Confirmation of Candidature

# Causability in Predictive Black Boxes: Theory, Algorithms and Applications

**Student:** Yu-Liang (Leon) Chou
**Student ID:** n9799362

Submitted in partial fulfillment of the requirement for the degree of
IF49 Doctor of Philosophy

School of Information Systems
Faculty of Science
Queensland University of Technology

August 11, 2022

# Contents

## Proposed Thesis Title

Causability in Predictive Black Boxes: Theory, Algorithms and Applications

## Proposed Supervisory Team

**Principal Supervisor:** Dr Catarina Moreira

**Associate Supervisor:** Prof. Peter Bruza

## Thesis type

Traditional Thesis by Monograph

# 1 Introduction

Artificial intelligence facilitates an individual's life rapidly, it exceeds human performance in many tasks includes categorization, recommendation and game playing. Although this success, the internal mechanisms of these technologies are an enigma because humans cannot scrutinize how these intelligent systems do what they do. This is known as the *black-box* Lipton (2018). Consequently, humans are reliant to blindly accept the answers produced by machine intelligence without understanding how that outcome came to be. There is growing uneasiness about this state of affairs as intelligent technologies increasingly support human decision-makers in high-stakes contexts such as the battlefield, law courts, operating theatres, etc.

## 1.1 The need for Explainability

Several factors motivated the raise of approaches that attempt to turn predictive black-boxes transparent to the decision-maker Doran et al. (2017). One of these factors is the recent European General Data Protection Regulation (GDPR) Goodman and Flaxman (2017), increasing the demand for the ability to question and understand Machine Learning (ML) systems. These regulations have a direct impact on worldwide businesses, because GDPR applies to not only to data being used by European organisations, but also to European data being used by other organisations. Another concerned is discrimination (such as gender, and racial bias) (O'Neil, 2017). Studies suggest that predictive algorithms widely used in healthcare, for instance, exhibited racial biases that prevented minority societal groups from receiving extra care (Obermeyer et al., 2019) or display cognitive biases associated with medical decisions (Lau and Coiera, 2007; Saposnik et al., 2016). In medical Xray images, it was found that deep learning models have learned to detect a mental token that technicians use to visualise the X-ray images, making this feature impacting the predictions of the algorithm (Zech et al., 2016). Other studies revealed gender and racial biases in automated facial analysis algorithms made available by commercial companies (Buolamwini and Gebru, 2018), gender biases in textual predictive models (Bolukbasi et al., 2016; Garg et al., 2018; Caliskan et al., 2017) or even more discriminatory topics such as facial features according to sexual orientation (Kosinski and Wang, 2018).

The black-box problem and the need for interpretability motivated an extensive novel body of literature in machine learning that is focused on developing new algorithms and approaches that not only can interpret the complex internal mechanisms of machine learning predictions, but also explain and provide understanding to the decision-maker of the *why* these predictions (Lakkaraju et al., 2019; Guidotti et al., 2018).

The overarching goal of XAI is to generate human-understandable explanations of the *why* and the *how* of specific predictions from machine learning or deep learning (DL) systems. Páez (2019) extends this goal by adding that explainable algorithms should offer a pragmatic and naturalistic account of understanding in AI predictive models and that explanatory strategies should offer well-defined goals when providing explanations to its stakeholders. For a model to be interpretable, it must suggest explanations that make sense to the decision-maker and ensure that those explanations accurately represent the true reasons for the model's decisions Serrano and Smith (2019).

## 1.2 From Explainability to the need of Causability

Causation is a ubiquitous notion in Humans' conception of their environment (Pearl, 1988). Humans are extremely good at constructing mental decision models from very few data samples because people excel at generalising data and tend to think in cause/effect manners (Byrne,

2019). There has been a growing emphasis that AI systems should be able to build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems Lake et al. (2017). For decision-support systems, whether in finance, law or even warfare, understanding the causality of learned representations is a crucial missing link Pearl (2009); Gershman et al. (2015); Peters et al. (2017). But when considering machines, how can we make computer-generated explanations causally understandable by humans? This notion was recently put forward by Holzinger et al. (2019) in a term coined *causability*.

The ability to find causal relationships between the features and the predictions in observational data is a very challenging problem and constitutes a fundamental step towards explaining predictions (Longo et al., 2020). In this paper, we argue that the causal approaches should be emphasized in XAI to promote a higher degree of interpretability to its users and avoid biases and discrimination in predictive black-boxes (Kilbertus et al., 2017).

### 1.3 Concepts and Definitions in Explainable AI

In this section, we propose the major concepts that are used throughout the work that correspond to the entire thesis:

- **Interpretability** is defined as the extraction of relevant sub-symbolic information from a machine-learning model concerning relationships either contained in data or learned by the model (Murdoch et al., 2019).
- **Explainability**, on the other hand, refers to the ability to translate this sub-symbolic information in a comprehensible manner through human-understandable language expressions (Holzinger et al., 2019).
- **Black box predictor:** It is a machine learning opaque model, whose internals are either unknown to the observer or they are known but uninterpretable by humans.
- **Causal inference:** Refers to the process where causes are inferred from data (Pearl, 2019).
- **Causality:** Is the efficacy by which one process or state (a cause), contributes to the production of another process or state, (an effect), where the cause is partly responsible for the effect, and the effect is partly dependent on the cause (Pearl, 2019).

## 2 Literature Review

Currently, multiple approaches have been presented in the literature to solve the issue of interpretability in machine learning. Generally, this problem can be classified into two major models: Interpretable models that inherently transparent and model agnostic also referred to as post-hoc models (which target extracting explanations out of opaque models). From our systematic literature review, these approaches can be categorized within the taxonomy presented in Figure 1.

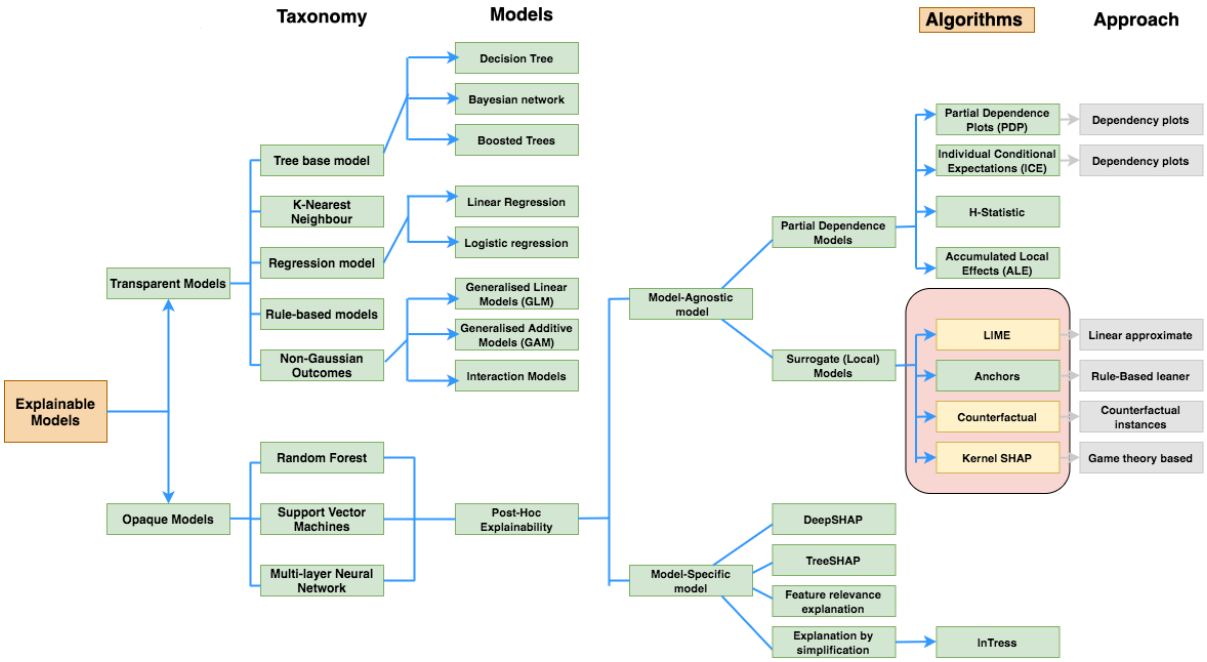


Figure 1: Taxonomy of explainable artificial intelligence

Interpretable models are by design already interpretable, providing the decision-maker with a transparent white box approach for prediction (Molnar, 2020). Decision trees, logistic regression, and linear regression are commonly used interpretable models. These models have been used to explain predictions of specific prediction problems (Siering et al., 2018). Model-agnostic approaches, on the other hand, refer to the derivation of explanations from a black-box predictor by extracting information about the underlying mechanisms of the system (Kim et al., 2020).

Model-agnostic models (post-hoc) are divided into two major approaches: Partial dependency plots and surrogate models. The partial dependency plots can only provide pairwise interpretability by computing the marginal effect that one or two features have on the prediction. Surrogate models, on the other hand, consist on training a new local model that approximates the predictions of a black-box. Moreover, the model-agnostic model have the flexibility of being applied to any predictive model as compared to model-specific model. The two most widely cited post-hoc models in the literature include LIME (Ribeiro et al., 2016) and Kernel SHAP (Lundberg and Lee, 2017). Counterfactuals are also model-agnostic post-hoc approaches, which will be detailed in Section 5.4 as the phase one deliverable.

## 2.1 LIME - Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) explains the predictions of any classifier by approximating it with a locally faithful interpretable model. Hence, LIME generates local interpretations by perturbing a sample around the input vector within a neighborhood of a local decision boundary (Ribeiro et al., 2016; Radwa Elshawi and Sakr, 2019). Each feature is associated with a weight that is computed using a similarity function that measures the distances between the original instance prediction and the predictions of the sampled points in the local decision boundary. An interpretable model, such as linear regression or a decision tree, is able to learn the local importance of each feature. This translates into a mathematical optimization problem expressed as

$$explanation(x) = argmin_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{1}$$

where $\mathcal{L}$ is the loss function which measures the similarity of the explainable model in the boundary of a perturbed data point $z$, $g(z)$, to the original black-box prediction, $f(z)$:

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \left(f(z) - g(z)\right)^2. \tag{2}$$

In Equations 1 and 2, $x$ is the instance to be explained and $f$ corresponds to the original predictive black-box model (such as a neural network). $G$ is a set of interpretable models, where $g$ is an instance of that model (for instance, linear regression or a decision tree). The proximity measure $\pi_x$ defines how large the neighborhood around instance x is that we consider for the explanation. Finally, $\Omega(g)$ corresponds to the model complexity, that is, the number of features to be taken into account for the explanation (controlled by the user) (Molnar, 2020).

## 2.2 Approaches Based on LIME

LIME has been extensively applied in the literature. For instance, Stiffler et al. (2018) used LIME to generate salience maps of a certain region showing which parts of the image affect how the black-box model reaches a classification for a given test image (Zeiler and Fergus, 2014; Lapuschkin et al., 2019). Tan et al. (2019) applied LIME to demonstrate the presence of three sources of uncertainty: randomness in the sampling procedure, variation with sampling proximity, and variation in the explained model across different data points.

In terms of image data, the explanations are produced by creating a set of perturbed instances by dividing the input image into interpretable components (contiguous superpixels) and runs each perturbed instance via the model to get a probability Badhrinarayan et al. (2020). After that, a simple linear model learns on this data set, which is locally weighted. At the end of the process, LIME presents the superpixels with the highest positive weights as an explanation. Preece (2018) proposed a CNN-based classifier, a LIME-based saliency map generator, and an R-CNN-based object detector to enable rapid prototyping and experimentation to integrate multiple classifications for interpretation.

Other researchers propose extensions to LIME. Turner (2016) derived a scoring system for searching the best explanation based on formal requirements using Monte Carlo algorithms. They considered that the explanations are simple logical statements, such as decision rules. (Osbert et al., 2017) utilized a surrogate model to extract a decision tree that represents the model behavior. (Thiagarajan et al., 2019) proposed an approach for building TreeView visualizations using a surrogate model. LIME has also been used to investigate the quality of predictive systems in predictive process analytics (Sindhgatta et al., 2020a). In Sindhgatta et al. (2020b) the authors found that predictive process mining models suffered from different biases,

including data leakage, and revealed that LIME could be used as a tool to debug black box models.

Lastly, a rule-based approach extension for LIME is Anchor (Ribeiro et al., 2018). Anchor attempts to address some of the limitations by maximizing the likelihood of how a certain feature might contribute to a prediction. Anchor introduces IF-THEN rules as explanations as well as the notion of coverage, which allows the decision-maker to understand the boundaries in which the generated explanations are valid.

## 2.3 SHAP - SHapley Additive exPlanations

The SHAP (SHapley Additive exPlanations) is an explanation method that uses Shapley values Shapley (1952) from coalitional game theory to fairly distribute the gain among players, where contributions of players are unequal (Lundberg and Lee, 2017). Shapley values are a concept in economics and game theory and consist of a method to fairly distribute the payout of a game among a set of players. One can map these game-theoretic concepts directly to an XAI approach: a game is the prediction task for a single instance; the players are the feature values of the instance that collaborate to receive the gain. This gain consists of the difference between the Shapley value of the prediction and the average of the Shapley values of the predictions among the feature values of the instance to be explained Strumbelj and Kononenko (2013).

In SHAP, an explanation model, $g(z')$ is given by a linear combination of Shapley values $\phi_j$ of a feature $j$ with a coalitional vector, $z'_j$, of maximum size $M$,

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j. \tag{3}$$

Strumbelj and Kononenko (2013) claim that in a coalition game, it is usually assumed that $n$ players form a grand coalition that has a certain value. Given that we know how much each smaller (subset) coalition would have been worth, the goal is to distribute the value of the grand coalition among players fairly (that is, each player should receive a fair share, taking into account all sub-coalitions). Lundberg and Lee (2017) on the other hand, present an explanation using SHAP values and the differences between them to estimate the gains of each feature.

To fairly distribute the payoff amongst players in a collaborative game, SHAP makes use of four fairness properties: (1) Additivity, which states that amounts must sum up to the final game result, (2) Symmetry, which states that if one player contributes more to the game, (s)he cannot get less reward, (3) Efficiency, which states that the prediction must be fairly attributed to the feature values, and (4) Dummy, which says that a feature that does not contribute to the outcome should have a Shapley value of zero.

## 2.4 Approaches Based on SHAP

SHAP has been applied in a wide range of experiments throughout the literature. del Pozo et al. (2011) and Michalak et al. (2013) showed the capabilities of SHAP for identifying important members of large social networks. Livshits et al. (2019) used Shapley values to measure how much a database tuple contributes to an answer to a query. In Parsa et al. (2020), the authors emphasize that SHAP could generate insightful meaning in interpreting prediction results. They evaluated the evaluating the global importance of the impacts of features on the outcome of the model and also extracted complex and non-linear joint impact of local features. In Peer et al. (2004), the authors use Shapley values to assign importance to protein interactions in large, complex biological interaction networks. More recently, Keane and Smyth (2020) used Shapley values to measure causal effects in neurophysical models. Hilde et al. (2019) explored

the impacts of decision-making in fraud cases. They aimed to increase the trust of domain experts in AI models analyzing transaction frauds in banking. In their research, they offer a case-based SHAP explanation based on neighborhood that enables experts to visualize similar instances to an observation for which a fraud alert was issued. JP et al. (2019) used SHAP in a case of insurance claim by exploring why the particular warranty claims are marked as anomalies by the predictive machine learning model.

## 2.5 Counterfactuals as Means to Achieve Causability

Causality is a fundamental concept to gain *intellectual understanding* of the universe and its contents, it is concerned with establishing cause-effect relationships (Hoque and Mueller, 2021). Causal concepts are central to our practical deliberations, health diagnosis, etc (Holzinger et al., 2019; Ramaravind K. Mothilal, 2020). Even when one attempts to explain certain phenomena, the explanation produced must acknowledge, to a certain degree, the causes of the effects being explained (Halpern and Pearl, 2005). However, the nature and definition of causality is a topic that has promoted a lot of disagreement throughout the centuries in Philosophical literature. While Bertrand Russel was known for being the most famous denier of causality, arguing that it constituted an incoherent topic (Psillos, 2002), it was mainly with the philosopher and empiricist David Hume that the concept of causation started to be formally analyzed 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 $E$ (Psillos, 2002). Hume was also one of the first philosophers to identify causation through the notion of counterfactual: a cause to be an object followed by another in which if the first object (the cause) had not occurred, then the second (the effect) would never exist (Hume, 1739). This concept started to gain more importance in the literature with the works of Lewis (1973a,b, 1986).

Counterfactuals are then defined as a conditional assertion whose antecedent is false and whose consequent describes how the world would have been if the antecedent had occurred (a *what-if* question). In the field of XAI, counterfactuals provide interpretations as a means to point out which changes would be necessary to accomplish the desired goal (prediction), rather than supporting the understanding of why the current situation had a certain predictive outcome (Wachter et al., 2018). While most XAI approaches tend to focus on answering *why* a certain outcome was predicted by a black-box, counterfactuals attempt to answer this question in another way by helping the user understand *what features does the user need to change to achieve a certain outcome* (Poyiadzi et al., 2020). For instance, in a scenario where a machine learning algorithm assesses whether a person should be granted a loan or not, a counterfactual explanation of *why* a person did not have a loan granted could be in a form of a scenario *if your income was greater than* $\$15,000$ *you would be granted a loan* (Ramaravind K. Mothilal, 2020).

# 3 Research Problem

Current XAI models (LIME, Anchor, Kernel SHAP) attempt to decipher a black-box that is already trained and only computes correlations between individual features to approximate the predictions of the black-box, instead of reflecting the true underlying mechanisms of the black box (as pointed out by Rudin (2019)). Therefore, we argue that the inability to disentangle *correlation from causation* can deliver sub-optimal or even erroneous explanations to decision-makers (Richens et al., 2020). Causal approaches should be emphasized in XAI to promote a higher degree of interpretability to its users and avoid biases and discrimination in predictive black-boxes (Kilbertus et al., 2017).

Moreover, the need for interpretability for black box motivated an extensive novel body of literature in machine learning which focused on developing new algorithms and approaches that not only can interpret the complex internal mechanisms of machine learning predictions, but also explain and provide understanding to the decision-maker of the *why* these predictions (Lakkaraju et al., 2019; Guidotti et al., 2018). In this sense, interpretability and explainability have become the main driving pillars of explainable AI (XAI) (Gilpin et al., 2018).

The overarching goal of XAI is to generate human-understandable explanations of the *why* and the *how* of specific predictions from machine learning or deep learning (DL) systems. Páez (2019) extends this goal by adding that explainable algorithms should offer a pragmatic and naturalistic account of understanding in AI predictive models and that explanatory strategies should offer well-defined goals when providing explanations to its stakeholders.

This phenomenon leads to the following research question that we aim to address.

## 3.1 Research Question

- **Research Question 1:** *How to induce a causal abstract model (CAM) from a predictive black box that promotes causability?*

  To provide the user with causal understandings of why certain features contributed to a certain prediction, a theoretical causal abstract model should be defined. This model should be able to compute cause and effect relationships between the permuted features based on a strong mathematical theory and based on probabilistic graphical models. Therefore, we hypothesise that the CAM should provide better insights and consequently better decisions due to the graphical probabilistic nature of the model and the incorporation of causability. A human expert can then offer more precise and reliable explanations for the predicted result.

- **Research Question 2:** *What metrics and/or criteria would be suitable to access the quality of explanations generated from the proposed CAM in the context of a prediction?*

  How to establish metrics that can assess how one explanation can be understood to be better than the other is one of the critical open research questions. By proposing novel and standard metrics for explainability, one can promote the reliability of interpretable models, leading to a higher trust of human decision-makers. It is necessary to develop new standards and norms in terms of evaluation metrics for XAI which can assess both the interpretable model and the explainable model.

- **Research Question 3:** *What is the effectiveness of the proposed causal abstract model when applied to specific predictive domains such as medical decision-making?*

  Determined the effectiveness of the proposed causal abstract model in a certain domain is a crucial aspect after defining the theory and the evaluation metrics. For instance, medical

decision-making is an important domain that highly benefits from explainable models. Holzinger et al. (2019) theorized what could be an interpretable model to interpret medical images of human pathologies and the urgent need for such systems. Finance is another field that should require interpretable models for customer credit evaluation.

# 4 Program of Research

## 4.1 Objectives

The main focus of this research project rests in the development of theoretical models and associated algorithms for interpretability and explanability grounded on the notion of causability, rather than in the development of human interactive explainable interfaces. However, a simple explainability dashboard will also be developed as part of this project, to visualize the output of the causal explanations of the proposed model. This research work will have three major components: the theoretical contribution of a causal abstract model for interpretability; the algorithmic contribution that promotes causal understandable explaplanations (causability) ; and, the application contribution, which will consist in the utilisation of the proposed model and algorithms in a specific domain with standardised evaluation metrics.

To generate an interpretable model with a reliable notion of causability four phases should be covered:

- **Aim 1 (Theory):** Create a new theoretical model (defined as causal abstract model) Which represents causal relationships between features for predictions.
- **Aim 2 (Algorithms):** Enhance the causal abstract model with causability, to furnish an explainability for the user
- **Aim 3 (Evaluation):** Evaluate the quality of causability base explanations in a variety of empirical settings
- **Aim 4 (Application):** Develop an open-source framework, which will operationalise the proposed approaches to yield explainable predictions and render the results to users in an understandable, visual, and interactive manner

## 4.2 Research Methodology

Design science research methodology (DSRM) is ideal to be utilized to achieve the project aims, this methodology focuses on the development and performance of artefacts with the explicit intention of developing functional performance and offers specific guidelines for evaluating and iterate the project. Based on the seven DSRM guidelines presented by **?**, the project can be delivered in an iterative and efficiency way. As the guideline requires, an artefact must be evaluated to ensure its utility for a specific problem. Besides, to form a novel research contribution, the artefact must be created to address a problem that has not yet been solved or offer a more effective solution. Additionally, both the construction and evaluation of the artefact must be done strictly, and the outcome of research supposed to be presented effectively both to technology-oriented and management-oriented audiences.

Figure 2 illustrates the combination of the project aims with DSRM. In the left section, a theoretical methodology and a casual abstract model base algorithm will first be created. In the next steps, a quantitative methodology (metrics) will be adopted to evaluate an artifact (theory & algorithm) developed from the prior step. The quantitative methods will also involve the empirical evaluation of the theory, and these methods are established in the field of Machine Learning ethics performance assessment. While the theory and algorithms present well after evaluation, an application is expected to be generated and applied to the expert and user at the end of the research.

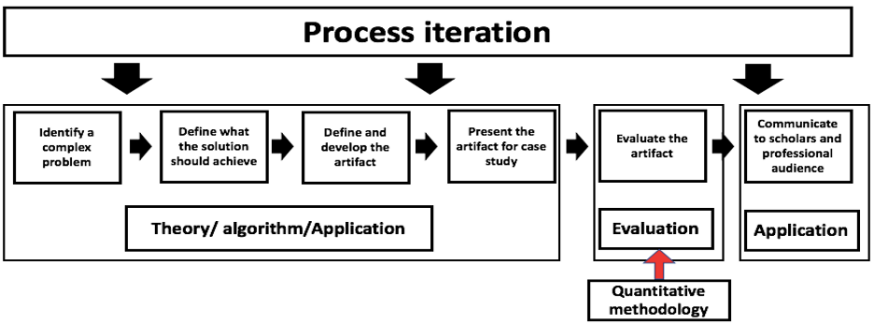


Figure 2: Generic Process Model for a Design Science Research Methodology (DSRM) with research objective

## 4.3 Research Plan

The following breakdown list showing how the project shall be carried out. The task has been grouping into 4 major phases that correspond to the components presented in the prior section. Each task has been assigned an identifier Task1-Task15 for later reference in the project plan.

**Phase I: Background research, problem definition, literature review [Theory]**

**Phase II: Theoretical/ algorithm development [Interpretation]**

**Phase III Testing and evaluate the application [Explanation]**

**Phase IV: Open-Source Framework, complete thesis write up, and publication**

The table listing the research timeline with sub 15 tasks segmented from four Phases in the research plan.

| TASK DESCRIPTION | | | Quarter | | | | Quarter | | | | Quarter | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Start | End | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Phase I: Background research, problem definition, literature review** | | | | | | | | | | | | | | |
| Task 1: Problem identification | 1 | 1 | ■ | | | | | | | | | | | |
| Task 2: QUT Online Course: AIRS001 | 1 | 1 | ■ | | | | | | | | | | | |
| Task 3: Stage 2 (Milestone) Submission | 1 | 1 | ■ | | | | | | | | | | | |
| Task 4: Develop the conceptual framework foundations of a causal abstract model | 2 | 3 | | ■ | ■ | | | | | | | | | |
| Task 5: Literature review | 1 | 9 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | |
| **Phase II: Theoretical/ Algorithm development [interpretation]** | | | | | | | | | | | | | | |
| Task 6: Define and develop a theory for a causal abstract model | 2 | 5 | | ■ | ■ | ■ | ■ | | | | | | | |
| Task 7: Annual report | 4 | 5 | | | | ■ | ■ | | | | | | | |
| Task 8: Annual seminar | 5 | 5 | | | | | ■ | | | | | | | |
| **Phase III: Testing and evaluate the application [Explanation]** | | | | | | | | | | | | | | |
| Task 9: Map interpretations from CAM to causal understandable explanations (Causability) | 4 | 7 | | | | ■ | ■ | ■ | ■ | | | | | |
| Task 10: Build a standard for evaluating the theory-base algorithm | 6 | 7 | | | | | | ■ | ■ | | | | | |
| Task 11: Apply the model in different domains: medical decision-making, etc. | 6 | 8 | | | | | | ■ | ■ | ■ | | | | |
| **Phase IV: Open-Source Framework** | | | | | | | | | | | | | | |
| Task 12: Development of an open-source explainable framework | 6 | 9 | | | | | | ■ | ■ | ■ | ■ | | | |
| Task 13: Thesis publication | 8 | 8 | | | | | | | | ■ | | | | |
| Task 14: Thesis finalisation & submission | 8 | 11 | | | | | | | | ■ | ■ | ■ | ■ | |
| Task 15: Final seminar | 12 | 12 | | | | | | | | | | | | ■ |

Figure 3: Research plan.

## 4.4 Resources and Funding Required

Not applicable

Human evaluation (measure the human understanding by proposed model base on holzinger's framework)

## 4.5 Individual Contribution to the Research Team

The hypothesis that we put forward in the systematic review research is that the inability to disentangle *correlation from causation* can deliver sub-optimal or even erroneous explanations to decision-makers (Richens et al., 2020). Causal approaches should be emphasized in XAI to promote a higher degree of interpretability to its users. In other words, to achieve a certain degree of human-understandable explanations, causability should be a necessary condition.

Given that there is not a clear understanding of the current state of the art concerning causal and causability approaches to XAI, we firstly making a systematic review and critical discussion of the diverse existing body of literature on these topics. This systematic review will introduce researchers in the field of XAI that are interested in focusing their knowledge on the current state-of-the-art approaches currently present in the literature.

In summary, it contributes to a literature review with discussions under three paradigms:

- **Theory.** Survey and formalize the most important theoretical approaches that ground current explainable AI models in the literature that promote the causality.
- **Algorithms.** Understand what are the main algorithms that have been proposed in the XAI literature that is based on probabilistic approaches for causality and which ones have the potential to achieve a certain degree of causability.
- **Applications.** A continuous use case analysis to understand what are the main domains and fields where XAI algorithms that promote causability are emerging and what are the potential advantages and disadvantages of such approaches in real-world problems, namely in the mitigation of biased predictions.

# 5 Progress To date

From literature review section 2, we argued that current XAI algorithms (LIME, SHAP, and Anchor) only indicate a correlation instead of causation for helping humans achieving a higher level of explainability. Therefore, we explored another XAI approach, *counterfactual* by developing systematic review research. This research aims to investigate the theories, algorithms, and applications that underpin XAI approaches which have the potential to achieve *causability*.

## 5.1 Systematic Literature Review Towards Counterfactuals and Causability in XAI

To help researchers identify knowledge gaps in the area of interest by extracting and analyzing the existing approaches, the systematic research will survey the approaches in the extensive body of literature that are primarily based on causality and counterfactual.

Our systematic literature review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework as a standardized way of extracting and synthesizing information available from existing studies with respect to a set of research questions. More specifically, we followed the PRISMA checklist[1] with the study search process presented in the PRISMA flow diagram[2].

Based on the PRISMA, the procedure of systematic review can be separated into several steps: (1) definition of the research questions; (2) description of the literature search process and strategy. Inspired in the recent work of Teh et al. (2020), we conducted a topic modeling analysis to refine the search results using the Latent Dirichlet Allocation (LDA) algorithm together with an inclusion and exclusion criteria to assist with the selection of relevant literature; (3) extraction of publication data (title, abstract, author keywords and year), systemization, and analysis of the relevant literature on counterfactuals and causality in XAI; (4) Lastly, we conducted identification of biases and limitations in our review process.

## 5.2 Research Questions for systematic review

The following systematic review research question not only answer the project research question on *How to induce a causal abstract model (CAM) from a predictive BlackBox that promotes causability*, but also help researchers identify knowledge gaps in the area of causality, causability, and counterfactuals in XAI.

- RQ1: What are the main theoretical approaches for counterfactuals in XAI (Theory)?
- RQ2: What are the main algorithms in XAI that use counterfactuals as a means to promote causal understandable explanations (Algorithms)?
- RQ3: What are the sufficient and necessary conditions for a system to promote causability (Applications)?
- RQ4: What are the pressing challenges and research opportunities in XAI systems that promote Causability?

### 5.2.1 Search Process

To address the proposed research questions, in this paper, we used three well-known Computer Science academic databases: (1) Scopus, (2) IEEE Xplore, and (3) Web of Science (WoS). We

[1] http://www.prisma-statement.org/documents/PRISMA%202009%20checklist.pdf
[2] http://prisma-statement.org/documents/PRISMA%202009%20flow%20diagram.pdf

considered these databases because they have good coverage of works on artificial intelligence, and they provide APIs to retrieve the required data with few restrictions. We used the following search query to retrieve academic papers in the field of artificial intelligence related to explainability or interpretability and causality or counterfactuals.

( *artificial* AND *intelligence* ) AND ( *xai* OR *explai** OR *interpretab** ) AND ( *caus** OR *counterf** )

This query allowed us to extract bibliometric information from different databases, such as publication titles, abstracts, keywords, year, etc. The initial search returned the following articles: IEEE Xplore (6878), Scopus (116), WoS (126). We removed duplicate entries in these results as well as results that had missing entries. In the end, we reduced our search process to IEEE Xplore (4712), Scopus (709), WoS (124). Our strategy is summarized in the PRISMA flow diagram illustrated in Figure 4.

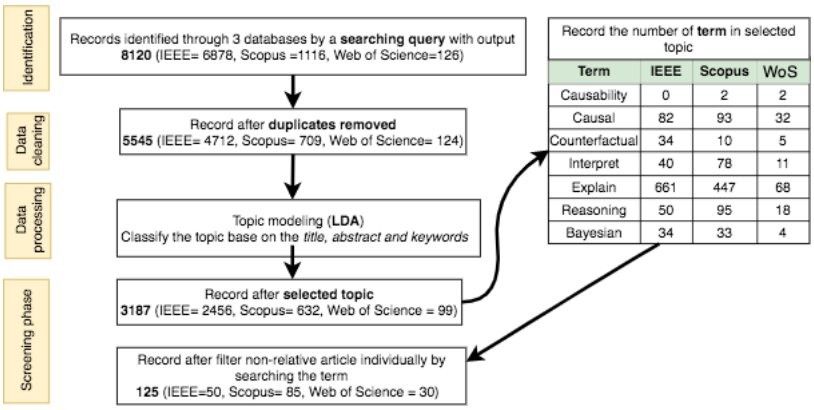


Figure 4: PRISMA flow diagram search results.

To guarantee that the initial query retrieved publications that match this review's scope, we conducted a topic modelling analysis based on Latent Dirichlet Allocation (LDA) to refine our search results.

### 5.2.2 Topic Modelling

Topic modelling is a natural language processing technique that consists of uncovering a document collection's underlying semantic structure based on a hierarchical Bayesian analysis. LDA is an example of a topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions. In our search strategy, LDA enabled us to cluster words in publications with a high likelihood of term co-occurrence and allowed us to interpret the topics in each cluster. This will guarantee that the papers classified within a topic contain all the relevant keywords to address our research questions.

In this paper, we used the title, abstract, and authors' keywords retrieved from the proposed query, and applied several text mining techniques, such as stop word removal, word tokenisation, stemming, and lemmatisation. We then analysed the term co-occurrences with LDA for each database. The best performing model contained a total of 4 topics. The LDA model's output is illustrated in Figure 5 with the inter-topic distance showing the marginal topic distributions (left) and the top 10 most relevant terms for each topic. Analysing Figure 5, Topic 1 contained all the words that are of interest for the research questions proposed in this survey paper: explainability, causality, and artificial intelligence. Topic 2, on the other hand,

has captured words that are primarily related to data management and technology. Topic 3 has words that relate to the human aspect of explainable AI, such as cognition, mental, and human. Finally, Topic 4 contains words associated with XAI in healthcare. For this survey paper, we chose all the publications classified as either Topic 1 or Topic 3. In the end, we were able to reduce our search results to IEEE Xplore (3187), Scopus (632), WoS (99). After manually looking at these publication records and selecting articles about "causability", "causal", "counterfactual", we obtained our final set of documents for analysis: IEEE Xplore (125), Scopus (85), WoS (30).

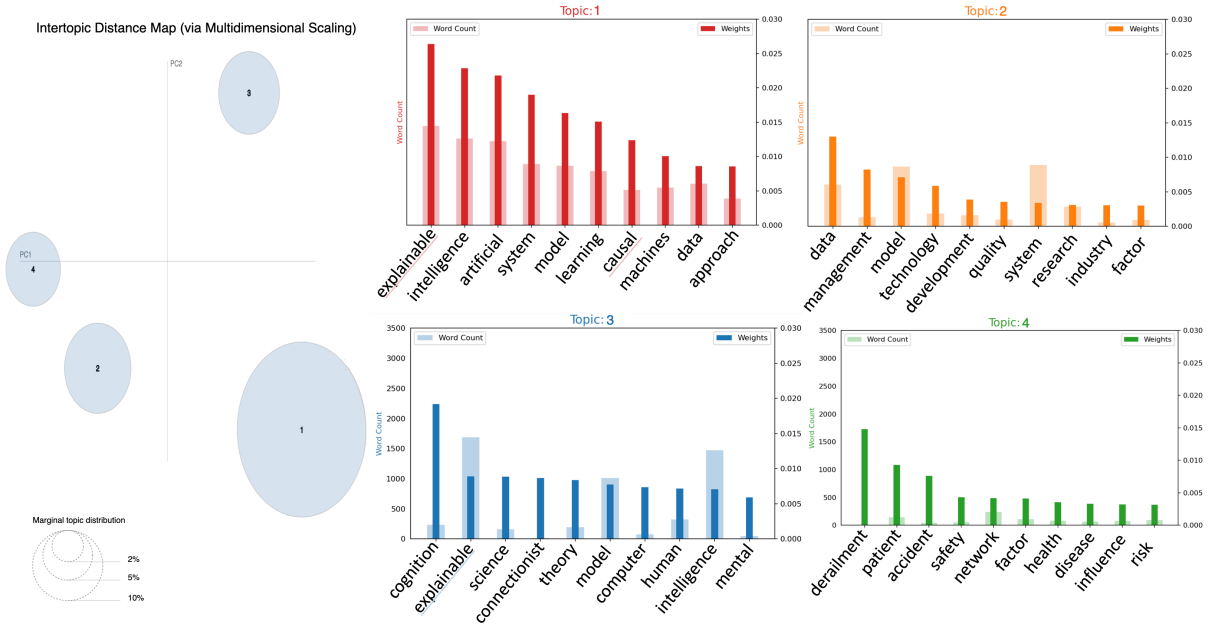


Figure 5: Best performing LDA topic model for Scopus database, using 709 titles, abstracts, and authors keywords found from the proposed search query. Figure also shows the top 10 most relevant words for each Topic.

### 5.2.3 Word Co-Occurrence Analysis

In our survey, we are interested in understanding the necessary and sufficient conditions to achieve causability and how current approaches can promote it. We started to analyse the keyword co-occurrence in the returned documents from our search query to achieve this understanding. We collected the title, abstract, and authors' keywords from the search results in *Scopus*, and filtered the results from using three different keywords of interest: explainable AI, counterfactuals, and causality. This resulted in three different *Scopus* files with the keywords of interest.

To visualise the results, we used the graphical capabilities of *VOS Viewer*[3], which is a software tool for constructing and visualising bibliometric networks.

Figure 6 represents the co-occurrence of authors' keywords regarding the field of XAI. The density plot reveals a shift in research paradigms evolving from *machine-centric* topics to more *human-centric* approaches involving intelligent systems and cognitive systems, to the need of *explainability* in autonomous decision-making.

It is interesting to note that machine-centric research interests (such as pattern recognition or computer-aided diagnostic systems) started to change around 2016. The European Union

[3]https://www.vosviewer.com/

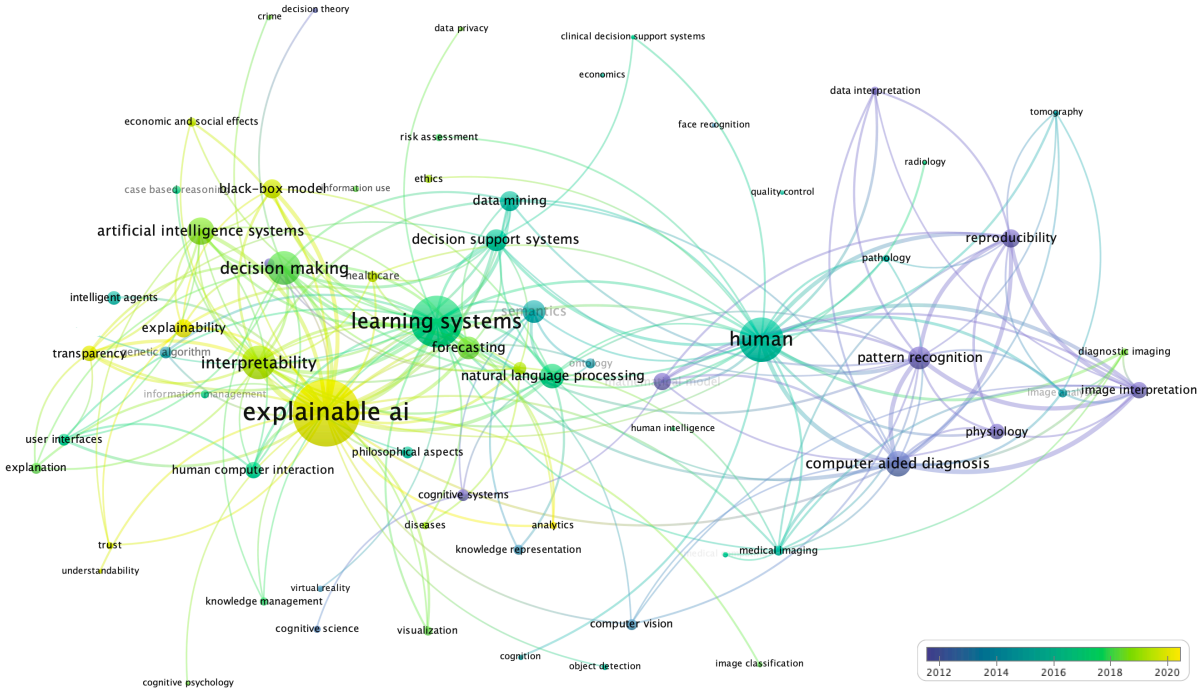


Figure 6: Network visualization of co-occurrence between keywords in articles about XAI.

Commission started to put forward a long list of regulations for handling consumer data, the GDPR. In that year, publications start shifting their focus from fully autonomous systems to a human-centric view of learning systems with a need for interpretability in decision-making. Figure 6 also shows another shift of research paradigms around 2018 towards *explainable AI*, which coincides with the year where GDPR was put into effect in the European Union, imposing new privacy and security standards regarding data access and usage. One of these standards is Article 22, which states that an individual has "the right not to be subject to a decision based solely on automated processing"[4]. In other words, an individual has the right for explainability whenever a decision is computed from an autonomous intelligent system. Given that these systems are highly opaque with complex internal mechanisms, there has been a recent growing need for *transparent* and *interpretable* systems that are able to secure *ethics* and promote user *understandability* and *trust*.

Some researchers argued that for a machine to achieve a certain degree of human intelligence, and consequently, explainability, then counterfactuals need to be considered (Holzinger et al., 2019, 2020). Recently, Miller (2019) stated that explanations need to be counterfactuals ("contrary-to-fact") (Byrne, 1997), since they enable mental representations of an event that happened and also representations of some other event alternative to it (Stepin et al., 2021). Counterfactuals describe events or states of the world that did not occur and implicitly or explicitly contradict factual world knowledge. For instance, in cognitive science, counterfactual reasoning is a crucial tool for children to learn about the world (Wesberg and Gopnik, 2013). The process of imagining a hypothetical scenario of an event that is contrary to an event that happened and reasoning about its consequences is defined as *counterfactual reasoning* (Pereira and Lopes, 2020). We investigated the word co-occurrence in articles involving *explainable AI* and *counterfactuals* to understand how the literature is progressing in this area. Figure 7 shows

[4] https://www.privacy-regulation.eu/en/article-22-automated-individual-decision-making-including-profiling-GDPR.htm

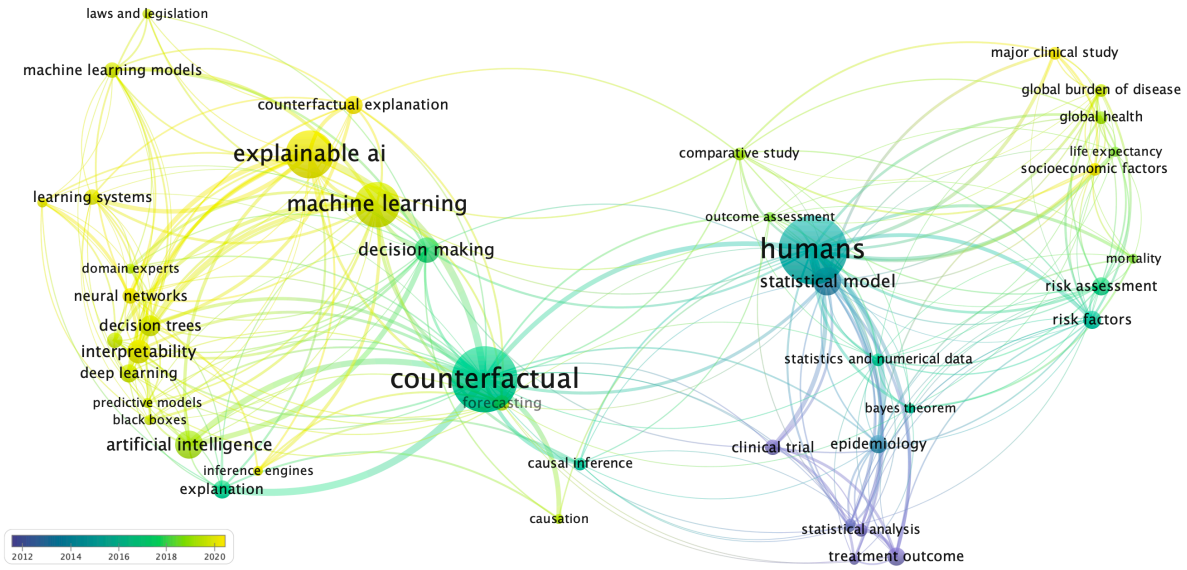


Figure 7: Network visualization of co-occurrence between keywords in articles about counterfactuals in XAI.

the obtained results.

Lau and Coiera (2007)

In the density plot in Figure 7, one can see that counterfactual research in XAI is a topic that has gained interest in the scientific community very recently, with most of the scientific papers dating from 2019 on-wards. This reflects the need for supporting explanations with contrastive effects: by asking ourselves what would have been the effect of something if we had not taken action, or vice versa. Creating such hypothetical worlds may increase the user's understanding of how a system works. The figure seems to be suggesting that the recent body of literature concerned with counterfactuals for XAI is motivated by the medical decision-making domain since we can see relevant keywords such as *patient treatment*, *domain experts*, and *diagnosis*. There is also a recent body of literature in clinical research supporting the usage of counterfactuals and causality to provide interpretations and understandings for predictive systems (Prosperi et al., 2020).

Some researchers also argued that for a machine to achieve a certain degree of human intelligence, then causality needs to be incorporated in the system (Pearl, 2019). Others support this idea in the context of XAI, where they argue that one can only achieve a certain degree of explainability if there is a causal understanding of the explanations, in order words, if the system promotes causability (Holzinger et al., 2019). In this sense, we also analysed co-occurrence between keywords in articles about causality in XAI. Figure 8 illustrates the results obtained.

In terms of causality, Figure 8, one can draw similar conclusions. Although the figure shows a clear connection between Artificial Intelligence and causality (causal reasoning, causal graphs, causal relations), the literature connecting causal relations to explainable AI is scarce. This opens new research opportunities in the area, where we can see from Figure 8 a growing need for counterfactual research. Literature regarding *causability* seems to be also very scarce and very recent. New approaches are needed in this direction, and it is the purpose of this systematic review to understand which approaches for XAI are underpinned by causal theories.

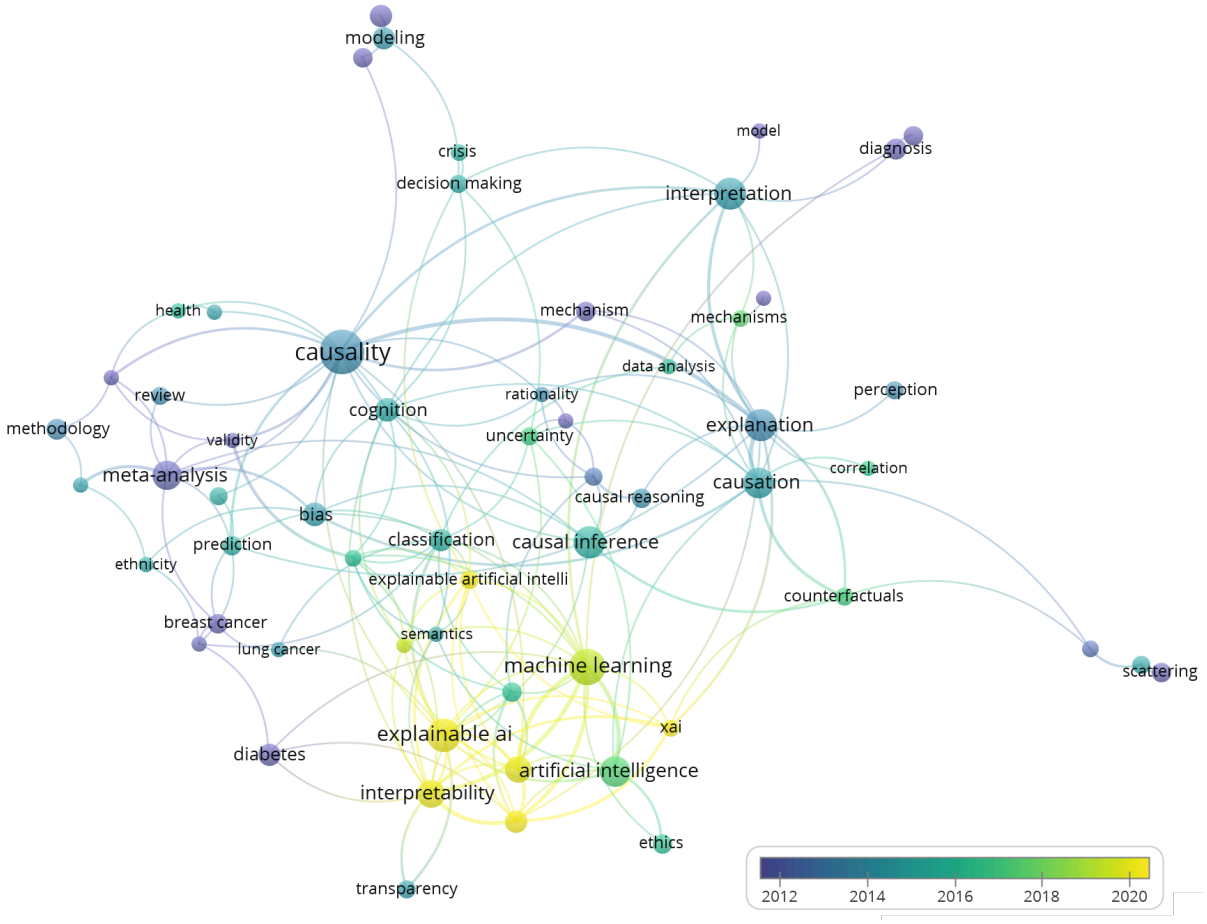


Figure 8: Network visualization of co-occurrence between keywords in articles about causality in XAI.

### 5.2.4 Inclusion and exclusion criteria

To select relevant literature from the obtained search results, we had to consider which papers should be included in our analysis and which ones should be excluded in order to be able to address the proposed research questions. Table 1 summarises the selected criteria.

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Papers about causality in XAI | Papers about causal machine learning |
| Papers about counterfactuals in XAI | Papers about causality |
| Papers about causability | Papers not in English |
| Papers about main algorithms in XAI | Papers without algorithms for XAI |

Table 1: Inclusion and exclusion criteria to assess the eligibility of research papers to analyse in our systematic literature review.

### 5.2.5 Risk of bias

As with any human-driven task, the process of finding relevant research is affected by cognitive biases. In this systematic review, we acknowledge that limiting our search to three databases (Scopes, Web of Science, and IEEE) might have contributed to some missing articles. Databases

that could have complemented our search could be Google Scholar, SpringerLink and PubMed. Another consideration is that we did not extract the references from the collected papers to enrich our search. The collection of retrieved documents was already too big, and we found that doing this would increase exponentially the complexity of the LDA topic analysis that we conducted. Finally, the search query was restricted to keywords that we found relevant to collect the papers of interest. These keywords, however, might have limited our search, and we might have missed relevant articles.

### 5.3 Different types of counterfactual

The systematic review that we conducted allowed us to understand the different counterfactual approaches for XAI. As mentioned throughout this article, counterfactuals have been widely studied in different domains, especially in philosophy, statistics, and cognitive science. Indeed, researchers are arguing that counterfactuals are a crucial missing component that has the potential to provide a certain degree of human intelligence and human-understandable explanations to the field of XAI (Holzinger et al., 2019). Other researchers state that counterfactuals are essential to elaborate predictions at the instance-level (Sokol and Flach, 2020) and to make decisions actionable (Fernandez et al., 2019). Other researcher claimm that counterfactuals can satisfy GDPR's legal requirements for explainability (Wachter et al., 2018).

Figure 9 shows an illustration of several counterfactual candidates for a data instance $x$ according to different works in the literature (Poyiadzi et al., 2020).

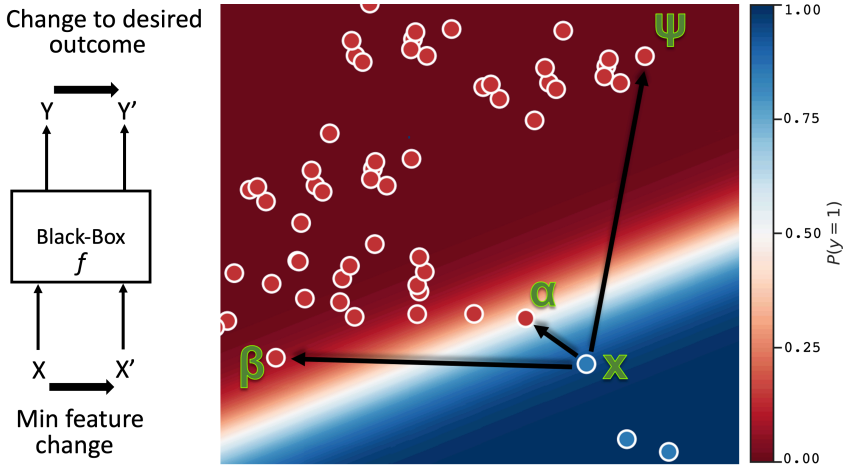


Figure 9: Different counterfactual candidates for data instance $x$. According to many researchers, counterfactual $\alpha$ is the best candidate, because it has the smallest Euclidean distance to $x$ (Wachter et al., 2018). Other researchers argue that counterfactual instance $\gamma$ is the best choice since it provides a feasible path from $x$ to $\gamma$ (Poyiadzi et al., 2020). Counterfactual $\beta$ is another candidate of poor quality because it rests in a less defined region of the decision boundary.

#### 5.3.1 The Importance of Distance Functions in Counterfactual Approaches for XAI

The definition of counterfactual as the minimum distance (or change) between a data instance and a counterfactual instance goes back to the theory proposed by Lewis (1973b). Given a data point $x$, the closest counterfactual $x'$ can be found by solving the problem where $d(.,.)$ is a

measurement for calculating the distance from initial point to generated point.

$$argmin_{x'} d(x, x') \tag{4}$$

One important question that derives from Equation 4 is *what kind of distance function should be used?* Different works in the literature address this optimization problem by exploring different distance functions and $L_p$-norms. This section will review different norms used as distance functions in the literature of XAI and their properties.

In general, a norm measures the size of a vector, but it can also give rise to distance functions. The $L_p$-norm of a vector $x$ is defined as:

$$\|x\|_p = \left( \sum_{i=1}^{n} |x_i|.^p \right)^{1/p} \tag{5}$$

Equation 5 shows that different values of $p$ yields a different distance function with specific properties. The systematic literature review revealed that most works in XAI used either the $L_0$-norm (which is not a norm by definition), the $L_1$-norm (also known as Manhattan distance), the $L_2$-norm (known as the Euclidean distance, and the $L_\infty$-norm. Figure 10 shows a graphical representation of the different norms and the respective contours.

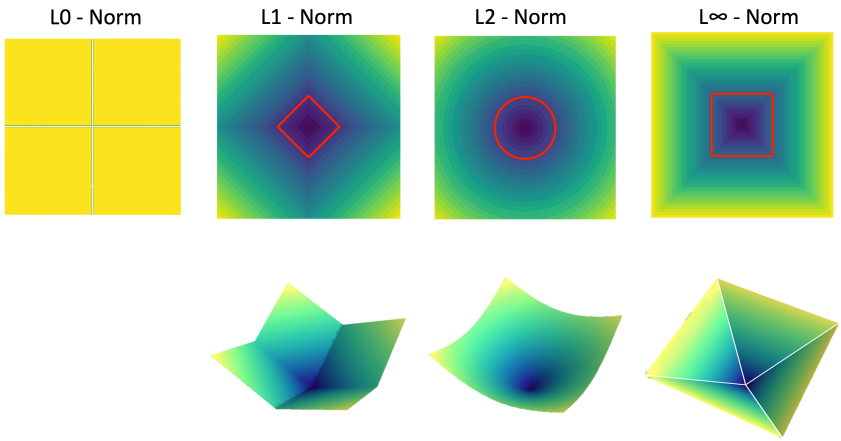


Figure 10: Graphical visualisation of different $L_p$-norms: $L_0$-norms (which is not really a norm by definition), the $L_1$-norm (also known as Manhattan distance), the $L_2$-norm (known as the Euclidean distance, and the $L_\infty$-norm.

- $L_0$**-norm**. The $L_0$-norm has been explored in the context of conterfactuals in XAI primarily by Dandl et al. (2020) and Karimi et al. (2020). Given a vector $x$, it is defined as

$$\|x\|_0 = \sqrt[0]{\sum_i x_i^0}. \tag{6}$$

  Intuitively, $L_0$-norm is the number of nonzero elements in a vector, and it is used to count the number of features that change between the initial instance $x$ and the counterfactual candidate $x'$, resulting in sparse counterfactual candidates (Karimi et al., 2020). Figure 10 shows a visualisation of the $L_0$-norm where one can see that the function is completely undifferentiable, making it very hard to find efficient solutions to minimize it.

- $L_1$**-norm**. The $L_1$-norm (also known as the Manhattan distance) has been the most explored distance function in the literature of counterfactuals in XAI. Wachter et al. (2018) argued that the $L_1$-norm provides the best results for finding good counterfactuals, since it induces sparse solutions. Given a vector $x$, the $L_1$-norm is defined as

$$\|x\|_1 = \sum_i |x_i|. \tag{7}$$

  Intuitively, $L_1$-norm is used to restrict the average change in the distance between the initial instance $x$ and the counterfactual candidate $x'$. Since the $L_1$-norm gives an equal penalty to all parameters and leads to solutions with more large residuals, it enforces sparsity. In Figure 10, one can see that the major problem with $L_1$-norms is its diamond shape, which makes it hard to differentiate.

- $L_2$**-norm**. The $L_2$-norm (also known as the Euclidean distance) has been one of the most explored distance function in the literature of counterfactuals in XAI, although it does not provide sparse solutions as the $L_1$ or $L_0$-norm. Given a vector $x$, the $L_1$-norm is defined as

$$\|x\|_2 = \sqrt{\sum_i x_i^2}. \tag{8}$$

  Intuitively, the $L_2$-norm measures the shortest distance between two points and can detect a much larger error than the $L_1$-norm, making it more sensitive to outliers. Although the $L_2$-norm does not lead to sparse vectors, it has the advantage that it is differentiable. As one can see in Figure 10, its smoothness and its rotational invariance (a circle or a hyper-sphere in higher dimensions) are both desirable properties in many optimization problems, making it computationally efficient.

- $L_\infty$**-norm**. The $L_\infty$-norm has been explored in the context of counterfactuals in XAI primarily by Karimi et al. (2020). Given a vector $x$, it is defined as

$$\|x\|_\infty = \sqrt[\infty]{\sum_i x_i^\infty} = max(|x_i|). \tag{9}$$

  Intuitively, $L_\infty$-norm is used to restrict maximum change across features. the maximum change across the features between the initial instance $x$ and the counterfactual candidate $x'$ (Karimi et al., 2020). Computationally, the $L_\infty$-norm is differentiable in every point, except when at least two features $x_i$ have the same absolute values $|x_i|$, which is illustrated in Figure 10. By minimizing the L-infinity norm, we are penalizing the cost of the largest feature, which leads to less sparse solutions when compared to $L_0$-norm or $L_1$-norm.

Distance functions in counterfactuals are associated with the sparsity of the vector, which is a highly desirable property to have when looking for counterfactuals. The minimum the changes we can have in the features, the better and more human interpretable counterfactuals we will find. In the next section, we will present the main properties that a theory for counterfactuals in XAI should satisfy.

### 5.3.2 Properties to Generate Good Counterfactuals

The main properties for generating an optimal counterfactual explanation as listed below:

- **Proximity.** Proximity calculates the distance of a counterfactual from the input data point while generating a counterfactual explanation, (Verma et al., 2020). As mentioned in Section 5.3.1, many different distance functions can be used to measure proximity, resulting in counterfactual candidates with different properties. Other works in the literature consider another type of proximity measures such as Nearest Neighbour Search (Keane and Smyth, 2020) or cosine similarity (Martens and Provost, 2014).

- **Plausibility.** This property is similar to the terms *Actionability* and *Reasonability* referred in Verma et al. (2020); Keane and Smyth (2020). It emphasizes that the generated counterfactuals should be legitimate, and the search process should ensure logically reasonable results. This means that a desirable counterfactual should never change *immutable* features such as gender or race. When explaining a counterfactual, one cannot have explanations like *"if you were a man, then you would be granted a loan"*, since these would show an inherent bias in the explanation. Mutable features, such as income, should be changed instead to find good counterfactuals.

- **Sparsity.** This property is related to the methods used to efficiently find the minimum features that need to be changed to obtain a counterfactual (Keane and Smyth, 2020).

  In cognitive science, counterfactuals are used as a process of a imagining hypothetical scenario of an event that is contrary to an event that happened, and reasoning about its consequences (Pereira and Lopes, 2020). To be human-understandable and interpretable, it is desired that these counterfactuals are sparse and with the fewest possible changes in their features to be effective. In Mothilal et al. (2020), for instance, the authors elaborate that sparsity is assessing how many features a user needs to change to transition to the counterfactual class. On the other hand, Verma et al. (2020) argues that sparsity can be seen as a trade-off between the number of features and the total amount of change made to obtain the counterfactual. Wachter et al. (2018) also stands on this idea and asserts that pursuing the "closest possible world", or the smallest (minimum-sized) change to the world that can be made to obtain a desirable outcome.

- **Diversity.** This property was introduced in the work of Russell (2019) and also explored in Ramaravind K. Mothilal (2020); Karimi et al. (2020). Since finding the closest points of an instance $x$ according to a distance function can lead to very similar counterfactual candidates with small differences between them, diversity was introduced as the process of generating a set of diverse counterfactual explanations for the same data instance $x$ (Karimi et al., 2020). This would allow the user to choose counterfactuals that are more understandable and interpretable, leading to a higher degree of explainability.

- **Feasibility.** This property was introduced by Poyiadzi et al. (2020) as an answer to the argument that finding the closest counterfactual to a data instance does not necessarily lead to a feasible change in the features. Going back to Figure 9 in Section **??**, one can see different counterfactual candidates. The closest counterfactual to the data instance $x$ is $\alpha$, however, it falls in the decision boundary, which could lead to the biased counterfactual, since the black-box is not very certain about its class. To address this, Poyiadzi et al. (2020) argues that counterfactual $\gamma$ is a better one because it falls in a well-defined region of the decision boundary and also corresponds to the point that has the shortest path to $x$. This way, it is possible to generate more confident counterfactuals with the least possible feature changes.

Given the above properties, in the next sections, we will classify the different algorithms found in the literature by (1) their underlying theory (Section 5.4), and by (2) the above

properties (Section 5.5).

## 5.4 Counterfactual Approaches to Explainable AI: The Theory

The systematic literature review contributed to developing a new taxonomy for the model-agnostic counterfactual approaches for XAI. Throughout the review process, we noticed that many algorithms derived from similar theoretical backgrounds. In total, we analysed 19 algorithms. We created a set of six different categories representing the *"master theoretical algorithm"*(Domingos, 2017) from which each algorithm derived. These categories are (1) instance-centric approaches, (2) constraint-centric approaches, (3) genetic-centric approaches, (4) regression-centric approaches, (5) game theory centric Approaches, and (6) Case-Based Reasoning Approaches. Figure 11 presents the proposed taxonomy as well as the main algorithms that belong to each category.

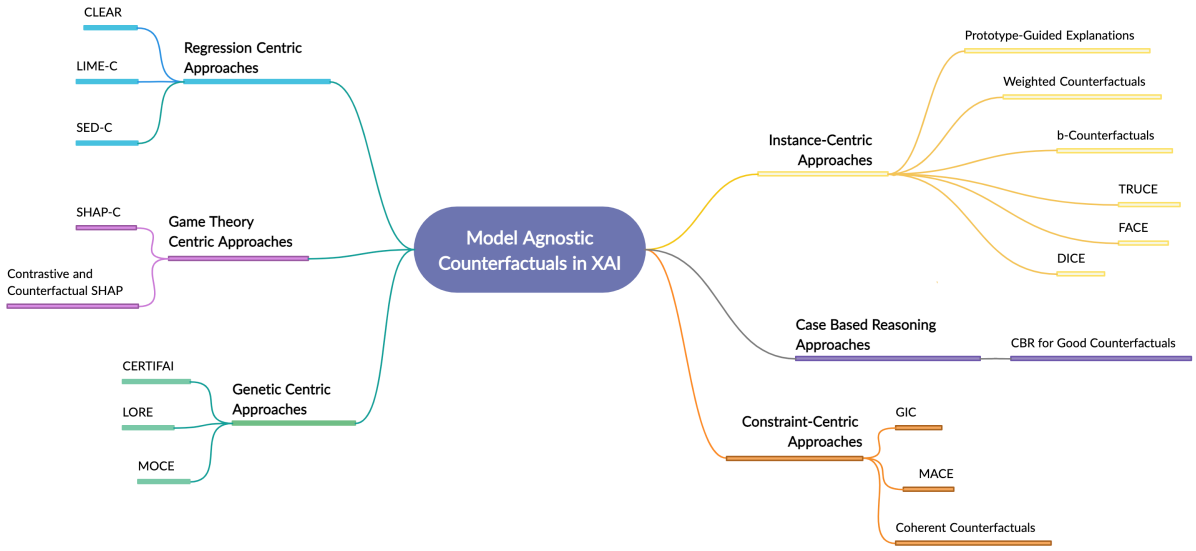


Figure 11: Taxonomy for model-agnostic counterfactual approaches for XAI.

- **Instance-Centric.** Correspond to all approaches that derive from the counterfactual formalism proposed by Lewis (1973b) and Wachter et al. (2018). These approaches are based on random feature permutations and on finding counterfactuals closed to the original instance by some distance function. Instance-centric algorithms seek novel loss functions and optimization algorithms to find counterfactuals. These approaches are susceptible to fail the plausibility and the diverse properties of counterfactuals, although some algorithms incorporate mechanisms in their loss functions to overcome these issues.

- **Constraint-Centtric.** Corresponds to all approaches that are modeled as a constraint satisfaction problem. Algorithms that fall in this category use different strategies to model the constraint satisfaction problem, such as satisfiability modulo theory solvers. The big advantage of these approaches is that they are general and can easily satisfy different counterfactuals properties such as diversity and plausibility in their model.

- **Genetic-Centric.** Corresponds to all approaches that use genetic algorithms as an optimization method to search for counterfactuals. Since genetic search allows feature vectors to crossover and mutate, counterfactual algorithms based on these approaches often satisfy properties such as diversity and plausibility.

- **Regression-Centric.** Corresponds to all approaches that generate explanations by using the weights of a regression model. These approaches are very similar to LIME. The intuition is that an interpretable model (in this case, linear regression) fits the newly generated data after permuting the features, and the weights of each feature presented explanations. Counterfactuals based on these approaches have difficulties satisfying several properties such as plausibility and diversity.

- **Game Theory Centric.** Corresponds to all approaches that generate explanations by using Shapley values. These approaches are very similar to SHAP. Algorithms that fall in this approach mainly extend the SHAP algorithm to take into consideration counterfactuals. Counterfactuals based on these approaches have difficulties satisfying several properties such as plausibility and diversity.

- **Case-Based Reasoning.** Corresponds to all approaches inspired in the case-based reasoning paradigm of artificial intelligence and cognitive science that models the reasoning process as primarily memory-based. These approaches often solve new problems by retrieving stored *cases* describing similar prior problem-solving episodes and adapting their solutions to fit new needs. In this case, the CBR system stores good counterfactual explanations. The counterfactual search process consists of retrieving from this database the closest counterfactuals to a given query. CBR approaches can easily satisfy different counterfactual properties, such as plausibility and diversity.

### 5.5 Counterfactual Approaches to Explainable AI: The Algorithms

In this systematic review, we found 18 model agnostic XAI counterfactual algorithms. We analyzed each algorithm in-depth and classified them according to the different properties presented in Section 5.3. We also classified them in terms of their applications, either for classification/regression problems and the supporting data structures. We complemented the analysis with the information of whether the algorithm is publicly available. Table 2 presents a classification of collected model-agnostic counterfactual algorithms for XAI based on different properties, theoretical backgrounds, and applications.

In the next sections, each algorithm of Table 2 is analysed relatively to its grounding *theoretical master algorithm*.

### 5.6 Instance-Centric Algorithms

In this section, we summarize the algorithms that we classified as instance-centric using the proposed taxonomy. By definition, these algorithms are very similar, diverging primarily on the loss function description with the corresponding optimization algorithm and the distance function specification.

- $b$-**Counterfactuals by Wachter et al. (2018).** $b$-Counterfactuals correspond to one of the first algorithms in model-agnostic counterfactuals for XAI. It extends the notion of a minimum distance between datapoints, originally proposed by Lewis (1973b), by adding a trade-off term between the quadratic distance between the model prediction for the counterfactual candidate $x'$ and the desired outcome $y$, and the distance between the counterfactual candidate $x'$ and the original data instance $x$.

  - **Loss function.** The loss function proposed by (Wachter et al., 2018) takes as input the data instance to be explained, $x$, the counterfactual candidate, $x'$, and a parameter $\lambda$, balances the distance in the prediction (first term) against the distance in feature

values (second term) (Molnar, 2020). The higher the value of $\lambda$, the closer the counterfactual candidate, $x'$, is to the desired outcome, $y'$. Equation 10 presents the loss function and respective optimization problem proposed by (Wachter et al., 2018).

$$\begin{aligned} &\mathcal{L}(x, x', y', \lambda) = \lambda (f(x') - y')^2 + d(x, x') \\ &argmin_{x'} max_{\lambda} \mathcal{L}(x, x', y', \lambda) \end{aligned} \tag{10}$$

- **Distance function.** Wachter et al. (2018) argue that the $L_1$-norm normalized by the inverse of the median absolute deviation of feature $j$ over the dataset is one of the best performing distance functions because it ensures the sparsity of the counterfactual candidates. Equation 11 presents the distance function used in their loss function.

$$\begin{aligned} d(x, x') &= \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}, \text{ where} \\ MAD_j &= median_{i \in \{1,\dots,n\}} \left( \left| x_{i,j} - median_{l \in \{1,\dots,n\}}(x_l, j) \right| \right) \end{aligned} \tag{11}$$

- **Optimization algorithm:** The Adam Gradient descent algorithm is used to minimize Equation 10.

- **Prototype Guided Explanations by Looveren and Klaise (2019).** Class prototypes were originally proposed by Looveren and Klaise (2019) which consist in guiding the perturbations towards an interpretable counterfactual. This method has the advantage to eliminate the computational problem from the optimization process that transpires by numerical gradient calculation for black-box models.

  - **Loss function**: Prototype Guided Explanations has two aims for defined a loss function: (1) guide the perturbations $\delta$ towards an interpretable counterfactual $x_{cf}$ which falls in the distribution of counterfactual class $i$ (2) Accelerate the counterfactual searching process which achieved by:

$$Loss = c.L_{pred} + \beta.L_1 + L_2 + L_{AE} + L_{proto}, \tag{12}$$

    The $L_{pred}$ measures the divergence between the class prediction probabilities, $L_1$ and $L_2$ correspond to the elastic net regularizer proposed by **?**, $L_{AE}$ represents an autoencoder loss term that penalizes out-of-distribution counterfactual candidate instances (which can lead to uninterpretable counterfactuals), and finally $L_{proto}$ is used to speed up the search process by guiding the counterfactual candidate instances towards an interpretable solution.

  - **Distance function.**Looveren and Klaise (2019) use the $L_2$-norm to find the closest encoding of perturbed instances, $ENC(x + \delta)$ of that data instance, $x$, to its prototype class, $proto_i$. This is given by Equation **??**.

$$L_{proto} = \theta \, \| ENC(x + \delta) - proto_i \|_2^2 \tag{13}$$

  - **Optimization function**: Looveren and Klaise (2019) the optimize function was triggered by adopted a fast integrative threshold algorithm (FISTA) which helps the perturbation parameter $\delta$ to reach momentum for N optimization steps. The $L1$ regularization has been adopted into the optimization function.

- **Weighted Counterfactuals by Grath et al. (2018).** Weighted counterfactuals proposed by Grath et al. (2018) that extends the *b*-counterfactual approach in two dimensions: (1) It proposes the concepts of positive and weighted counterfactuals (2) proposes two weighting strategies to generate more interpretable counterfactual.

  While traditional counterfactuals address the question *why my load was not granted?* through a hypothetical *what-if* scenario, positive counterfactuals address the question *How much was I accepted a loan by?* when the desired outcome is reached.

  - **Loss function.** The weighted counterfactuals are computed in the same way as the *b*-counterfactuals (Wachter et al., 2018) as expressed in Equation 10.
  - **Distance function.** The distance function used to compute weighted counterfactuals is the same as in *b*-counterfactuals (Wachter et al., 2018), with the addition of a weighting parameter $\theta_j$,

$$d(x, x') = \sum_{j=1} \frac{\left|x_j - x'_j\right|}{MAD_j} \theta_j. \tag{14}$$

  - **Optimization algorithm.** While Wachter et al. (2018) used gradient descent to minimize the loss function, Grath et al. (2018) used the Nelder-Mead algorithm, which was originally suggested in the book of Molnar (2020) and is used to find the minimum of the loss function in a multidimensional space. The Nelder-Mead algorithm is a better algorithm to deal with the $l_1$-norm since it works well with nonlinear optimization problems for which derivatives may not be known.

  Experiments conducted by (Grath et al., 2018) showed that weights generated from feature importance lead to more compact counterfactuals, and consequently offered more human-understandable interpretable features.

- **Feasible and Actionable Counterfactual Explanations () by Poyiadzi et al. (2020).**

  FACE created by Poyiadzi et al. (2020) aims to build coherent and feasible counterfactuals by using the shortest path distances defined via density-weighted metrics. This approach allows the user to impose additional feasibility and classifier confidence constraints naturally and intuitively. Moreover, FACE uses Dijkstra's algorithm to find the shortest path between existing training datapoints and the data instance to be explained (Verma et al., 2020).

  - **Main Function.** The main function of FACE's algorithm is given by Equation 15, where $f$ corresponds to a positive scalar function and $\gamma$ is a function that connects the path between a data instance $x_i$ and a counterfactual candidate instance $x_j$.

$$\begin{aligned} \hat{\mathcal{D}}_{f,\gamma} &= \sum_i f_p\left(\frac{\gamma(t_{i-1}) + \gamma(t_i)}{2}\right) . \left\| . \gamma(t_{i-1}) - \gamma(t_i) \right\|, \text{where} \\ \hat{\mathcal{D}}_{f,\gamma} &= \int_\gamma f(\gamma(t)) . \left|\gamma'(t)\right| dt. \end{aligned} \tag{15}$$

    When the partition $\hat{\mathcal{D}}_{f,\gamma}$ converges, Poyiadzi et al. (2020) suggest, for a given threshold $\epsilon$, using weights of the form

$$\begin{aligned} w_i, j = f_p\left(\frac{x_i + x_j}{2}\right) . \left\|x_i - x_j\right\|, \\ \text{when} \quad \left\|x_i - x_j\right\| \leq \epsilon. \end{aligned} \tag{16}$$

The $f$-distance function is used to quantify the trade-off between the path length and the density in the path. This can subsequently be minimized using Dijkstra's shortest path algorithm by approximating the $f$-distance using a finite graph over the data set.

- **Distance Function.** Poyiadzi et al. (2020) used the $l_2$-norm in addition to Dijkstra's algorithm to generate the shortest path between a data instance $x_i$ and a counterfactual candidate instance $x_j$.
- **Optimization Function.** Poyiadzi et al. (2020) suggest three approaches that can be used to estimate the weights in Equation 16:

$$\begin{aligned} w_{i.j} &= f_{\hat{p}}\left(\frac{x_i + x_j}{2}\right).\left\|x_i - x_j\right\| \\ w_{i.j} &= \tilde{f}\left(\frac{r}{\left\|x_i + x_j\right\|}\right).\left\|x_i - x_j\right\|, \quad r = \frac{k}{N.\eta_d} \\ w_{i.j} &= \tilde{f}\left(\frac{\varepsilon^d}{\left\|x_i + x_j\right\|}\right).\left\|x_i - x_j\right\| \end{aligned} \tag{17}$$

The first equation requires using Kernel Density Estimators to allow convergence, the second requires a k-NN graph construct, and the third equation requires $\epsilon$-graphs. In their experiments, Poyiadzi et al. (2020) found that the third weight equation together with $\epsilon$-graphs generated the most feasible counterfactuals.

- **Diverse Counterfactual Explanations (DiCE) by Ramaravind K. Mothilal (2020).** Originally proposed by (Ramaravind K. Mothilal, 2020), DICE is an extension and improvement of the $b$-counterfactuals (Wachter et al., 2018) throughout different properties: Diversity, proximity, and sparsity. DiCE generates a set of diverse counterfactual explanations for the same data instance $x$, allowing the user to choose counterfactuals that are more understandable and interpretable. Diversity is formalized as a determinant point process, which is based on the determinant of the matrix containing information about the distances between a counterfactual candidate instance and the data instance to be explained.

  - **Loss Function.** In DiCE, the loss function is given as a linear combination of a hinge loss function that is a metric that minimizes the distance between the user prediction $f(.)'$ for $c_i s$ and an ideal outcome $y$ $loss(f(c_i), y)$, a proximity factor, which is given by a distance function, and a diversity factor $dpp\_diversity(c_1, ..., c_k)$:

$$\begin{aligned} C(x) = \underset{c_1,...,c_k}{\operatorname{argmin}} \frac{1}{k}\sum_{i=1}^{k} y loss(f(c_i), y) + \frac{\lambda_1}{k}\sum_{i=1}^{k} dist(c_i, x) \\ -\lambda_2\, ddp\_diversity(c_1, ..., c_k) \end{aligned} \tag{18}$$

  - **Distance Function.** DiCE uses the $L_1$-norm normalized by the inverse of the median absolute deviation of feature $j$ over the dataset, just like in $b$-counterfactual (Wachter et al., 2018).
  - **Optimization Function.** Gradient descent is used to minimize Equation 18.

- **Unjustified Counterfactual Explanations (TRUCE)** Laugel et al. (2019) generated an instance-based approach named growing sphere for an explanation. It represents a

principle that consists of determining the minimal changes to alter a prediction. It offers post-hoc explanations for a data instance through the comparison of output with its closest data point(feature).

– **Loss Function.** To simplify the process for reaching a closest desirable feature, Laugel et al. (2019) presented a formalization for binary classification by finding an observation value $e$, then classified it into a different class other than $x$. For instance, $f(e) \neq f(x)$, indicates that the observation has been classified into the same class as $x$ by the classifier, and a desirable feature has been found if it is classified to the other class. For the next step, a function has been defined $c : X \times X \rightarrow R^+$ such that $c(x, e)$ is the cost of moving from observation $x$ to enemy $e$.

$$\begin{aligned} e^* &= \arg\min_{e \in X} \{c(x,e) \quad | \quad f(e) \neq f(x)\} \\ c(x,e) &= \|x-e\|_2 + \gamma \|x-e\|_0 \end{aligned} \tag{19}$$

The first equation is a minimization function which need to be addressed, this problem arising from defining the cost function $c$ is that, even if the classifier is created for learning and optimizing certain loss function, the considered black-box hypothesis is complied to select a different metric. In the second equation Laugel et al. (2017) associate $c$ with the weight associated to the vector sparsity and $\|.\|_2$ the Euclidean norm.

– **Distance Function.** $l_2$ and $L_0$ were considered for measure distance, $l_2$ norm has been used in the vector $e - x$ as a component of a cost function to calculate the proximity between $x$ and $e$. On the other hand, $l_0$ norm is used in the cost function $c$ and combined with $l2$ as a weighted average to guarantee the explanation is human-interpretability.

– **Optimization Function.** A growing sphere algorithm was conducted to handle the problem which shows in the first equation as a greedy method to find the closest feature in all possible directions until the decision boundary has been reached which successfully minimized the $l_2$ norm. Besides, a growing sphere could help the model remove several features by minimizing the $L_0$ of the cost function and contributes the final solution in $e^*$.

### 5.6.1 Constraint-Centric Approaches

In this section, we summarize the algorithms that we classified as constraint centric using the proposed taxonomy.

- **Model-Agnostic Counterfactual Explanations for Consequential Decisions by Karimi et al. (2020)** Originally proposed by Karimi et al. (2020), MACE maps the problem of counterfactual explanation search into a satisfiability modulo theory (SMT) model. More specifically, given a sequence of satisfiability problems expressing the predictive model, $f$, the distance function, $d$, and the constraint functions, the goal of MACE is to map these sequences into logical formulae, and verify if there exists a counterfactual explanation that satisfies a distance smaller than some given threshold. The constraints that are taken into consideration in this approach are *plausibility* and *diversity*. This is achieved in the following way. Given the counterfactual logical formula $\phi_{CF_f(\hat{x})}$, the distance formula $\phi_{d,\hat{x}}$,

constraints formula $\phi_{g,\hat{x}}$, and a threshold $\epsilon$, they are combined into the counterfactual formula, $\phi_{\hat{x},\delta}(x)$, given by

$$\phi_{\hat{x},\delta}(x) = \phi_{CF_f(\hat{x})}(x) \wedge \phi_{d,\hat{x}} \wedge \phi_{g,\hat{x}}, \tag{20}$$

and used as input for a SMT solver, $SAT(\phi_{\hat{x},\delta}(x))$, which will find counterfactuals that will satisfy the conditions with a distance smaller than $\epsilon$. MACE is a general algorithm that supports any $L_p$-norm as a distance function, as well as any number of constraints.

- **Coherent counterfactuals by Russell (2019)**

  Originally proposed by Russell (2019), this approach focuses on generating diverse counterfactuals based on "mixed polytope" methods which aims to handle complex data with contiguous range or an additional set of discrete states. Russell (2019) created a concrete method for generating diverse counterfactual to map back into the same space as the original data. The author also provides a novel set of criteria for generating diverse counterfactuals and integrate them with the mixed polytope method. Before achieving the two targets (coherent and diversity), Russell (2019) firstly offers a solution on generating a counterfactual which is based on Wachter et al. (2018)'s $b$-counterfactuals. To increase the stability when using the Lagrangian approach to generate counterfactual explanations, Russell (2019) changed a linear program when $f$ is linear and distance function $d$ takes the form of a weighted $l_1$ norm. Where they occur, binary constraints (such as this variable must take only values 0 or 1) are treated as integer constraints. The following equation obeys on Wachter et al. (2018)'s suggestion, making use of the $L_1$ norm, weighted by the inverse Median Absolute Deviation, which present as $||||1, MAD$.

$$\min_{\mathrm{x'}} \max_{\lambda} \; ||\mathrm{x\text{-}x'}||1, \mathrm{MAD} + \lambda \; \mathrm{f(x)\text{-}c})^2 \tag{21}$$

  *Coherent*:

  To achieve coherent counterfactual while generating a counterfactual, Russell (2019) set a proper encoding which enables a linear classifier to be trained on it instead of original data points. This could help the model output a substantially higher performance with coherent counterfactual by tackling an issue that the extra degrees of freedom allow unrealistic counterfactual states which do not map back into the original data space when using embedding into higher-dimensional spaces for computing a counterfactual.

- **Generalized Inverse Classification by Lash et al. (2017)**

  Inverse classification is similar to the sub-discipline of sensitivity analysis, which examines the impact of predictive algorithm input on the output. Lash et al. (2017) proposed Inverse Classification framework which mainly focuses on optimizing the generation through a process of perturbing an instance, and this task is achieved by operating on features that could act immediately and tracking each change that leads to an individual cost. Every change in perturbing an instance is subject to happen within a certain level of cumulative change. This result helps the framework estimate the feature by changes a consequence of direct action taken.

  To assess the capability of the GIC, Lash et al. (2017) applied five methods including three heuristic-based methods to solve the generalized inverse classification problem including hill-climbing + local search (HC+LS), a genetic algorithm (GA), and a genetic algorithm + local search (GA+LS). The other two methods applied are sensitivity analysis-based methods such as Local Variable Perturbation–Best Improvement (LVP-BI) and Local

Variable Perturbation–First Improvement (LVP-FI). These five algorithms were conducted into an experiment for examining the average likelihood of test instances conforming to a non-ideal class over varying budget constraints. And the final result shows that LVP-FI outperforms all other methods, while LVP-BI is comparable to GA and GA+LS. HC+LS has the worst performs.

The GIS permits the use of virtually any classification function, requiring only a simple non-prohibitive assumption. Besides, Lash et al. (2017) refined an existing inverse classification framework to include non-linear cost-to-change functions.

### 5.6.2 Genetic-Centric Approaches

In this section, we summarize the algorithms that we classified as genetic-centric using the proposed taxonomy.

- **Local Rule-Based Explanations of Black Box Decision Systems by Guidotti et al. (2018)**

  Originally proposed by Guidotti et al. (2018), this approach attempts to provide interpretable and faithful explanations by learning a local interpretable predictor on a synthetic neighborhood generated by a genetic algorithm. Explanations are generated by decision rules that derive from the underlying logic of the local interpretable predictor.

  LORE works as follows. Given a black-box predictor and a local counterfactual instance, $x$, with outcome $y$, an interpretable predictor is created by generating a balanced set of neighbor instances of the given instance $x$ using an ad-hoc genetic algorithm. The interpretable model used to fit the data corresponds to a decision-tree from which sets of counterfactual rules can be extracted as explanations.

  The distance function used in this algorithm is given by Equation 22. The fitness function used corresponds to the distance of $x$ to a generated counterfactual candidate $z$, $d(x,z)$. This algorithm also considers the mixed types of features by a weighted sum of the simple matching coefficient for categorical features, and by using the $L_2$-norm to normalize the continuous features. Assuming $h$ corresponds to categorical features and $m-h$ to continuous ones, then the distance function is given by

$$d(x,z) = \frac{h}{m}.SimpleMatch(x,z) + \frac{m-h}{m}.NormEuclid(x,z). \tag{22}$$

- **Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models (CERTIFAI) by (Sharma et al., 2019)** CERTIFAI is a custom genetic algorithm based explanation proposed by Sharma et al. (2019) with several strengths including the capability of evaluating the robustness of a machine learning model (CERScore) and assessing fairness with linear and non-linear models and any input form (from mixed tabular data to image data) without any approximations to or assumptions for the model.

  Establishing a CERTIFAI framework comes with several steps which start by creating an original genetic framework, select of distance function, and improving counterfactuals with constraints. In the first stage, a custom genetic algorithm was made by considering $f$ as a classifier for a black box model and an instance $x$ as an input. In this formalist, consider $c$ as the counterfactual candidate instance of $x$ and $d(x,c)$ the distance between them. The distance function the distance used is the $L_1$ norm normalized by the median absolute deviation (MAD), as proposed by (Wachter et al., 2018), and the goal is to minimize the distance between $x$ and $c$,

To minimise this distance function, Sharma et al. (2019) applied a genetic algorithms to search for counterfactuals.

Given a variable $W$ defined as space from which individuals can be generated, to ensure feasible solutions. multiple constraints need to be created for matching with continuous, categorical, and immutable features. For instance, $W_1, W_2..., W_n$ is define for continuous features constraint as $W_i \in W_{imin,W_{imax}}$ and $W_i \in W_1, W_2, ..., W_j$ is for categorical variable. Finally, a feature $i$ for an input $x$ should be muted by setting $W_i = X_i$.

The robustness and fairness of the population of counterfactuals generated if given by

$$CERScore(model) = \frac{E}{x}[d(x, c^*)] \tag{23}$$

Fairness ensures that the solutions generated contain different counterfactuals with multiple values of an unchangeable feature (e.g., gender, race).

- **Multi-Objective Counterfactual Explanations (MOCE) by (Dandl et al., 2020).** Dandl et al. (2020) proposed a multi-objective counterfactual explanation algorithm which translates the counterfactual search into a multi-objective optimization problem based on genetic algorithms. This approach not only brings the benefit of providing a diverse set of counterfactuals with a variety of trade-offs between the proposed objectives but also maintains diversity in feature space at the same time.

  Dandl et al. (2020) proposed a four-objective loss equation to achieve an to present an explanation:

$$L(x, x', y', X^{obs}) = (o_1(f\hat{}(x', y'), o_2(x, x'), o_3(x, x'), o_4(x', X^{obs}))) \tag{24}$$

  In the proposed equation the four objectives $o_1 too_4$ represent one of the four criteria: Objective 1, $o_1$, focuses on generating the close as possible result from a prediction of counterfactual $x'$ to the desired prediction $y'$. It minimizes the distance between $f(x')$ and $y'$, and calculates it by the $L_1$1-norm. Objective 2 reflects that the ideal counterfactual should be as similar as possible to instance $x$. It quantifies the distance between $x'$ and $x$ as the Grower distance. Objective 3, $o_3$ is generated to calculate the sparse feature changes by the $L_0$-norm due to a grower distance that can handle both numerical and categorical features, but lack of capability to count how many features were changed. Finally, Objective 4 reflects that the ideal counterfactual should have likely feature value combinations. The solution to measure how "likely" a data point is using the training data or other dataset which can be inferred, then, Dandl et al. (2020) denote this dataset as $Xobs$ as an approximation for the likelihood and used $o4$ measures the average Grower distance between $x'$ and the nearest observed data point $x[1] \in Xobs$.

  It is noticed that MOCE has no balancing and weighting terms like $\lambda$ as proposed by Wachter et al. (2018), to optimize all four objects $o_1 too_4$ in the same time, Dandl et al. (2020) used the Nondominated Sorting Genetic Algorithm or short NSGA-II which is a nature-inspired algorithm and applies Darwin's law of the survival of the fittest and denote the fitness of a counterfactual by its vector of objectives values $(o1, o2, o3, o4)$, this solution helps to produce a fitter counterfactual result by showing the lower counterfactuals four objectives.

  Dandl et al. (2020) emphasize that MOCE is still limit on the ability to assess other data than binary classification, it remains an open question on how to offer users select the counterfactual that meet their prior unknown to trade off between objective.

### 5.6.3 Regression-Centric Approaches

In this section, we summarize the algorithms that we classified as regression-centric using the proposed taxonomy:

- **CLEAR.**: Counterfactual Local Explanations via Regression was introduced by White and d'Avila Garcez (2019) to underlying a classifier for local explanation models. This method aims to provide a counterfactual that is explained by regression coefficients including interaction terms and significantly improve the fidelity of regression by b-counterfactual value. White and d'Avila Garcez (2019) firstly generates boundary counterfactual explanations which state minimum changes necessary to flip a prediction's classification, then builds local regression models using the boundary counterfactual to measure and improve the fidelity of its regressions.

  A special function of CLEAR is the feature of fidelity measurement, this function based on the concept of b-perturbation and compares each $b$-perturbation with an estimation of that value which named $estimated b$-perturbation, it has been calculated by local regression to output a *counterfactual fidelity error*, $CFE$ which referred to simply as 'fidelity errors' as follow:

  $$CFE = |estimated\ b-perturbation - b-perturbation| \tag{25}$$

  The generation of CLEAR falls into a few steps: (1). Determine $x's$ boundary perturbations. (2). Generate labeled synthetic observations. (3). Create a balanced neighborhood dataset. (4). Perform a step-wise regression on the neighborhood dataset, under the constraint that the regression curve should go through $x$. (5). Estimate the b-perturbations by substituting $x'sb$-counterfactual values from min $f(x)$, other than for feature $f$, into the regression equation and calculating the value off. See the example below. (6). Measure the fidelity of the regression coefficients. (7). Iterate it to the best explanation. (8). Lastly, CLEAR also presents the option of adding $x'sb$-counterfactuals, min $f(x)$, to $x'$s neighborhood dataset.

  The performance in terms of fidelity of CLEAR has been compared with LIME by White and d'Avila Garcez (2019). The result shows that regressions are found to have significantly higher fidelity than LIME with five case studies and produce averaging over 40 percent higher in their research.

- **LIME-C.**: LIME-C is a hybrid solution proposed by Ramon et al. (2020) which distinct in connects additive feature attribution explanations(LIME) with counterfactual. The motivation for this hybrid solution start from an assumption which Ramon et al. (2020) assumes that if the importance-rankings of features are sufficiently accurate, it may be possible to compute counterfactual from them which, therefore, become the reason that the feature which raking on the top was considered to remove until the predicted class.

  Additive feature attribution methods use an explanation model $g$ as an interpretable approximation of the trained classification model $C$ which can be show as a linear model as below Ramon et al. (2020):

  $$g(x') = \phi_0 + \sum_{j=1}^{m} \phi_j x'_j \tag{26}$$

  The equation, $x'_j \in 0,1$ is the binary representation of $xj$ (where $x'_j$ is 1, if $x_j$ is non-zero, else it equals 0), $m$ is the number of features of instance $x$, and $\phi_0, \phi_j \in R$. To generate

this hybrid method, a linear algorithm for finding counterfactuals (lin-SEDC) has been retrieved from Martens and Provost (2014) to applied on the ranked list for counterfactual generation.

Ramon et al. (2020) points out that this method is stable effectiveness for all data and models, and even for very large data instances that require many features to be removed for a predicted class change, LIME-C computes counterfactuals relatively fast.

- **SEDC**

  Martens and Provost (2014) proposed a Search for Explanations for Document Classification (SEDC) which can counterfactually explain predictions of any classification model, this model output minimum-size explanations for linear models by ranks all words appearing in the document regarding the product $\beta_{jX_{ij}}$, where $\beta_j$ is the linear model coefficient. The explanation with the top-ranked words is an explanation of the smallest size.

  Moreover, SEDC is a model-agnostic algorithm for counterfactual which can manage behavioral and textual data sources, it conducts a best-first search with local improvement.

### 5.6.4 Game Theory Centric Approaches

In this section, we summarize the algorithms that we classified as Game Theory Centric using the proposed taxonomy:

- **SHAP-C.** It is an additive feature attribution explanation proposed by Ramon et al. (2020). It has a similar generating function with LIME at the beginning by random sampling to generate counterfactuals. Both algorithm start from mapping the instance to a binary representation $X' = (x'_1, ..., x'_m)$ by a function $h(x) = x'$. After that, instances have been perturbed and generated from $x'$ and all the perturbed instance $z'$ has been assigned a distance weight $\pi_{x'}(z')$. Different from LIME, SHAP uses Shapley values to determine the local feature by generating estimating distance weights for different subject sizes. For each subset size $s$, a distance weight is estimated, then the method samples $\tilde{n}$ perturbed instances from the subset spaces, starting from the smallest (and largest) subsets, then it trained the model by $l1$ regularized linear regression.

  An experiment hosted by Ramon et al. (2020) comparing the model performance between LIME-C, SHAP-C, as a result, SHAP-C presents well with a highly unbalanced data set. However, it is much sensitive to the number of active features compared to LIME-C.

- **SHAP-CC.** Rathi (2019) The notion of contrastive and counterfactual SHAP was created by Rathi (2019) which attempts to generate partial post-hoc contrastive explanations with a corresponding counterfactual. Rathi (2019) used a P-contrastive methodology for generating an explanation that allows the user viewing the change visibly to achieve a certain output.

  The main idea of this explanation is considered a P-contrast question which equivalent to the format *"Why [predicted-class]"* shows instead of *[desired class]?*. This has been conducted into Rathi (2019)'s explanation to estimate its Shapely values for each of the possible target classes. With the Shapely values, a negative Shapely value indicates the features that have negatively contributed to the specific class classification. Moreover, Rathi (2019) generate a "Why P not Q" explanation by break down into "why P?" and "Why not Q", these two segments are the answer for constructed using the Shapley values for class P and class Q and returned. A description to generate the Natural language

explanation in SHAP and the Contrastive datapoint as following commend: (1). **Data**: $dp = Input()$ is the datapoint, (2). $Q = Input()$ is the desired class $\neq$ P. (3). **Result:** Shapley values for a given datapoint generate from the SHAP toolest (4). $p < classifier(dp), Q < Q, SV < SHAP(dp)$ (5). **return** $P, Q, SV$

### 5.6.5 Case-Based Reasoning Approaches

In this section, we summarize the algorithms that we classified as Case-Based Reasoning using the proposed taxonomy:

- **Case based reasoning(CBR) for Good counterfactual**

  Keane and Smyth (2020) argue that the most of techniques for generating XAI used random perturbations which have an issue on sparsity and plausibility. Whereas, case-based reasoning has successfully explained predictions using factual cases with well sparsity and plausibility. One prior condition for using CBR is to ensure the case-base is contained a good counterfactual. Keane and Smyth (2020) consequently proposed a mechanism that considers the feasibility of counterfactual from a CBR perspective.

  The mechanism of generating a good Counterfactual on a case-based approach should firstly be a constraint on finding a minimally-different case which near to the decision boundary within two features. The method for pairing a case and its corresponding good counterfactual as an explanation case or XC has been defined into an equation by Keane and Smyth (2020). The equation is present as symmetric, either of the cases can be viewed as the query or counterfactual and each XC is corresponding to a set of match-features that are the same between the query and counterfactual. Most importantly, the XC could reveal which features should be changed when generating a new counterfactual in the feature space near a query case.

$$xc(c, c') \iff class(c) \neq class(c') \& diffs(c, c') \leq 2 XC(C) = (c, c') : c, c'' \in C \& xc(c, c') \quad (27)$$

### 5.6.6 Summary

Table 2 summarizes all the algorithms that we analysed in this section in therms of underlying theories, algorithms, applications and properties.

## 5.7 Counterfactual Approaches to Explainable AI: Applications

Turning a counterfactual algorithm into an application becomes a target for a recent researcher, our systematic review result shows that the each counterfactual approaches applied its formats

[5] https://github.com/SeldonIO/alibi
[6] https://github.com/SeldonIO/alibi
[7] https://github.com/thibaultlaugel/truce
[8] https://github.com/interpretml/DiCE
[9] https://github.com/amirhk/mace
[10] https://bitbucket.org/ChrisRussell/diverse-coherent-explanations/src/master/
[11] https://github.com/susanne-207/moc
[12] https://github.com/Ighina/CERTIFAI
[13] https://github.com/riccotti/LORE
[14] https://github.com/yramon/LimeCounterfactual
[15] https://github.com/yramon/edc
[16] https://github.com/ClearExplanationsAI
[17] https://github.com/yramon/ShapCounterfactual

| Theory / Approach | Algorithms | Ref. | Applications | Code? | Properties | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Prolixity | Plausibility | Sparsity | Diversity | Feasibility | Optimization | Causal? |
| Instance-Centric | $b$-Counterfactuals | (Wachter et al., 2018) | C [Tab / Img] | Yes[5] [Algo: CF] | ✓ [$L_1$-norm] | ✗ | ✓ | ✗ | ✗ | Gradient Descent | ✗ |
| | Prototype Counterfactuals | (Looveren and Klaise, 2019) | C [Tab / Img] | Yes[6] [Algo: CFProto] | ✓ [$L_1$/$L_2$-norm] | ✓ | ✓ [kd-trees] | ✓ | ✗ | FISTA | ✗ |
| | FACE | (Poyiadzi et al., 2020) | C [Tab / Img] | No | ✗ | ✓ | ✗ | ✓ | ✓ | $\epsilon$ -graphs | ✗ |
| | Weighted Counterfactual | (Grath et al., 2018) | C [Tab] | No | ✓ [$L_1$-norm] | ✗ | ✓ | ✗ | ✗ | Gradient Descent | ✗ |
| | TRUCE | (Laugel et al., 2019, 2018) | C [Tab / Txt / Img] | Yes[7] | ✓ [$L_0$-norm] | ✗ | ✓ | ✗ | ✗ | Growing Spheres | ✗ |
| | DICE | (Mothilal et al., 2020) | C [Tab] | Yes[8] | ✓ [$L_1$-norm] | ✗ | ✓ [hinge loss] | ✓ | ✓ | Gradient Descent | ✗ |
| Constraint-Centric | GIC | (Lash et al., 2017) | C [Tab] | No | ✓ | ✗ | ✓ | ✗ | ✗ | Hill Climbing / Genetic Algorithms | ✗ |
| | MACE | (Karimi et al., 2020) | C [Tab] | Yes[9] | ✓ [$L_0$/$L_1$/$L_i nf$-norm] | ✗ | ✓ [constraint satisfaction] | ✓ | ✓ | SMT | ✗ |
| | Coherent Counterfactuals | (Russell, 2019) | C /R [Tab / Txt] | Yes[10] | ✓ [$L_1$-norm] | ✓ | ✓ [mixed polytopes] | ✓ | ✓ | Gurobi Optimization | ✗ |
| Genetic-Centric | MOCE | (Dandl et al., 2020) | C [Tab] | Yes[11] | ✓ [$L_0$-norm] | ✗ | ✓ [ min feature changes] | ✗ | ✗ | NSGA-II | ✗ |
| | CERTIFAI | (Sharma et al., 2019) | C [Tab / Img ] | Yes[12] | ✓ [$L_1$-norm / SSIM] | ✗ | ✗ | ✓ [mutations] | ✗ | Fitness | ✗ |
| | LORE | (Guidotti et al., 2018) | C [Tab] | Yes[13] | ✓ [$L_2$-norm / Match] | ✗ | ✗ | ✓ [mutations] | ✗ | DecisionTree Model | ✗ |
| Regression-Centric | LIME-C | (Ramon et al., 2020, 2019) | C / R [Tab / Txt / Img] | Yes[14] | ✗ | ✗ | ✗ | ✗ | ✗ | Additive Feature Attribution | ✗ |
| | SED-C | (Martens and Provost, 2014) | C [Txt] | Yes[15] | ✗ [cosine similarity] | ✗ | ✗ | ✗ | ✗ | - | ✗ |
| | CLEAR | (White and d'Avila Garcez, 2019) | C [Tab] | Yes[16] | ✓ [$L_2$-norm] | ✗ | ✓ [ min feature changes] | ✗ | ✗ | Regression | ✗ |
| Game Theory Centric | SHAP-C | (Ramon et al., 2020, 2019) | C / R [Tab / Txt / Img] | Yes[17] | ✗ | ✗ | ✗ | ✗ | ✗ | Shapley Values | ✗ |
| | SHAP-CC | (Rathi, 2019) | C / R [Tab] | No | ✗ | ✗ | ✗ | ✗ | ✗ | Shapley Values | ✗ |
| Case Based Reasoning | CBR for Good Counterfactuals | (Keane and Smyth, 2020) | C [Tab / Txt] | No | ✓ [$L_1$-norm] | ✓ | ✓ [counterfactual potential] | ✓ | ✓ | Nearest Unlikely Neighbour | ✗ |

Table 2: Classification of collected model-agnostic counterfactual algorithms for XAI based on different properties, theoretical backgrounds and applications. In this table, C stands for Classification, R for Regression. In terms of data structures, *tab* refers to tabular data, *img* to image data, and *txt* to text data. Finally, the column *Ref.* stands for reference to the original paper where the algorithm was proposed.

such as numerical, image-based and text-based and model types includes classification and regression, we notice that the most of the counterfactual thesis comply the counterfactual through a table, image recognition is almost an equivalent popular that widely accept in counterfactual conception. Whereas, the SED-Counterfactual only implement the counterfactual notion into a textual operation.

A relatively consummate counterfactual application has appraise for loan credit evaluation which proposed by Grath et al. (2018), this application has capability for giving a user what-if explanation by conducted a HELOC,(Home Equity Line of Credit) credit application dataset for demonstration, it display the correlation between the input variables and showing the explanation through an interface. "What-IF tool" Wexler et al. (2019) is an other open-source application that allows practitioners to probe, visualize, and analyze machine learning systems, with minimal coding, this tool not only help a user shows counterfactual reasoning in both numeric and categorical feature values, but also increase the possibility for user to investigate decision boundaries, and explore how general changes to data points affect prediction. It is observed that the loan credit and what if tool has well-establish interface to display the counterfactual hypothesis with interaction between human and computer, but both of their works are investigating the correlation between the input features instead of joint the causability into the application to promote the user's causal understanding.

Although the demand for providing XAI systems that promote the causability, the literature is very scarse in this aspect. We only found one recent article that proposed an explanation framework (FATE) based on causability Shin (2021). This framework focuses on human interaction, and the authors used the system causability scale proposed by Holzinger et al. (2019) to validate the effectiveness of their system's explanations.

Shin (2021) highlight that causability represents the quality of explanations and emphasize that it is an antecedent role to explainability. Furthermore, they found that properties such as transparency, fairness, and accountability, play a critical role in improving user trust in the

explanations. In general, this framework is a guideline for developing user-centred interface design from the perspective of user interaction and examines the effects of explainability in terms of user trust. This framework does not refer to XAI algorithms underpinned by a theory of causality, neither on how to achieve causability from such mathematical constructs. In the next section, we provide a set of conditions that we find are crucial elements for an XAI system to promote causability. However, FATE causability system is not underpinned by any formal theory of causality, and the causability metrics applied in this work focused on the interaction of the human with the system.

Holzinger et al. (2019) proposed a theoretical framework with a set of guidelines to promote causability in XAI systems in the medical domain. One of the policies put forward is in creating new visualization techniques that can be trainable by medical experts, as the specialists can survey the underlying explanatory factors of the data. Another point is to formalize a structural causal model of human decision-making and delineating features in the model. Holzinger et al. (2019, 2020) argue that a human-AI interface with counterfactual explanations will help achieve causability. An open research opportunity is to extend human-AI explainable interfaces with causability by allowing a domain expert to interact and ask "what-if" questions (counterfactuals). This will enable the user to gain insights into the underlying explanatory factors of the predictions. Holzinger et al. (2020) propose a system causability scale framework as an evaluation tool for causability in XAI systems.

### 5.7.1 The submitted articles

So far, I have involved in two articles that are under review:
**Article1:** Date: 16/07/2020 — Decision in process
Journal: Decision Support Systems —impact factor: 4.721
Title: An Interpretable Probabilistic Approach for Demystifying Black-box Predictive Models
Authors: Catarina Moreira; **Yu-Liang Chou**; Mythreyi Velmurugan; Chun Ouyang; Renuka Sindhgatta; Peter Bruza

**Article2:** Date:20/02/2021
Journal: Information Fusion — Impact factor: 13.669
Title: Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications
Authors: **Yu-Liang Chou**; Catarina Moreira; Peter Bruza; Chun Ouyang; Joaquim Jorge

# 6 Future work

## 6.1 Towards Causability: Opportunities for Research

Our current work is motivated by Holzinger et al. (2019) hypothesis, which states that for a system to provide understandable human explanations, the user needs to achieve a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use (Holzinger, 2020; Holzinger et al., 2020, 2021, 2017; Holzinger, 2018; Xu et al., 2020).

We support this hypothesis, and we analysed how humans could generate causal mental representations in a system. We found that people find it helpful to engage in counterfactual thinking when considering complex causal scenarios. We also found that children as young as three years old can consider counterfactual scenarios to figure out both what has caused a particular outcome and how it could be prevented (Wesberg and Gopnik, 2013). In other words, explanations based on counterfactuals seem to render causal understandings in people. They are also interpretable because they provide the minimum change in a feature that leads to an alternative hypothetical scenario. In his ladder of causality, Pearl defines counterfactuals as the theory of the world that explains why certain actions have specific effects and what happens in the absence of such actions under a structured probabilistic causal model (Pearl, 2019, 2000).

We conducted a thorough systematic literature review guided by the argument that to measure the causability of an XAI system, then the system needs to be underpinned by a probabilistic causal framework such as the one proposed in Pearl (2000). We believe that counterfactual reasoning could provide human causal understandings of explanations (i.e., causability). However, we found no counterfactual generation approach grounded on a formal and structured theory of causality. In other words, current counterfactual explanation generation approaches for XAI are unable to promote causability and they are based on spurious correlations, rather than cause-effect relationships. This inability to disentangle correlation from causation can deliver sub-optimal, erroneous or even biased explanations to decision-makers, as Richens et al. (2020) highlighted in his work about medical decision making. This lack of formal causal approaches in XAI opens a new path and research directions, as well as new challenges, since learning causal relationships from observational data is a very difficult problem (Zhao and Hastie, 2019; Peters, 2019). So, what properties should an XAI system have to promote causability? In the next section, we provide some guidance on how to answer this question, based on the performed systematic literature review.

## 6.2 The Main Characteristics of a Causability System

We conclude our systematic literature review by highlighting what properties should an XAI system have to promote causability. We find that the process of generating explanations that are human-understandable needs to go beyond the minimisation of some loss function as proposed by the majority of the algorithms in the literature. Explainability is a property that implies the generation of human mental representations that can provide some degree of human-understandability of the system and, consequently, allow users to trust it. As Guidotti et al. (2018) stated, explainability is the ability to present interpretations in a meaningful and effective way to a human user. We argue that for a system do be both explainable and promote causability, then it cannot be resumed to a minimisation optimisation problem. Doing so would imply a simplistic and objective explanation process that needs to be necessarily human-centric to achieve human understandability (Confalonieri et al., 2021). We argue that, for a system to promote causability, the following properties should be satisfied:

- **Causality.** The analysis we conducted revealed that current model-agnostic explainable

AI algorithms lack a foundation on a formal theory of causality. Causal explanations are a crucial missing ingredient for opening the black box to render it understandable to human decision-makers since knowing about the cause/effect relationships of variables can promote human understandability. We argue that causal approaches should be emphasised in XAI to promote a higher degree of interpretability to its users and causability.

- **Counterfactual.** Explanations generated by a causability system needs to be counterfactual. Cognitive scientists agree that counterfactual reasoning is a crucial ingredient in learning and a key for explaining adaptive behaviour in a changing environment (Paik et al., 2014). Counterfactual reasoning induces mental representations of an event that happened and representations of some other event alternative to it (Stepin et al., 2021). It follows that for a machine to achieve a certain degree of human intelligence, then explainability systems need to provide counterfactual explanations. Additionally, for a system to achieve causability, the counterfactual explanations need to be underpinned by a formal theory of causality (Holzinger et al., 2019, 2020). Properties to generate good counterfactuals, such as diversity, feasibility, and plausibility, should also be considered to increase the level of human understanding.

- **Human-Centric.** Explanations need to be adapted to the information needs of different users. For instance, in medical decision-making, a doctor is interested in certain aspects of an explanation, while a general user is interested in other types of information. Adapting the information for the type of user is a crucial and challenging point that is currently missing in XAI literature. There is the need to bring the human user back to the optimisation process with human-in-the-loop strategies (Holzinger, 2016; Holzinger et al., 2019) containing contextual knowledge and domain-specific information. This interactive process has the potential to promote causability, since it will allow the user to create mental representations of the counterfactual explanations in a symbiotic process between the human and the counterfactual generation process.

- **Inference.** To promote the system's user understandability, we argue that a causability framework should be equipped with causal inference mechanisms where the user can interact with the system and ask queries to the generated explanations. Queries such as "given that I know my patient has a fever, what changes this information induces in the explanation?". This type of interaction can be highly engaging for the user and promote more transparency in the system. It can enable more human-centric understandability of the system since the user asks questions (performs inference) over variables of interest.

- **Semantic annotations.** One of the major challenges in XAI and a current open research problem is to convert the sub-symbolic information extracted from the black-box into human-understandable explanations. Incorporating semantic contextual knowledge and domain-specific information are crucial ingredients that are currently missing in XAI. We argue that story models and narratives are two important properties that need to be considered to generate human-understandable and human-centric explanations. Story models and narratives can promote higher degrees of believability in the system and consequently achieve causability.

### 6.3 The causal Abstract model

How to endow machine intelligence with capabilities to explain the underlying predictive mechanisms in a way that helps decision-makers understand and scrutinize the machine-

learned decisions. In the following, we plan to propose an extended framework of Bayesian Networks for generating post hoc local interpretations of black-box predictive models. It supports extracting a Bayesian network as an approximation (or an abstraction) of a black-box model for a specific prediction learned from any given input. Note that explanations can be constructed from the graphical representations of the causal abstract model, and we will address the explanation generation component as a direction for future work.

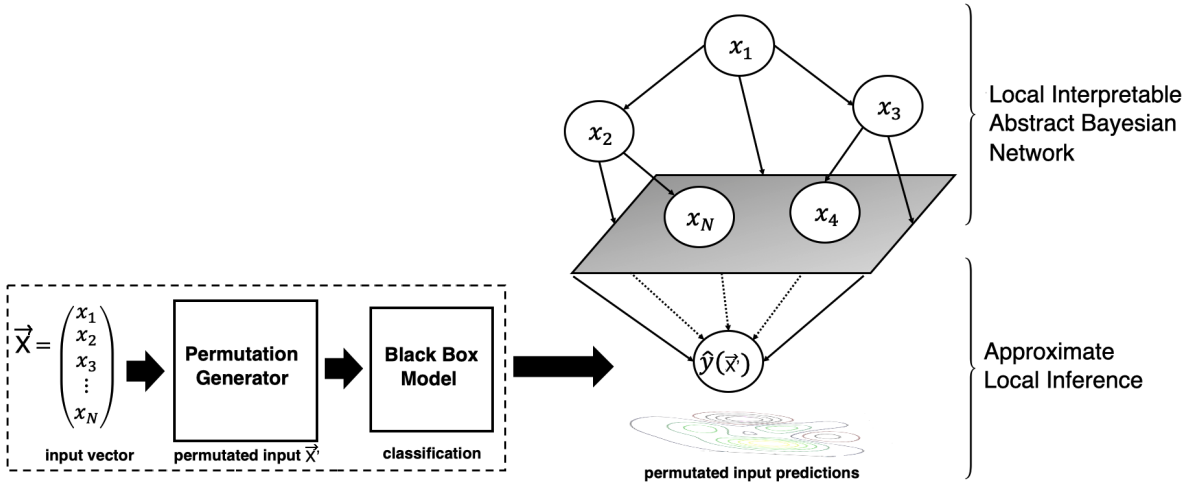


Figure 12: An general illustration of Causal abstract model

The basic idea behind the proposed causal abstract model rests in three main steps: i) permutation generation, ii) Bayesian network learning, and iii) computation of the class variable (representing the result of a prediction). It is important to stress that the proposed model aims to augment a decision-maker's intelligence towards a specific decision problem, providing interpretations that can either reinforce the predictions of the black-box or lead to a complete distrust in these predictions (identification of misclassifications). Figure 12 shows a general illustration of the proposed framework.

# 7 Conclusion

In this report, we emphasize the demand for explainability in an XAI model and identify the research problem for an XAI system. To address the difficulty, we investigated what element can promote the causability that helps people obtaining a causal understanding from a black box prediction. The element derives from formal theories of causality such as causal inference, counterfactuals and probabilistic graphical models. To obtaining a thorough understanding of the field, we define the research aim and a methodology to manage the research project. After that, we segment the research plan into four phases with a project timeline for evaluating the deliverable. Phase I and II have been done by answer how to induce a causal abstract model to promote the XAI system, we probed the explainability of the current representative XAI model(LIME, Anchor, and SHAP). Due to the inability of providing a causal explanation on LIME and SHAP, we conducted a systematic literature review to determine the modern theories underpinning model-agnostic counterfactual algorithms for XAI and analyze if any of the existing algorithms can promote causability.

From the systematic review, we extended the current literature by proposing a new taxonomy for model-agnostic counterfactuals based on six approaches: instance-centric, constraint-centric, genetic- centric, regression-centric, game theory centric, and case-based reasoning centric. Our

research also showed that model-agnostic counterfactuals are not based on a formal and structured theory of causality as proposed by (Pearl, 2000). For that reason, we argue that these systems cannot promote a causal understanding to the user without the risk of the explanations being biased, sub-optimal, or even erroneous. Current systems determine relationships between features through correlation rather than causation.

Finally, this future Ph.D. research will firstly focus on developing a causal abstract model grounded on probabilistic theories of causality and graphical models. Then, generate an explainable counterfactual dashboard that bases on the proposed probabilistic causal grounded framework. And we will consider using the system causability scale proposed by Holzinger et al. (2019) to validate the effectiveness of their system's explanations as we scheduled in the project plan.

## References


Z. C. Lipton, The mythos of model interpretability, Communications ACM 61 (2018) 36–43.

D. Doran, S. Schulz, T. R. Besold, What does explainable ai really mean? a new conceptualization of perspectives, 2017. arXiv:1710.00794.

B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Magazine 38 (2017) 50–57.

C. O'Neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Broadway Books, 2017.

Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453.

A. Lau, E. Coiera, Do people experience cognitive biases while searching for information?, Journal of the American Medical Informatics Association 14 (2007) 599 – 608.

G. Saposnik, D. Redelmeier, C. C. Ruff, P. N. Tobler, Cognitive biases associated with medical decisions: a systematic review, BMC Medical Informatics and Decision Making 16 (2016) 138.

J. Zech, M. Badgeley, M. Liu, A. Costa, J. Titano, E. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, PLOS Medicine 15 (2018) 1-17.

J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, A. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Proceedings of the 30th Conference on Neural Information Processing Systems, 2016.

N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academies of Science of the United States of America 115 (2018) 3635–3644.

A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

M. Kosinski, Y. Wang, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, Journal of Personality and Social Psychology 114 (2018) 246–257.

H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Faithful and customizable explanations of black box models, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2019, pp. 131–138.

R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, arxiv: 1805.10820 (2018).

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, CoRR abs/1806.00069 (2018).

W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences 116 (2019) 22071 – 22080.

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

A. Páez, The pragmatic turn in explainable artificial intelligence (xai), Minds and machines (Dordrecht) 29 (2019) 441–459.

S. Serrano, N. A. Smith, Is attention interpretable?, in: Proc. of the 57th Conference of the Association for Computational Linguistics, ACL, Association for Computational Linguistics, 2019, pp. 2931–2951.

M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD, 2016, pp. 1135–1144.

S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS), 2017.

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

J. G. Richens, C. M. Lee, S. Johri, Improving the accuracy of medical diagnosis with causal machine learning, Nature Communications 11 (2020) 3923–3932.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, Avoiding discrimination through causal reasoning, in: Proceedings of the 31st Conference on Neural Information Processing Systems, 2017.

L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: Machine Learning and Knowledge Extraction, 2020, pp. 1–16.

J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann Publishers, 1988.

R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6276–6282.

B. Lake, T. Ullman, J. Tanenbaum, S. Gershman, Building machines that learn and think like humans, Brain and Behavioural Sciences 40 (2017) e253.

J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2009.

S. Gershman, E. Horvitz, J. Tenenbaum, Computational rationality: A converging paradigm for intelligence in brains, minds, and machines, Science 349 (2015) 273–278.

J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference Foundations and Learning Algorithms, MIT Press, 2017.

A. Holzinger, Explainable ai and multi-modal causability in medicine, i-com 19 (2020) 171 – 179.

D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai, International Journal of Human-Computer Studies 146 (2021) 102551.

M. N. Hoque, K. Mueller, Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making, arxiv: 2101.00633 (2021).

C. T. Ramaravind K. Mothilal, Amit Sharma, Examples are not enough, learn to criticize! criticism for interpretability, in: Proceedings of the 2020 Conference on Fairness, Accountability, and TransparencyJanuary, 2020.

J. Halpern, J. Pearl, Causes and explanations: A structural-model approach. part i: Causes, The British Journal for the Philosophy of Science 56 (2005) 889–911.

S. Psillos, Causation and Explanation, MPG Books Group, 2002.

D. Hume, A Treatise of Human Nature, London: John Noon, 1739.

D. Lewis, Causation, Journal of Philosophy 70 (1973a) 113–126.

D. Lewis, Counterfactuals, Oxford: Blackwell, 1973b.

D. Lewis, Causation, in: Philosophical Papers, Volume II, Oxford University Press, 1986.

S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black-box: Automated decisions and the gdpr, Harvard journal of law & technology 31(2) (2018).

R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, Face: Feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on ai, ethics, and society, 2020, pp. 344–350.

T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1 – 38.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018) 93:1–93:42.

A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, 2020. arXiv:2006.11371.

A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82 – 115.

S. Mohseni, N. Zarei, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, CoRR cs.HC/1811.11839 (2020) 1–45.

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

R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, 2019. arXiv:1812.04608.

D. Alvarez-Melis, T. S. Jaakkola, On the robustness of interpretability methods, 2018. arXiv:1806.08049.

J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, 2020. arXiv:2010.03240.

M. Siering, A. V. Deokar, C. Janze, Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews, Decision Support Systems 107 (2018) 52 – 63.

B. Kim, J. Park, J. Suh, Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information, Decision Support Systems 134 (2020) 113302.

M. A.-M. Radwa Elshawi, Youssef Sherif, S. Sakr, Interpretability in healthcare a comparative study of local machine learning interpretability techniques, in: Proceedings of IEEE Symposium on Computer-Based Medical Systems (CBMS), 2019.

C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Leanpub, 2020.

M. Stiffler, A. Hudler, E. Lee, D. Braines, D. Mott, D. Harborne, An analysis of the reliability of lime with deep learning models, in: Proceedings of the Dstributed Analytics and Information Science International Technology Alliance, 2018.

M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, 2014, pp. 818–833.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, The pragmatic turn in explainable artificial intelligence (xai), Nature communications 10 (2019) 1096–1096.

H. F. Tan, K. Song, M. Udell, Y. Sun, Y. Zhang, Why should you trust my interpretation? understanding uncertainty in lime predictions, 2019.

M. Badhrinarayan, P. Ankit, K. Faruk, Explainable deep-fake detection using visual interpretability methods, in: 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 289–293.

A. Preece, Asking 'why' in ai: Explainability of intelligent systems - perspectives and challenges, International journal of intelligent systems in accounting, finance & management 25 (2018) 63–72.

R. Turner, A model explanation system, in: IEEE 26th International Workshop on Machine Learning for Signal Processing, 2016.

B. Osbert, K. Carolyn, B. Hamsa, Interpretability via model extraction, arxiv: 1705.08504 (2017).

J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, N. Ramamurthy, Karthikeyan, Treeview: Peeking into deep neural networks via feature-space partitioning, Nature communications (2019).

R. Sindhgatta, C. Moreira, C. Ouyang, A. Barros, Interpretable predictive models for business processes, in: Proceedings of the 18th Internation Conference on Business Process Management (BPM), 2020a.

R. Sindhgatta, C. Ouyang, C. Moreira, Exploring interpretability for predictive process analytics, in: Proceedings of the 18th International Conference on Service Oriented Computing (ICSOC), 2020b.

M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the 32nd AAAI International Conference on Artificial Intelligence, 2018.

L. S. Shapley, A value for n-person games, Rand coporation (1952) 15.

E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowledge and Information Systems 41 (2013) 647–665.

M. del Pozo, C. Manuel, E. González-Arangüena, G. Owen, Centrality in directed social networks. a game theoretic approach, Social networks 33 (3) (2011) 191–200.

T. P. Michalak, K. V. Aadithya, S. P. L, B. Ravindran, N. R. Jennings, Efficient computation of the shapley value for game-theoretic network centrality, The Journal of artificial intelligence research 46 (2013) 607–650.

E. Livshits, L. Bertossi, B. Kimelfeld, M. Sebag, The shapley value of tuples in query answering, arxiv: 1904.08679 (2019).

A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, A. (Kouros)Mohammadian, Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis, Accident Analysis & Prevention 136 (2020) 105405.

B. Peer, J. L. J, von Mering Christian, R. A. K, L. Insuk, E. M. Marcotte, Protein interaction networks from yeast to human, The Journal of artificial intelligence research 14 (2004) 292–299.

M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), arxiv: 2005.13997 (2020).

J. W. Hilde, W. van Ipenburg, M. Pechenizkiy, A human-grounded, evaluation of shap for alert processing, arxiv: 1907.03324 (2019).

W. H. JP, W. van Ipenburg, M. Pechenizkiy, A human-grounded evaluation of shap for alert processing, CoRR: arxiv: 1907.03324 (2019).

H. Y. Teh, A. W. Kempa-Liehr, K. I.-K. Wang, Sensor data quality: a systematic review, Journal of Big Data 7 (2020) 11.

A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (scs), KI - Künstliche Intelligenz 34 (2020) 193–198.

R. Byrne, Cognitive processes in counterfactual thinking about what might have been, The psychology of learning and motivation: Advances in research and theory 37 (1997) 105–154.

I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

D. Wesberg, A. Gopnik, Pretense, counterfactuals, and bayesian causal models: Why what is not real really matters, Cognitive Science 37 (2013) 1368–1381.

L. M. Pereira, A. B. Lopes, Cognitive prerequisites: The special case of counterfactual reasoning, Machine Ethics. Studies in Applied Philosophy, Epistemology and Rational Ethics 53 (2020).

J. Paik, Y. Zhang, P. Pirolli, Counterfactual reasoning as a key for explaining adaptive behavior in a changing environment, Biologically Inspired Cognitive Architectures 10 (2014) 24–29.

M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, J. Bian, Causal inference and counterfactual prediction in machine learning for actionable healthcare, Nature Machine Intelligence 2 (2020) 369–375.

J. Pearl, The seven tools of causal inference, with reflections on machine learning, Communications of ACM 62 (2019) 7.

K. Sokol, P. Flach, Explainability fact sheets: a framework for systematic assessment of explainable approaches, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?, IEEE computational intelligence magazine 14(1) (2019).

S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, Lecture Notes in Computer Science (2020) 448–469.

A.-H. Karimi, G. Barthe, B. Balle, I. Valera, Model-agnostic counterfactual explanations for consequential decisions, in: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), 2020, pp. 895–905.

S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, arxiv: 2010.10596 (2020).

D. Martens, F. Provost, Explaining data-driven document classifications, MIS quarterly 38(1) (2014).

M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), arxiv: 2005.13997 (2020).

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

C. Russell, Efficient search for diverse coherent explanations, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 20–28.

P. Domingos, The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World, Penguin, 2017.

A. V. Looveren, J. Klaise, Interpretable counterfactual explanations guided by prototypes, 2019. arXiv:1907.02584.

R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, F. Lecue, Interpretable credit application predictions with counterfactual explanations, in: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NIPS), 2018.

T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, Inverse classification for comparison-based interpretability in machine learning, arxiv: 1712.08443 (2017).

M. T. Lash, Q. Lin, N. Street, J. G. Robinson, J. Ohlmann, Generalized inverse classification, Proceedings of the 2017 Society for Industrial and Applied Mathematics International Conference on Data Mining (2017) 162–170.

S. Sharma, J. Henderson, J. Ghosh, Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, arxiv: 1905.07857 (2019).

A. White, A. d'Avila Garcez, Measurable counterfactual local explanations for any classifier, arxiv: 1908.03020 (2019).

Y. Ramon, D. Martens, F. Provost, T. Evgeniou, A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c, Advances in Data Analysis and Classification 1(1) (2020).

S. Rathi, Generating counterfactual and contrastive explanations using shap, 2019. arXiv:1906.09293.

M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: Case-Based Reasoning Research and Development, Springer International Publishing, 2020.

A. V. Looveren, J. Klaise, Interpretable counterfactual explanations guided by prototypes, arxiv: 1907.02584 (2019).

T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, Comparison-based inverse classification for interpretability in machine learning, in: Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations, 2018, pp. 100–111.

Y. Ramon, D. Martens, F. Provost, T. Evgeniou, Counterfactual explanation algorithms for behavioral and textual data, 2019. arXiv:1912.01819.

J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Transactions on Visualization and Computer Graphics (2019) 1–1.

A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Information Fusion 71 (2021) 28–37.

A. Holzinger, C. Biemann, C. Pattichis, D. Kell, What do we need to build explainable ai systems for the medical domain?, 2017. arXiv:1712.09923.

A. Holzinger, From machine learning to explainable ai, in: Proceedings of the 2018 World Symposium on Digital Intelligence for Systems and Machines, 2018.

G. Xu, T. D. Duong, Q. Li, S. Liu, X. Wang, Causality learning: A new perspective for interpretable machine learning, 2020. arXiv:2006.16789.

J. Pearl, Causality: Models, Representation and Inference, Cambridge University Press, 2000.

Q. Zhao, T. Hastie, Causal interpretations of black-box models, Journal of Business & Economic Statistics (2019) 1–10.

O. Peters, The ergodicity problem in economics, Nature Physics 15 (2019) 1216–1221.

R. Confalonieri, L. Coba, B. Wagner, T. R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Mining and Knowledge Discovery 11 (2021) e1391.

A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop?, Brain Informatics 3 (2016) 119–131.

A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G. C. Crişan, C.-M. Pintea, V. Palade, Interactive machine learning: experimental evidence for the human in the algorithmic loop, Applied Intelligence 49 (2019) 2401–2414.

Y. Wu, Y. Li, Y. Xu, Dual pattern-enhanced representations model for query-focused multi-document summarisation, Knowledge-Based Systems 163 (2019) 736–748.

R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, 2019. arXiv:1812.04608.

A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Information Fusion 71 (2021) 28–37.

J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating xai: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404.

M. N. Hoque, K. Mueller, Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making, 2021. arXiv:2101.00633.

S. Völkel, C. Schneegass, M. Eiband, D. Buschek What is "Intelligent" in Intelligent User Interfaces? A Meta-Analysis of 25 Years of IUI, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020