeleton of the causal graph by testing for statistical independence between variables. The algorithm checks the independence of pairs (X, Y) of variables conditioned by a third variable, Z, $X \perp\!\!\!\perp Y \mid Z$. If variable Z explains the correlation, the relation is called d-separated, and the edge is removed.
- The result of this step is an undirected graph that does not specify the directionality of the causal links.

Step 2: *Determine the Orientation of the Edges*

- The algorithm applies a set of orientation rules based on the concept of "immorality" in the graph. There are different possibilities to draw three variables: the chain $X \rightarrow Y \rightarrow Z$ or $X \leftarrow Y \leftarrow Z$; the fork $X \leftarrow Y \rightarrow Z$, and the collider $X \rightarrow Y \leftarrow Z$.
- Using the conditional independence tests again, if $X \perp\!\!\!\perp Y \mid Z$, we are in a fork chain. Otherwise, if $\sim X \perp\!\!\!\perp Y \mid Z$, it is the case of a collider. The algorithm adjusts the orientations of colliders to satisfy the global Markov property, a fundamental assumption in causal graphical models.

The PC algorithm's first step removes correlated edges given a third variable, and the second step studies what came first: the egg or the chicken. The PC algorithm is widely used in causal discovery due to its efficiency and ability to handle large datasets. It provides a valuable tool for understanding causal relationships and inferring causal structures from observational data.

5. CAUSAL INFERENCE

Causal inference is determining cause-and-effect relationships between variables in a system. It involves understanding how changes in one variable (treatment T) cause changes in another (outcome Y). Researchers often use experimental and statistical techniques to establish causality in causal inference.

The effect of the treatment (T) on the outcome (Y) can be expressed as the difference between the potential outcomes when the treatment is applied (Y(1)) and when it is not applied (Y(0)). The referred difference is known as the Average Treatment Effect (ATE):

$$ATE = E[Y(1) - Y(0)]$$

A widely used technique in experimental studies is randomized controlled trials, RCTs, randomly assigning treatments to study participants (like A/B testing), allowing the causal effect estimation.

A/B testing is the most common intervention, but it is impossible to apply A/B testing consistently. In some areas of social science and healthcare, the A/B test is not considered adequate because it is considered unfeasible or unethical. For instance, when studying the effect of smoking, it is unethical to ask users to consume tobacco to perform a controlled trial. Moreover, historical data does not present test and control samples in many economic activities. Instead of sampling and experimental design, the focus is on quasi-experimental design, mainly observational studies.

Given many causal inference techniques for observational studies, we present a taxonomy in the following subsection.

5.1. Taxonomy of Causal Inference Techniques

Unlike A/B testing, observational studies involve observing subjects in their natural environment without any manipulation or intervention from the researcher. In data science, the chosen technique is also very dependent on the available data. The presented taxonomy considers the accessible data. In Table 2, a synthesis of the causal inference techniques is presented. The following books are good references: Cunningham [2021] and Huntington-Klein[2022].

Table 2. Taxonomy of the causal inference techniques

category	variables	techniques
treatment T and outcome Y	X, T, Y	Propensity Score Matching and Stratification
previous variables with a constant	X, T, Y, cutoff	Regression Discontinuity and Difference-in-differences
previous variables with a third variable	X, T, Y, Z/M	Instrumental Variables and Mediation

Given the variables X (covariates), T (treatment), and Y(outcome), Propensity Score Matching can balance the covariates between treatment and control groups, mimicking a randomized control trial. Stratification, another method based on propensity scores, involves dividing the data into distinct strata based on the propensity scores and estimating the treatment effect within each stratum.

The second group of techniques adds a constant to the previous variables. Regression Discontinuity can be used when the assignment variable has a specific cutoff. Difference-in-differences is used when we have repeated cross-sectional data over time and are interested in the effect of a treatment applied at a specific time. It compares the outcome change over time between a group that received the treatment and a group that did not.

The last group of techniques adds a third variable to variables X, T, and Y. When a third variable, Z, is added. Instrumental Variables can be used when there is concern about unmeasured confounding variables affecting the treatment-outcome relationship. Adding a variable M, mediation analysis can be used to understand the indirect effects of the treatment on the outcome via the mediator.

Propensity Score Matching and Stratification

Given the fundamental problem of causal inference that it is impossible to observe both potential outcomes of an individual treated and not treated and to avoid confounding in the presence of many covariates, matching seems advisable. Propensity Score Matching involves pairing a treated individual with one non-treated individual with similar characteristics, given by the propensity score. Propensity Score Matching may not limit the matching to a 1:1 ratio. The 1:many matchings, where one treated unit is matched with several control units with similar propensity scores.

Stratification (or subclassification) is another method used in causal inference to deal with confounding variables when estimating the effect of a treatment. In propensity score stratification, individuals are divided into mutually exclusive groups (or "strata") based on their propensity scores. For example, one might create ten strata such that all individuals with propensity scores between 0 and 0.1 fall into the first stratum for ten deciles. Stratification is a relatively straightforward method of using propensity scores to estimate treatment effects. Stratification can be implemented with parametric (like Logistic Regression or Linear Discriminant Analysis) and non-parametric models (Machine Learning approaches like Decision Trees).

Regression Discontinuity and Difference-In-Differences

A significant advance in the operability of counterfactual reasoning took place in econometrics [Angrist, Pischke 2015]. Some counterfactual questions can be answered using appropriate statistical techniques, named counterfactual impact evaluation methods. Regression Discontinuity Design and Difference-in-differences allow an intuitive graphical representation [Crato, Paruolo 2019].

Given a set of observations and intervention, the Regression Discontinuity Design measures the impact of the intervention given by the discontinuity between two regressions. Since the observations contributing to identifying the counterfactual effect are mainly those around the intervention, the method may require large sample sizes.

The difference-in-differences method is a statistical technique that measures the impact evaluation of the relative impacts of two sets of observations, mimicking an experimental research design by studying the differential effect of an intervention on a treatment and control group. The difference in the slopes of the linear regressions gives the counterfactual change.

Instrumental Variables and Mediation

Since observational data is not always the outcome of a random trial, the instrumental variables [Angrist, Pischke 2015] approach is the most common technique to estimate causal effects using non-experimental data. Instrumental variables estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment.

The most relevant cause-effect relationship mechanism is known as the 'why?' which can be found in controlled experiments and observational studies.

One of the first controlled experiments was carried out by the physician James Lind in 1747, who administered six different treatments to sailors with scurvy, concluding that citrus fruits cured patients. The correspondent DAG is Citrus → ~Scurvy, that is, citrus are the cause of non-scurvy, or citrus avoid scurvy. However, the reason for the cure is not evident, and the answer would only come in the 1930s with the discovery of vitamin C. On the path (Citrus → Vitamin C → ~ Scurvy), the variable Vitamin C mediates the relation. Therefore, mediation analysis is the tool that allows a better understanding of the association of two variables.

Baron and Kenny [1986] define the principles for detecting measurement between three variables, making it one of the 33 most cited papers. Following the proposal in their seminal work, to test for mediation, one should estimate three regression equations: a) mediator M is explained by the independent variable T; b) dependent variable Y is explained by the independent variable T; c) dependent variable Y is explained by the independent variable T and the mediator M.

After the regressions are calculated, to establish mediation, the following conditions must be fulfilled: a) the independent variable T must affect the mediator M; b) the independent variable T must show to affect the dependent variable Y; c) the mediator M must affect the dependent variable Y.

6. RESPONSIBLE AND EXPLAINABLE AI

Causality and explainable AI are related because they provide insights into complex systems' inner workings. The why questions of Pearl and the right-to-explanation are on the same path.

Causality is concerned with identifying the causal relationships between variables in a system, such as the relationship between a particular input and its effect on an output. In the context of AI, causal inference can help us understand how the treatment feature influences the outcome.

Explainable AI, on the other hand, is focused on developing AI systems that can provide clear and understandable explanations for their outputs or predictions. Causal tools are more and more critical in situations where the decisions made by AI systems have significant consequences, such as in healthcare, finance, or legal contexts.

Causal inference is an essential tool of responsible AI since we can better understand how AI systems make their predictions and provide more trustworthy and transparent explanations. Causality helps build trust in AI systems, improve performance, and ensure they are used ethically and responsibly.

6.1 Responsible AI

As AI becomes more prevalent in various domains, upholding fairness, transparency, explanation, and responsibility becomes increasingly vital to ensuring that AI benefits all and does not lead to unintended negative consequences. Responsible AI refers to the ethical and accountable development, deployment, and use of artificial intelligence technologies.

Moniz Pereira argues the need to design machine ethics in the theoretical field, paving the way for a new scientific area beyond computational ethics [Pereira, Lopes 2020]. Machine ethics involves creating AI systems that can recognize morally relevant situations, understand the consequences of their actions, and make choices consistent with ethical principles. The goal is to build AI to navigate complex moral dilemmas and contribute positively to society.

Dignum [2019], during her talk, generalized with humor how AI is used in different countries like the United States of America, China, and European countries. In the United States, AI is widely used for commercial purposes and generating revenue. The Chinese government invested heavily in AI for social control and surveillance. European Union (EU) has taken a more cautious approach to AI, focusing on regulatory frameworks to address ethical, legal, and privacy concerns.

In Europe, AI is employed in various sectors, much like in the USA and China, but with a greater emphasis on adhering to regulatory guidelines. The aim is to balance encouraging innovation and transparency in AI. From 2018 onwards, the EU extended this requirement by imposing the so-called right-to-explanation in algorithm decision-making [European Commission 2020].

As a result of the EU's regulatory effort, the EU is further ahead than most countries. Its advances are reflected in the so-called Brussels Effect [Bradford 2012]. The Brussels Effect occurs when other countries replicate the regulatory decisions of the EU. Examples of this are the global application of the General Data Protection Regulation (GDPR) by Facebook or the impact of the European Union's Emissions Trading System on aviation services and industries.

Currently, AI is facing significant controversy over whether to prioritize regulation or certification as a means of governing AI technologies. For example, the Responsible Artificial Intelligence Institute (RAI Institute) provides the first independent, accredited certification program for responsible AI systems. Both regulation and certification have their advocates and detractors, each offering different perspectives on addressing the challenges posed by AI.

Those in favor of AI regulation argue that the rapid advancement of AI technologies requires clear and enforceable rules to mitigate potential risks. They believe comprehensive regulations can set ethical standards, ensure transparency, and protect against bias and discrimination. Proponents of regulation stress the need for government intervention to prevent AI misuse and promote Responsible AI development.

On the other hand, advocates of AI certification argue that a flexible, industry-led approach is better suited to promoting innovation and adaptation. They believe heavy regulation could stifle AI progress and hurt competition. Instead, they propose voluntary certification programs that establish best practices and guidelines, allowing companies to demonstrate their adherence to AI ethical principles without imposing strict legal requirements.

Explainable AI can be seen as a subset of Responsible AI. In the framework of Responsible AI, explainability is a critical factor in ensuring transparency and accountability. Without explainability, holding AI systems accountable for their decisions or ensuring their fairness would be challenging, as we

would not understand the reasoning behind their actions, decisions, or results. Similarly, explainability is critical for users and stakeholders to trust the system, providing them with the necessary insight into its decisions.

6.2. Explainable AI (xAI)

The interest in the interpretability of machine learning algorithms increased significantly after the publication of the 'Book of Why' [Pearl, Mackenzie 2018]. Concepts of interpretability are more oriented to white-box models such as decision trees, decision rules, and linear regression. On the other hand, explainable approaches focus on black-box methods, like neural networks, clarifying individual predictions with SHAP and LIME procedures [Molnar 2020].

A clear xAI taxonomy is presented by Belle and Papantonis [2020], differentiating algorithms (transparent and opaque), models (agnostic and specific), and a group of techniques: feature relevance explanation, local explanations, explanation by model simplification, and visual explanations.

Model-agnostic techniques are general approaches that can be applied to any machine learning model, regardless of its architecture. These methods provide insights into a model's functions without requiring specific knowledge of its internal workings. On the other hand, model-specific techniques are designed to explain the decisions of particular types of models. They exploit specific characteristics of the model to provide explanations.

The approaches to explain opaque algorithms use the reduction of the input data (feature or instances) or the simplifications of the model. Some of the most well-known groups of techniques for explanation include:

- Feature relevance: This technique involves analyzing the contribution of each input feature to the model's output. It can help identify the most critical features the model relies on to make predictions. Example: SHAP (SHapley Additive exPlanations).
- Local or counterfactual explanations: This technique involves generating alternative input scenarios that would result in a different output from the model. This approach is also known as sensitivity analysis or what-if analysis. It can help identify the factors most influential in the model's decision-making process and can also be used to test the model's robustness.
- Model simplification: This technique involves training a simpler, more interpretable model on a subset of the training data, aiming to approximate the original model's behavior in a specific region of the input space. It can help to provide insight into the decision-making process of the original model for a particular instance or set of instances. Example: LIME (Local Interpretable Model-agnostic Explanations).

Causality is not in the listed techniques since it is an exploring approach in xAI. Causality is already mentioned in recent works like Belle and Papantonis [2020] and Masís [2021].

Causality can be crucial in mitigating bias in xAI systems. The xAI goal is to provide explanations for model predictions and ensure that those explanations are fair, transparent, and unbiased. By incorporating causality into the xAI process, we can address potential sources of bias and achieve more reliable and interpretable results as follows:

- Identifying confounding variables: Causality allows us to distinguish between correlation and causation. Often, AI model biases result from correlations that might not be directly causal. We can use causal models to identify and control confounding variables, which are the hidden factors influencing the input features and the model's output.
- Counterfactual observations: Causality enables one to deal with counterfactual observations. By exploring counterfactuals, we can better understand the effect of a particular treatment variable. Counterfactual analysis can help identify and mitigate bias by understanding the causal relationships between features and outcomes and assessing fairness through counterfactual scenarios.

From the business point of view, in July 2023, billionaire Elon Musk announced the debut of a new AI company, XAI, intending to 'understand the true nature of the universe'. XAI stands for Explainable AI and sets itself apart from models like ChatGPT through several vital factors. Its primary focus is on explainability, ensuring it can provide clear justifications for its decisions and responses. Context awareness is another crucial aspect, allowing XAI to accurately understand and respond to nuanced conversations. Moreover, XAI emphasizes ethical considerations, aiming to align with ethical standards, address privacy and bias issues, and promote Responsible AI deployment.

Still in the business area, Gartner, a leading technology analysis company, included causal AI alongside the popular generative AI in its Hype Cycle for New Technologies 2023.

7. CONCLUSION

Judea Pearl has significantly impacted the understanding of causality in computer science through his ladder of causation framework [Pearl, Mackenzie 2018], Pearl [2019]. He emphasizes that despite Artificial Intelligence (AI) advancements, the field still struggles to understand causal relationships and their implications. Pearl's work highlights the significance of the causal revolution, emphasizing the need to improve AI's capacity for understanding cause-and-effect relationships. By doing so, we can unlock the potential for developing more robust and reliable AI systems that have profound and meaningful applications in the real world.

Business Experimentation (BE), as proposed by Thomke [2020], involves vital players GAFAM (Google, Amazon, Facebook, Apple, Microsoft) who play a significant role in this domain. The theoretical foundation of BE is intricately linked to the concept of intervention, as introduced by Pearl [2019]. The insights drawn from BE and its ties to Pearl's intervention concept contribute to the advancement of data-driven strategies and improvements in the overall business landscape.

This work describes the complementarity between causal discovery and causal inference. Moreover, introduces a novel taxonomy of causal inference methods based on the available data in the dataset. The taxonomy reveals three distinct groups: i) involving treatment T and outcome Y, ii) incorporating variables T and Y along with a time variable, and iii) including variables T and Y with an additional third variable, such as an instrumental variable or a mediator. This taxonomy clarifies the diverse approaches to infer causality from data and offers valuable insights for advancing causal analysis methodologies.

AI has rapidly advanced in recent years, transforming various industries and aspects of daily life. In Europe, it has developed several regulatory guidelines to balance innovation and transparency in AI, placing it ahead of many countries, reflected in the Brussels effect. In this work, after detailing the concept of causality, we described its implications in AI, namely in Responsible AI and explainable AI (xAI).

We identified the current groups of techniques in xAI and pointed out that the emergence of causality is crucial to mitigate bias in AI systems.

In summary, we believe that causality is the next step in AI in at least three ways:

- by creating new questions (and finding answers) that go beyond the classic machine learning prediction but measure the effect of the treatment variables in BE environments;
- by creating causal tools to respond to the right-to-explanation and the EU regulation;
- and by clarifying the opaque models with xAI approaches.

Finally, this paper clarifies causal models (discovery and inference) and shows the future implications of causality in different AI subareas.

REFERENCES

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering Metrics: The Path from Cause to Effect*. Princeton University Press.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. doi:10.1037/0022-3514.51.6.1173 PMID:3806354

Belle, V., & Papantonis, I. (2020). *Principles and practice of explainable machine learning*. arXiv:2009.11698.

Bland, D. J., & Osterwalder, A. (2020). *Testing Business Ideas: A Field Guide for Rapid Experimentation*. John Wiley and Sons.

Bradford, A. (2012). The Brussels Effect. *Northwestern University Law Review, 107*. https://ssrn.com/abstract=2770634

Cavique, L., Pinheiro, P., & Mendes, A. B. (2023). (accepted)). Data science maturity model: From raw data to Pearl's causality hierarchy. *WorldCIST*, 2023.

Crato, N., & Paruolo, P. (2019), Data-Driven Policy Impact Evaluation: How Access to Microdata is Transforming Policy Design. Springer. doi:10.1007/978-3-319-78461-8

Cunningham, S. (2021). *Causal inference: the mixtape*. Yale University Press.

Davenport, T. H. (2014). *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business School Publishing Corporation. doi:10.15358/9783800648153

Dignum, V. (2019). The responsibility is ours. In *Artificial intelligence: applications, implications and speculations*. Fidelidade-Culturgest Conferences and Debates.

Eppen, G. D., Gould, F. J., Schmidt, C. P., Moore, J. H., & Weatherford, L. R. (1998), Introductory Management Science: decision modeling with spreadsheets. Prentice-Hall International.

European Commission. (2020), *White Paper on Artificial Intelligence: a European approach to excellence and trust* (White Paper No. COM(2020) 65 final), European Commission. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Fisher, R. A. (1966). *The design of experiments* (8th ed.). Hafner Publishing Company.

Hernán, M. A. (2017), Causal Diagrams: Draw Your Assumptions Before Your Conclusions. EDX. https://www.edx.org/course/causal-diagrams-draw-your-assumptions-before-your.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.1080/01621459.1986.10478354

Huntington-Klein N. (2022). *The effect: an introduction to research design and causality.* Chapman and Hall/CRC.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567. doi:10.2307/2025310

Luca M., & Bazerman, M. H. (2020). *The Power of Experiments: Decision Making in a Data-Driven World.* MIT Press.

Masís S. (2021). *Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples*. Packt Publishing.

Milgram, S. (1967). The Small World Problem. *Psychology Today*, *1*(1), 60–67.

Molnar C. (2020). *Interpretable Machine Learning: a guide for making black box interpretable*. lulu.com.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Pearl J. (2019), The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*. ACM.

Pearl J., & Glymour, M. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pereira, L. M., & Lopes, A. (2020). Machine Ethics: From Machine Morals to the Machinery of Morality, book series Studies in Applied Philosophy, Epistemology and Rational Ethics, SAPERE. Springer.

Sharda, R., Delen, D., & Turban, E. (2017). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective* (4th ed.). Pearson.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT Press.

Thomke S.H. (2020). Experimentation Works: The Surprising Power of Business Experiments. *Harvard Business Review Press*.

Zanga, A., Ozkirimli, E., & Stella, F. (2022). A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning, 151.* Doi:10.1016/j.ijar.2022.09.004

KEY TERMS AND DEFINITIONS

Artificial Intelligence (AI): A branch of computer systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and decision-making.

Business Experimentation (BE): A strategic approach that systematically tests and analyzes different business strategies, processes, or innovations.

Causal Discovery: Aims to identify cause-and-effect relationships between variables in datasets, helping to uncover underlying patterns and dependencies.

Causal Inference: The process of determining cause-and-effect relationships between variables by analyzing data and employing statistical methods to understand how changes in one variable influence another.

Causality: Refers to the relationship between cause and effect, where one event (the cause) leads to another event (the effect).

Explainable AI (xAI): An approach in artificial intelligence that focuses on developing machine learning models and systems that provide transparent and interpretable explanations for their decisions and predictions, enhancing trust and understanding of AI systems by humans.

GAFAM: An acronym representing five of the largest and most influential technology companies in the world: Google, Amazon, Facebook, Apple, and Microsoft.

IABE: An acronym representing the Intelligence, Analytics, and Business Experimentation maturity model.

Responsible AI: This refers to the ethical and accountable development and deployment of artificial intelligence systems, considering factors like fairness, transparency, privacy, and social impact.