Information Fusion 81 (2022) 59–83

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications

Yu-Liang Chou [a], Catarina Moreira [a,b,*], Peter Bruza [a], Chun Ouyang [a], Joaquim Jorge [b]

[a] School of Information Systems, Queensland University of Technology, Brisbane, Australia
[b] INESC-ID Lisboa, Instituto Superior Técnico, ULisboa, Portugal

ARTICLE INFO

Keywords:
Deep learning
Explainable AI
Causability
Counterfactuals
Causality

ABSTRACT

Deep learning models have achieved high performance across different domains, such as medical decision-making, autonomous vehicles, decision support systems, among many others. However, despite this success, the inner mechanisms of these models are opaque because their internal representations are too complex for a human to understand. This opacity makes it hard to understand the *how* or the *why* of the predictions of deep learning models.

There has been a growing interest in model-agnostic methods that make deep learning models more transparent and explainable to humans. Some researchers recently argued that for a machine to achieve human-level explainability, this machine needs to provide human causally understandable explanations, also known as *causability*. A specific class of algorithms that have the potential to provide causability are *counterfactuals*.

This paper presents an in-depth systematic review of the diverse existing literature on counterfactuals and causability for explainable artificial intelligence (AI). We performed a Latent Dirichlet topic modelling analysis (LDA) under a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to find the most relevant literature articles. This analysis yielded a novel taxonomy that considers the grounding theories of the surveyed algorithms, together with their underlying properties and applications to real-world data.

Our research suggests that current model-agnostic counterfactual algorithms for explainable AI are not grounded on a causal theoretical formalism and, consequently, cannot promote causability to a human decision-maker. Furthermore, our findings suggest that the explanations derived from popular algorithms in the literature provide spurious correlations rather than cause/effects relationships, leading to sub-optimal, erroneous, or even biased explanations. Thus, this paper also advances the literature with new directions and challenges on promoting causability in model-agnostic approaches for explainable AI.

1. Introduction

Artificial intelligence, in particular, deep learning, has made great strides in equalling and even surpassing human performance in many tasks such as categorisation, recommendation, game playing, or even in medical decision-making [1]. Despite this success, the internal mechanisms of these technologies are an enigma because humans cannot scrutinise how these intelligent systems do what they do. This is known as the *black-box* problem [2]. Consequently, humans rely on blindly accepting the answers produced by machine intelligence without understanding how that outcome came to be. There is growing disquiet 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, among many other critical settings.

1.1. The need for explainability

Several factors motivated the rise of approaches that attempt to turn predictive black-boxes transparent to the decision-maker [3,4]. One of these factors is the recent European General Data Protection Regulation (GDPR) [5], which made the audit and verifiability of decisions from intelligent autonomous systems mandatory, increasing the demand for the ability to question and understand Machine Learning (ML) systems. These regulations directly impact worldwide businesses because GDPR applies not only to data being used by European bodies but also to European data being used by other organisations. Another critical factor is concerned with discrimination (such as gender and racial bias) [6]. Studies suggest that predictive algorithms widely used in healthcare, for

* Corresponding author at: School of Information Systems, Queensland University of Technology, Brisbane, Australia.
E-mail address: catarina.pintomoreira@qut.edu.au (C. Moreira).

https://doi.org/10.1016/j.inffus.2021.11.003
Received 19 February 2021; Received in revised form 20 September 2021; Accepted 6 November 2021
Available online 13 November 2021

instance, exhibited racial biases that prevented minority societal groups from receiving extra care [7] or display cognitive biases associated with medical decisions [8,9]. In medical X-ray images, it was found that deep learning models have learned to detect a metal token that technicians use to visualise the X-ray images, making this feature impacting the predictions of the algorithm [10]. Other studies revealed gender and racial biases in automated facial analysis algorithms made available by commercial companies [11], gender biases in textual predictive models [12–14] or even more discriminatory topics such as facial features according to sexual orientation [15].

The black-box problem and the need for interpretability have motivated an extensive novel body of literature in machine learning. Those publications focus on developing new algorithms and approaches that can not only interpret the complex internal mechanisms of machine learning predictions but also explain and make the decision-maker understand the why of these predictions [16,17]. In this sense, interpretability and explainability have become the main driving pillars of explainable AI (XAI) [18]. More precisely, we define interpretability and explainability in the following way.

- **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 [19].
- **Explainability**, on the other hand, refers to the ability to translate this sub-symbolic information in a comprehensible manner through human-understandable language expressions [20].

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 [21] extends this goal by adding that explainable algorithms should offer a pragmatic and naturalistic account of understanding in AI predictive models. Furthermore, explanatory strategies should offer well-defined goals when providing explanations to its stakeholders.

Currently, there is an extensive body of literature reviewing different aspects of XAI. In Miller [22], the author portraits the missing link between the current research on explanations from the fields of philosophy, psychology, and cognitive science. Miller [22] highlighted three main aspects that an XAI system must have in order to achieve explainability: (1) people seek explanations of the form *why some event happened, instead of another?*, which suggests a need for contrastive and counterfactual explanations; (2) recommendations can focus on a selective number of causes (not all of them), which suggests the need for causality in XAI; and (3) explanations should consist in conversations and interactions with a user promoting an explanation process where the user engages in and learns from the explanations. In Guidotti et al. [23], the authors survey black-box specific methods for explainability and propose a taxonomy for XAI systems based on four features: (1) the type of problem; (2) the type of explainer adopted; (3) the type of black-box model that the explainer can process; and (4) the type of data that the black-box supports. On the other hand, Das and Rad [24] proposed a taxonomy for categorising XAI techniques based on their scope of explanations, the methodology behind the algorithms, and explanation level or usage. Adadi and Berrada [25] classified explainable methods according to (1) the interpretability of the model to be explained, (2) the scope of the interpretability, and (3) if the black-box is dependent or not on any machine learning model.

Barredo Arrieta et al. [26] proposed and discussed a taxonomy related to the explainability of different machine learning and deep learning models. Additionally, the authors proposed new methodologies towards responsible artificial intelligence and discuss several aspects regarding fairness, explainability, and accountability in real-world organisations.

Some authors have reviewed evaluation methodologies for explainable systems and proposed a novel categorisation of XAI design goals and evaluation measures according to stakeholders [27]. In contrast, other authors identified objectives that evaluation metrics should achieve and demonstrated the subjectivity of evaluation measures regarding human-centred XAI systems [28]. Carvalho et al. [29] extensively surveyed the XAI literature with a focus on both qualitative and quantitative evaluation metrics as well as properties/axioms that explanations should have. For other works that survey evaluation in XAI, we refer the reader to Hoffman et al. [30], and Alvarez-Melis and Jaakkola [31].

In terms of the generation of explanations, Chen et al. [32] surveyed the literature in terms of biases. The authors identified seven different types of biases that were found in recommendations. They proposed a taxonomy on recommender systems and potential ways to *de-bias* them.

1.2. From explainability to the need of causability

For a model to be interpretable, it must offer explanations that make sense to the decision-maker and ensure that they accurately represent the actual reasons for the model's decisions [33]. Current XAI approaches that attempt to decipher a black-box that is already trained (also known as post-hoc, model-agnostic techniques) build models around local interpretations, providing approximations to the predictive black-box [34,35], instead of reflecting the true underlying mechanisms of the black-box (as pointed out by [36]). In other words, these algorithms compute correlations between individual features to approximate to the black-box predictions. In this paper, we argue that the inability to disentangle *correlation from causation* can deliver sub-optimal or even erroneous explanations to decision-makers [37]. Causal approaches should be emphasised in XAI to promote a higher degree of interpretability to its users and avoid biases in predictive black-boxes [38].

Finding causal relationships between features and predictions in observational data is a challenging problem and constitutes a fundamental step towards explaining predictions [20]. Causation is an ubiquitous notion in Humans' conception of their environment [39]. Indeed, 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 [40]. There has been a growing emphasis that AI systems should be able to build causal models of the world that support both explanation and understanding, rather than merely to solve pattern recognition problems [41]. For decision-support systems, whether in finance, law, or even warfare, understanding the causality of learned representations is a crucial missing link in XAI [42–44]. However, when considering machines, how can we make computer-generated explanations that are causally understandable by humans? This notion was recently put forward by Holzinger et al. [20] in a term coined *causability*. **Causability** corresponds to the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use [20]. In this sense, causability can be seen as a property of *human intelligence*, whereas explainability is a property of *artificial intelligence* [45]. Fig. 1 illustrates the notion of *causability* under the context of XAI.

1.3. 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 [47]. Causal concepts are central to practical deliberations, health diagnosis, and similar types of reasoning. [20,48]. Even when one attempts to describe certain phenomena, the evidence produced must acknowledge, to a certain degree, the causes of the effects being explained [49]. However, causality, its nature, and its definition have promoted much disagreement throughout the centuries in Philosophical literature. Bertrand Russel famously denied causality, arguing that it constituted an incoherent topic [50]. The philosopher

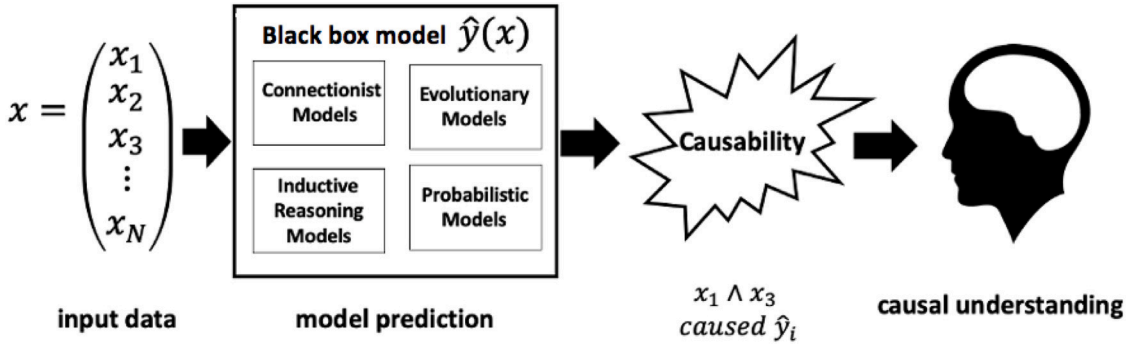

Fig. 1. The general notion of causability. Given a predictive black-box model, the goal is to create interpretable and explainable methods that will provide the user a causal understanding of why certain features contributed to a specific prediction [20,45,46].

and empiricist David Hume formalised causation in terms of *sufficient* and *necessary conditions*: an event c causes an event e if and only if there are event-types C and E such that C is necessary and sufficient for E [50]. 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 (the effect) such that, had the first object (the cause) not occurred, then the second (the effect) would never exist [51]. This concept started to gain more importance in the literature with the works of Lewis [52].

Counterfactuals are defined as conditional assertions 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 specific predictive outcome [53]. While most XAI approaches tend to focus on answering *why* a black-box predicted a particular outcome, counterfactuals attempt to answer this question in another way by helping the user understand *what features would need to change to achieve a desired outcome* [54]. For instance, suppose 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 can take the form of an alternative hypothetical scenario, such as *if your income were higher than* $15.000, *you would have been granted a loan* [48].

1.4. Contributions

The hypothesis that we put forward in this paper is that the inability to disentangle *correlation from causation* can deliver sub-optimal or even erroneous explanations to decision-makers [37]. Therefore, causal approaches should be emphasised in XAI to promote a higher degree of interpretability to people. 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 in XAI, it is the purpose of this paper to make a systematic review and critical discussion of the diverse existing body of literature on these topics. Recently, three papers surveying counterfactuals in XAI were proposed in the literature [55–57]. Our paper departs from the current body of literature by analysing current counterfactual model-agnostic approaches and how they could promote causability.

In summary, this paper contributes with a literature review with discussions under three paradigms:

- **Theory.** We survey and formalise the most important conceptual approaches that ground current explainable AI models in the literature based on counterfactual theories for causality.
- **Algorithms.** We study the main algorithms proposed in the XAI literature that use counterfactuals and discuss which ones have are based on probabilistic approaches to causality and which ones can achieve a certain degree of causability.
- **Applications.** We apply a continuous use case analysis to understand the main domains and fields where XAI algorithms that promote causability are emerging; and study the potential advantages and disadvantages of such approaches to real-world problems, namely in the mitigation of biased predictions.

1.5. Paper organisation

The remainder of this paper is organised as follows. In Section 2, we present the taxonomy of the current state-of-the-art algorithms in XAI. Then, in Section 3 we present a systematic review on counterfactual and causability approaches in XAI. In the following sections, we present the findings of our systematic literature review. First, in Section 4, we present properties that are used throughout the literature to assess what are good counterfactuals and a discussion on the impacts of different distance functions on XAI algorithms. Then, Section 5, we present our first contribution where we analyse the theories that underpin the different algorithms of the literature by introducing a novel taxonomy for model-agnostic counterfactuals in XAI. Next, in Section 6, we analyse the different algorithms of the literature based on the taxonomy that we proposed. In Section 7, we discuss the main applications of XAI algorithms together with recent developments in causability. Next, in Section 8, we present the main characteristics that should be part of a causability system for XAI. Finally, Section 9 contains answers to the proposed research questions, and Section 10 presents the main conclusions of the work.

2. A general overview of current model-agnostic approaches in XAI

Various approaches have been proposed in the literature to address the problem of interpretability in machine learning. Generally, this problem can be classified into two primary categories: interpretable models (inherently transparent) and model-agnostic approaches, also referred to as post-hoc models (which aim to extract explanations out of any opaque model). From our systematic literature review, these approaches can be categorised within the taxonomy presented in Fig. 2 Belle and Papantonis [58].

Interpretable models are by design already interpretable. They provide the decision-maker with a transparent white-box approach for prediction [59]. Decision trees, logistic regressions, and linear regressions are examples of interpretable models [60]. Model-agnostic approaches, on the other hand, refer to the derivation of explanations from a black-box predictor by extracting information about the system's underlying mechanisms [61].

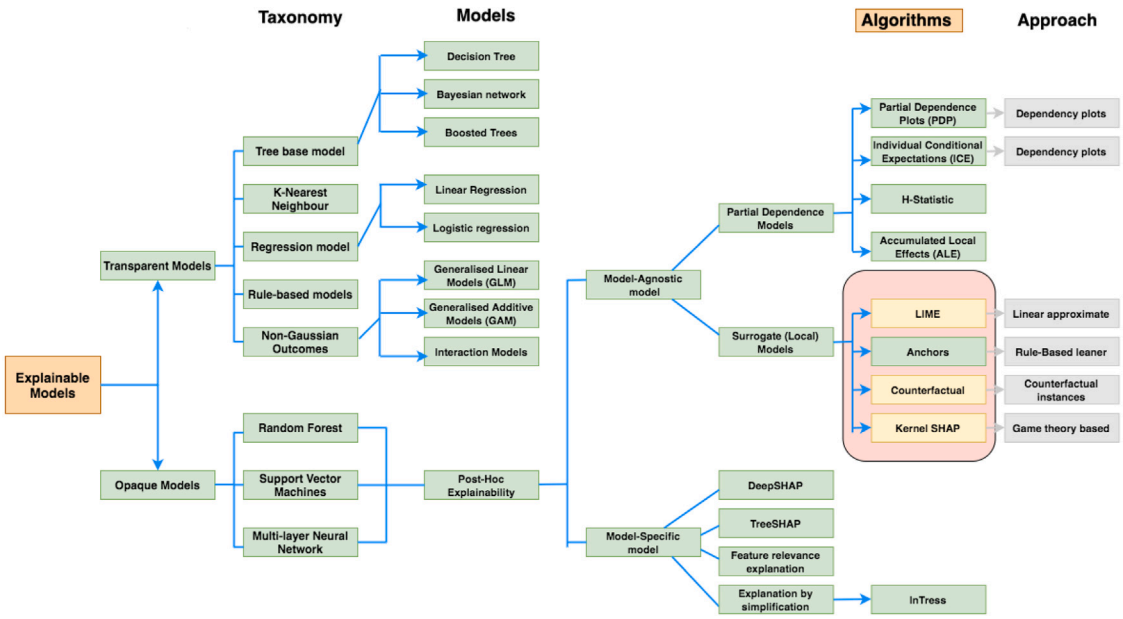

Fig. 2. Taxonomy of explainable artificial intelligence based on the taxonomy proposed by Belle and Papantonis [58]..

Model-agnostic methods (also post-hoc) are divided into two major approaches: partial dependency plots and surrogate models. Partial dependency plots can only provide pairwise interpretability by computing the marginal effect that one or two features have on the prediction. On the other hand, surrogate models consist of training a new local model that approximates the predictions of a black-box. Model-agnostic post-hoc methods have the flexibility of being applied to any predictive model compared to model-specific post-hoc approaches. The two most widely cited post-hoc models in the literature include LIME [34] and Kernel SHAP [35]. In XAI literature, counterfactual explanations are generated using a post-hoc approach, and they can also either be model-agnostic or model-specific. The main focus of this literature review is on model-agnostic post-hoc counterfactual models due to their flexibility and ability to work in any pre-existing trained model. This is detailed in Section 5.

2.1. LIME - Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) [34] can explain the predictions of any classifier. They do so by approximating it with a locally faithful interpretable model. Hence, LIME generates local interpretations by perturbing a sample around the input vector within a neighbourhood of a local decision boundary [34,62]. Each feature is associated with a weight computed using a similarity function. This function measures the distance between the prediction of the original instance and the predictions of the sampled points in the local decision boundary. An interpretable model, such as linear regression or a decision tree, can learn the local importance of each feature. This explanation translates into a mathematical optimisation 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 Eqs. (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 π_x defines how large the neighbourhood 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) [59].

In terms of image data, LIME creates explanations via perturbed instances. These are produced by dividing the input image into interpretable components (contiguous superpixels) and running each perturbed image instance through the model to yield a probability Badhrinarayan et al. [63]. After that, a simple linear model learns this locally-weighted dataset. At the end of the process, LIME presents the superpixels with the highest positive weights as an explanation.

While LIME can find local models with short and selective user-friendly explanations, it is also susceptible to sampling bias. Additionally, the explanations generated slightly change every time the algorithm is executed due to poor stability. LIME's popularity has motivated different approaches, as outlined in the next section.

2.2. Approaches based on LIME

LIME has been extensively applied in the literature. For instance, Stiffler et al. [64] used LIME to generate salience maps of a particular region showing which parts of the image affect how the black-box model reaches a classification for a given test image [65,66]. In addition, Tan et al. [67] 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.

Other researchers proposed extensions to LIME. Turner [68] 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. [69] utilised a surrogate model to extract a decision tree that represents its behaviour. Thiagarajan et al. [70] proposed an approach for building TreeView visualisations using a surrogate model. LIME has also been used to investigate the quality of predictive systems in predictive process analytics [71]. In Sindhgatta et al. [72] the authors found that predictive process mining models suffered from different biases, including data leakage. Additionally, they demonstrated that LIME could be used as a tool to debug black-box models.

Lastly, a rule-based approach extension for LIME is Anchor [73]. Anchor attempts to address some of LIME limitations by maximising the likelihood of how a given feature might contribute to a prediction. Anchor introduces IF-THEN rules as explanations and the notion of coverage, which allows the decision-maker to understand the regions where the generated explanations are valid. Although Anchors can generate rules as explanations, it is computationally intractable and has poor stability (the explanations change significantly each time the algorithm is executed).

2.3. SHAP — SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) is an explanation method that uses Shapley values [74] from coalitional game theory to fairly distribute the gain among players, where contributions of players are unequal [35]. 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 [75].

In SHAP, an explanation model, $g(z')$ is given by a linear combination of Shapley values ϕ_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 [75] claim that in a coalition game, it is usually assumed that n players form a grand coalition that has a specific 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 [35] 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, the player 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.

Compared with LIME, SHAP provides several advantages. The most important one is that it is highly stable, which means that even if the algorithm is executed several times, the generated explanations are always consistent. SHAP is, however, computationally more expensive than LIME. However, some kernels have been developed for specific algorithms, such as Deep SHAP [76] and Tree SHAP [77], that are computationally more efficient.

2.4. Approaches based on SHAP

In terms of related literature, Miller Janny Ariza-Garzón and Segovia-Vargas [78] adopted SHAP values to assess the logistic regression model and several machine learning algorithms for granting scores in P2P (peer-to-peer) lending; the authors point out SHAP values can reflect dispersion, nonlinearity, and structural breaks in the relationships between each feature and the target variable. They concluded that SHAP could provide accurate and transparent results on the credit scoring model. Parsa et al. [79] also highlights that SHAP could bring insightful meanings to interpret prediction outcomes. For instance, one of the techniques in the model, XGBoost, not only is capable of evaluating the global importance of the impacts of features on the output of a model, but it can also extract complex and non-linear joint impacts of local features.

Recently, Wang et al. [80] proposed to generalise the notion of Shapley value axioms for directed acyclic graphs. The new algorithm is called Shapley flow and relies on causal graphs in order to be able to compute the flow of Shapley values that describe the internal mechanisms of the black-box.

The following sections will expand the analysis of model-agnostic approaches for XAI by conducting a systematic literature review on counterfactual and causability approaches for XAI.

3. Systematic literature review towards counterfactuals and causability in XAI

The purpose of this systematic review paper is to investigate the theories, algorithms, and applications that underpin XAI approaches that have the potential to achieve *causability*. This paper will survey the approaches in the extensive body of literature that are primarily based on causality and counterfactuals to help researchers identify knowledge gaps in the area of interest by extracting and analysing the existing approaches.

Our systematic literature review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework as a standardised way of extracting and synthesising information from existing studies concerning 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 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. [81], we conducted a topic modelling 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), systematisation, 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.

3.1. Research questions

To help researchers identify knowledge gaps in the area of causality, causability, and counterfactuals in XAI, we proposed the following research questions:

- 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 understandable causal 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?

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

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

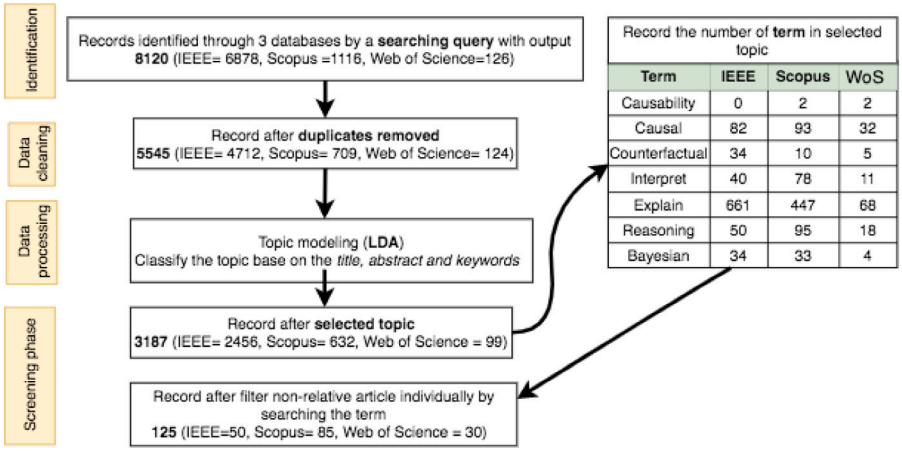

Fig. 3. PRISMA flow diagram search results.

3.2. 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 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 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 (1116), 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 summarised in the PRISMA flow diagram illustrated in Fig. 3.

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.

3.3. 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 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 process guaranteed that the papers classified within a topic contained all the relevant keywords to address our research questions.

This paper 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 Fig. 4 with the inter-topic distance showing the marginal topic distributions (left) and the top 10 most relevant terms for each topic. Analysing Fig. 4, Topic 1 contained all the words that are of interest to 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 related 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 (2456), 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 (50), Scopus (85), WoS (30).

3.4. Word co-occurrence analysis

Our survey focuses on understanding the necessary and sufficient conditions to achieve causability and how current approaches can promote it. We analysed 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 using three different keywords of interest: explainable AI, counterfactuals, and causality.

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

Fig. 5 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 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. Fig. 5 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

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

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

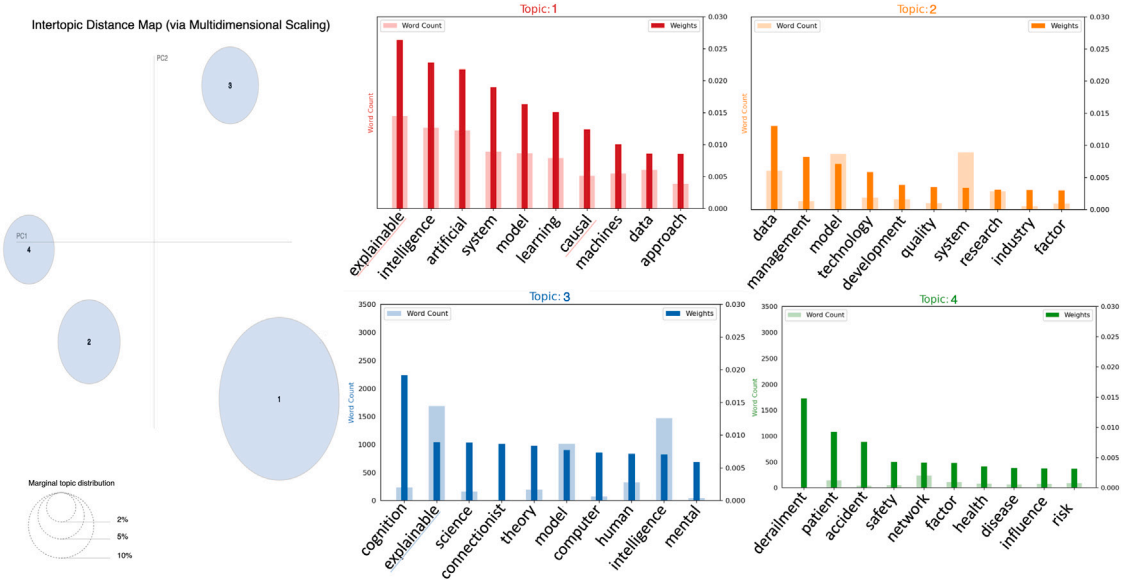

Fig. 4. Best performing LDA topic model for Scopus database, using 709 titles, abstracts, and authors keywords found from the proposed search query. The figure also shows the top 10 most relevant words for each Topic.

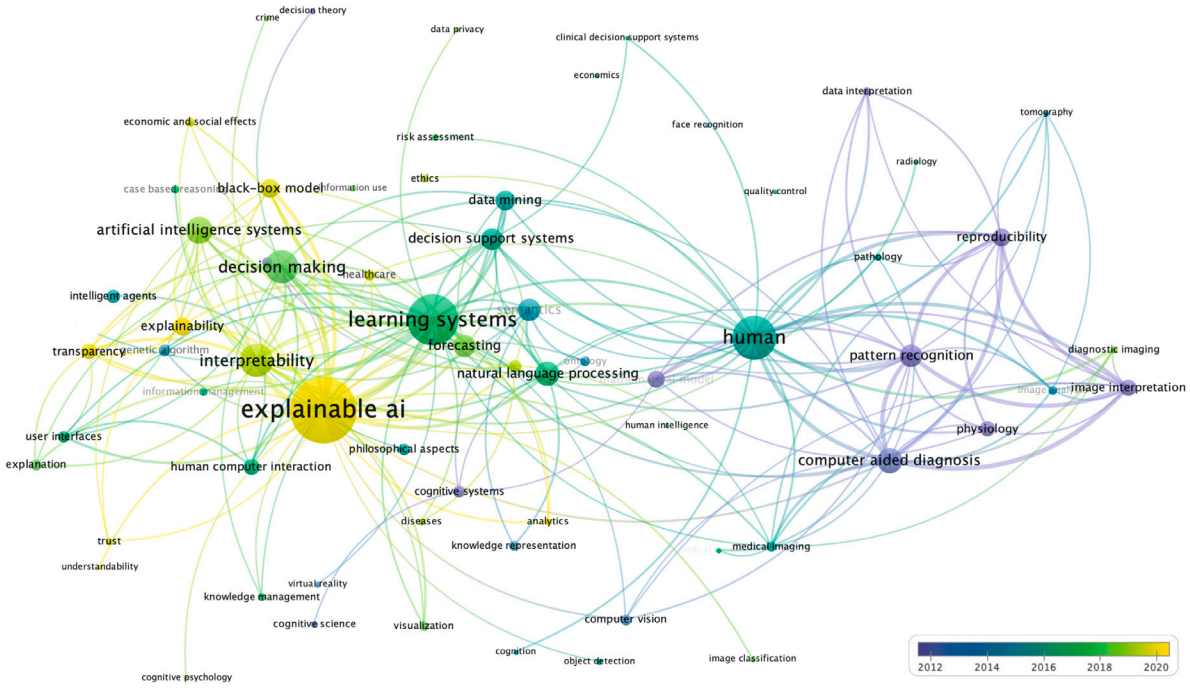

Fig. 5. Network visualisation of co-occurrence between keywords in articles about XAI.

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 [20,82]. Recently, Miller [22] stated that explanations need to be counterfactuals ("contrary-to-fact") [83], since they enable mental representations of an event that happened and also representations of some other event alternative to it [56]. 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 [84]. 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* [85]. Counterfactual reasoning is also critical for explaining adaptive behaviour in a changing environment [86]. We investigated the word co-occurrence in articles involving *explainable AI* and *counterfactuals* to understand how the literature is progressing in this area. Fig. 6 shows the obtained results.

In the density plot in Fig. 6, 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

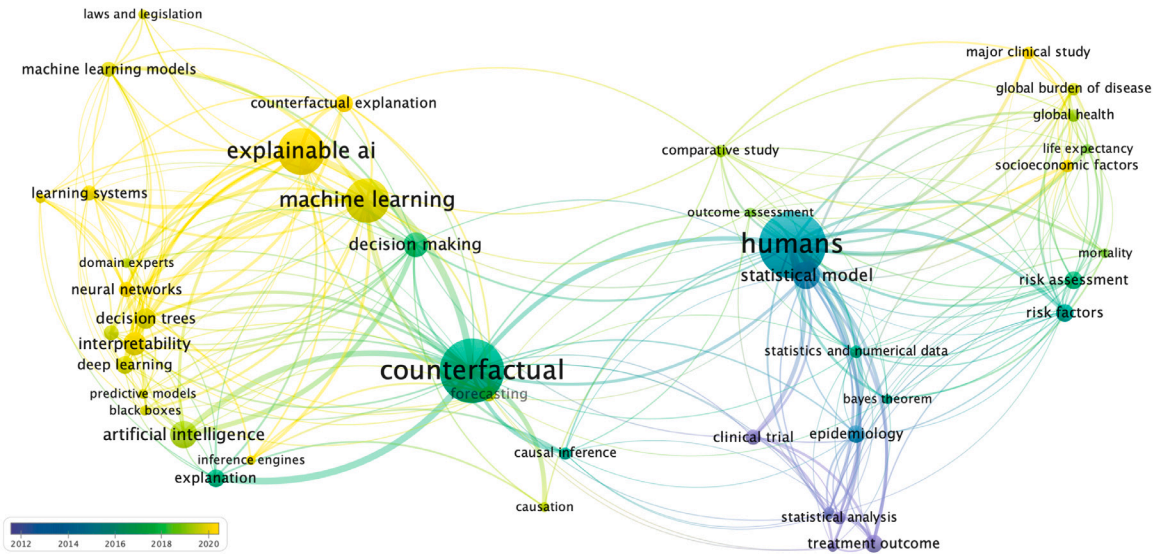

Fig. 6. Network visualisation of co-occurrence between keywords in articles about counterfactuals in XAI.

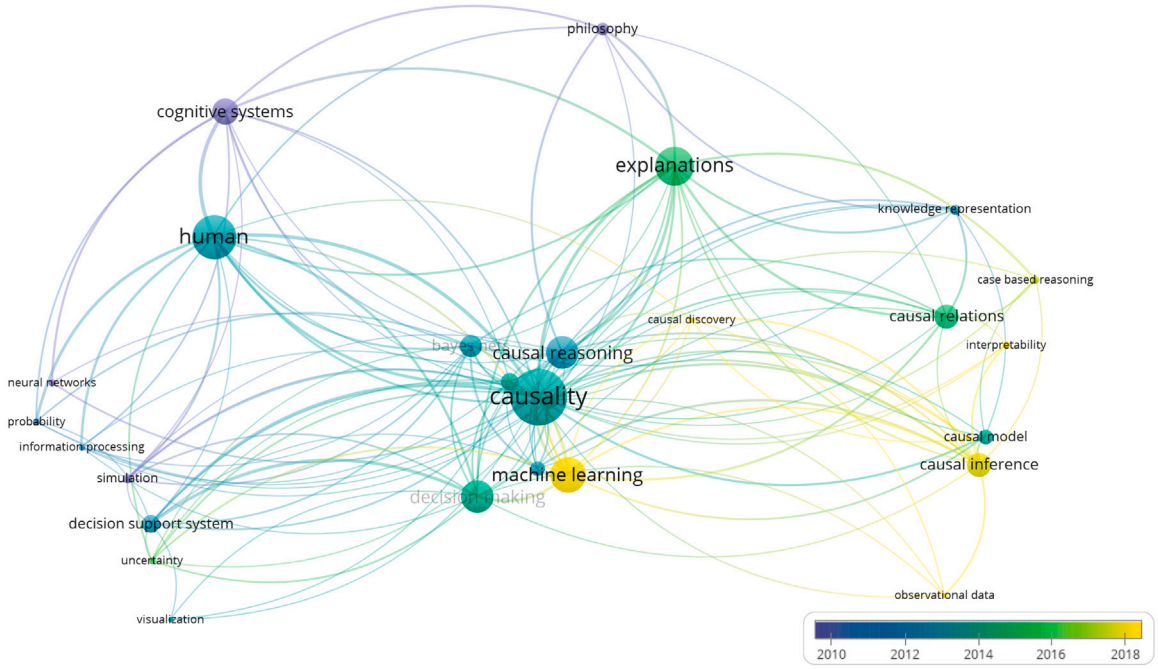

Fig. 7. Network visualisation of co-occurrence between keywords in articles about causality in XAI.

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 [87].

Some researchers also argued that for a machine to achieve a certain degree of human intelligence, causality must be incorporated in the system [88]. 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 [20]. In this sense, we also analysed co-occurrence between keywords in articles about causality in XAI. Fig. 7 illustrates the results obtained.

In terms of causality, in Fig. 7, 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 Fig. 7 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.

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

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

3.5. 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.

3.6. Risk of bias

As with any human-driven task, the process of finding relevant research is affected by cognitive biases. This systematic review acknowledges that limiting our search to three databases (Scopes, Web of Science, and IEEE) might have contributed to 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 extensive, and we found that doing this would exponentially increase the complexity of the LDA topic analysis that we conducted. Finally, the search query was restricted to keywords relevant to collecting the papers of interest. These keywords, however, might have limited our search, and we might have missed relevant articles.

4. Counterfactual approaches explainable AI: Properties for good counterfactuals

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. 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 [20]. Other researchers state that counterfactuals are essential to elaborate predictions at the instance-level [89] and to make decisions actionable [90]. Other researchers claim that counterfactuals can satisfy GDPR's legal requirements for explainability [53].

Most XAI algorithms attempt to achieve explainability by (1) perturbing a single data instance, generating a set of perturbed data points around the decision boundary, (2) passing these perturbed instances through the black-box, generating labels to these data points, and by (3) fitting an interpretable model (such as linear regression or a decision tree) to the perturbed data points [34]. Counterfactuals are classified as *example-based approaches* for XAI [59]. They are based on approaches that compute which changes should be made to the instance datapoint to flip its prediction to the desired outcome [53]. Fig. 8 shows an illustration of several counterfactual candidates for a data instance x according to different works in the literature [54].

4.1. 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 [91]. 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 the initial point to the generated candidate data point.

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

One important question that derives from Eq. (4) is *what kind of distance function should be used?* Different works in the literature address this optimisation 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:

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

Eq. (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 use 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_∞-norm. Fig. 9 shows a graphical representation of the different norms and the respective contours.

- **L_0-norm**. The L_0-norm has been explored in the context of counterfactuals in XAI primarily by Dandl et al. [92] and Karimi et al. [93]. 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 [93]. Fig. 9 shows a visualisation of the L_0-norm where one can see that the function is entirely undifferentiable, making it very hard to find efficient solutions to minimise 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. [53] 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 Fig. 9, 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 also been highly explored in the literature of counterfactuals in XAI, although it does not provide sparse solutions when compared with the L_1 or L_0-norm. Given a vector x, the L_2-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. Fig. 9 shows that smoothness and rotational invariance (a circle or a hyper-sphere in higher dimensions) are both desirable properties in many optimisation problems, making it computationally efficient.

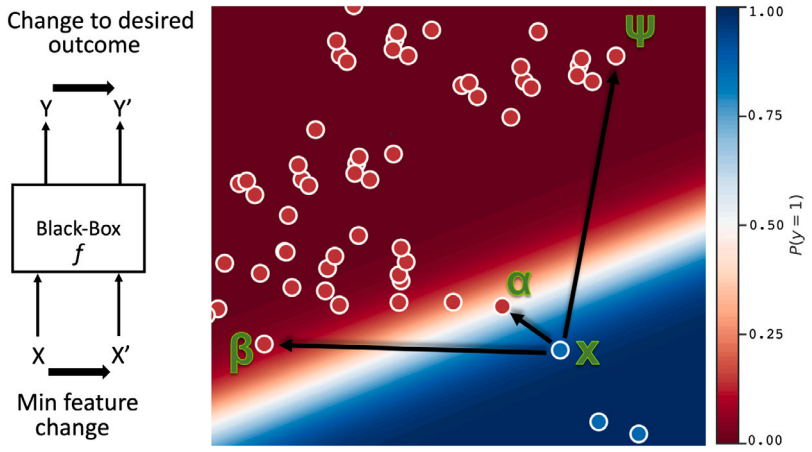

Fig. 8. The different counterfactual candidates for a data instance x. According to many researchers, counterfactual α is the best candidate because it has the smallest Euclidean distance to x [53]. Other researchers argue that counterfactual instance γ is the best choice since it provides a feasible path from x to γ [54]. Counterfactual β is another candidate of poor quality because it rests in a less defined region of the decision boundary.

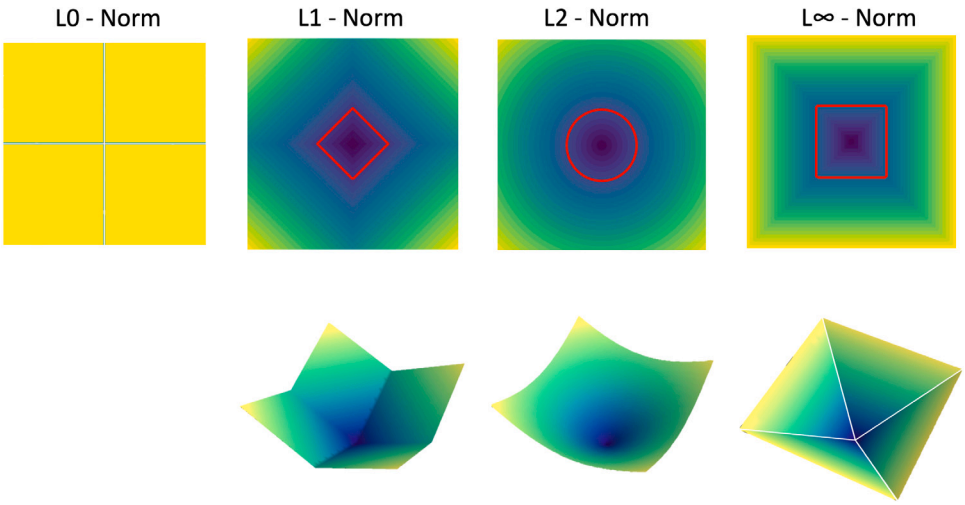

Fig. 9. Graphical visualisation of different L_p-norms: L_0-norms (which is not a norm by definition), the L_1-norm (also known as Manhattan distance), the L_2-norm (known as the Euclidean distance, and the L_∞-norm).

- **L_∞-norm**. The L_∞-norm has been explored in the context of counterfactuals in XAI primarily by Karimi et al. [93]. Given a vector x, it is defined as

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

  Intuitively, L_∞-norm is used to restrict maximum change across features, in other words the maximum change across the features between the initial instance x and the counterfactual candidate x' [93]. Computationally, the L_∞-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 Fig. 9. By minimising the L_∞ norm, we are penalising the cost of the largest feature, leading to less sparse solutions 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. The following section will present the main properties that a theory for counterfactuals in XAI should satisfy.

4.2. Properties to generate good counterfactuals

The literature suggests a set of properties that need to be satisfied in order to generate a good (interpretable) counterfactual:

- **Proximity.** This property consists in functions that calculate the distance of a counterfactual from the input data point while generating a counterfactual explanation, [55]. As mentioned in Section 4.1, many different distance functions can be used to measure proximity, resulting in counterfactual candidates with different properties. Other works in the literature consider other types of proximity measures such as Nearest Neighbour Search [94] or cosine similarity [95].
- **Plausibility.** This property is similar to the terms *Actionability* and *Reasonability* referred in [55,96]. It emphasises 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 number features that need to be changed to obtain a counterfactual [96].
  In cognitive science, counterfactuals are used as a process of imagining a hypothetical scenario contrary to an event that happened and reasoning about its consequences [85]. It is desired that counterfactuals are sparse, i.e., with the fewest possible changes in their features. This property leads to more effective, human-understandable, and interpretable counterfactuals. In Mothilal et al. [48], 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. [55] 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. [53] 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.
  Recently, Pawelczyk et al. [97] proposed a theoretical framework that challenges the notion that counterfactual recommendations should be sparse. The authors argue that the problem of predictive multiplicity can result in situations where there is no superior solution to a prediction problem for a measure of interest (e.g., error rate).
- **Diversity.** This property was introduced in the work of Russell [98] and also explored in Mothilal et al. [48], Karimi et al. [57]. Finding the closest points of an instance x according to a distance function can lead to very similar counterfactual candidates with minor differences between them. Diversity was introduced as the process of generating a set of diverse counterfactual explanations for the same data instance x [93]. This leads to explanations that are more interpretable and more understandable to the user.
- **Feasibility.** This property was introduced by Poyiadzi et al. [54] 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. In Fig. 8, one can see different counterfactual candidates. The closest counterfactual to the data instance x is α. However, this point falls in the decision boundary. Thus, the black-box is not very certain about its prediction, and may lead to biased counterfactual explanations. To address this problem, Poyiadzi et al. [54] argues that counterfactual γ is a better candidate 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 human-interpretable counterfactuals with the least possible feature changes that are achievable and understandable to the user.

From the definitions of plausibility and feasibility, one can conclude that they are related to each other: for a counterfactual to be feasible, it must first be plausible. Plausibility refers to a property that ensures that generated counterfactuals are legitimate. This means that a legitimate counterfactual should never change immutable features such as gender or race. On the other hand, feasibility is related to searching for a counterfactual that does not lead to "paradoxical interpretations". Using the example from Poyiadzi et al. [54], low-skilled unsuccessful mortgage applicants may be told to double their salary. This implies that they need to increase their skill level first, which may lead to counterfactual explanations that are impractical and, therefore, infeasible. Thus, satisfying feasibility automatically guarantees plausible counterfactuals, promoting a higher level of interpretability of counterfactual explanations.

Given the above properties, in the following sections, we will classify the different algorithms found in the literature by (1) their underlying theory (Section 5), and by (2) the above properties (Section 6).

5. Counterfactual approaches in 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 23 algorithms. We created a set of seven different categories representing the *"master theoretical algorithm"* [99] 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, (6) Case-Based Reasoning Approaches, and (7) Probabilistic-Centric approaches. Fig. 10 presents the proposed taxonomy as well as the main algorithms that belong to each category.

- **Instance-Centric.** Corresponds to all approaches that derive from the counterfactual formalism proposed by Lewis [91] and Wachter et al. [53]. These approaches are based on random feature permutations and consist of finding counterfactuals close to the original instance by some distance function. Instance-centric algorithms seek novel loss functions and optimisation algorithms to find counterfactuals. Thus, they are more susceptible to fail the plausibility, the feasibility, and the diversity properties, although some instance-centric algorithms incorporate mechanisms in their loss functions to overcome these issues. Among this subset, FACE and DiCE incorporate mechanisms in their loss functions to consider feasibility and diversity in the generation of counterfactuals.
- **Constraint-Centric.** It corresponds to all approaches that are modelled as a constraint satisfaction problem. Algorithms in this category use different strategies, such as satisfiability modulo theory solvers. The major advantage of these approaches is that they are general and can easily satisfy different counterfactuals properties such as feasibility, diversity, and plausibility.
- **Genetic-Centric.** Corresponds to all approaches that use genetic algorithms as an optimisation method to search for counterfactuals. Since genetic search allows feature vectors to crossover and mutate, these approaches often satisfy properties such as diversity.
- **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 are presented as explanations. Counterfactuals based on these approaches have difficulties satisfying several properties such as plausibility and diversity, and they usually have poor stability.
- **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 SHAP algorithm to take into consideration counterfactuals. Counterfactuals based on these approaches have difficulties satisfying properties such as plausibility and diversity. However, they usually have good stability.
- **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 proximity, plausibility, feasibility, and diversity.

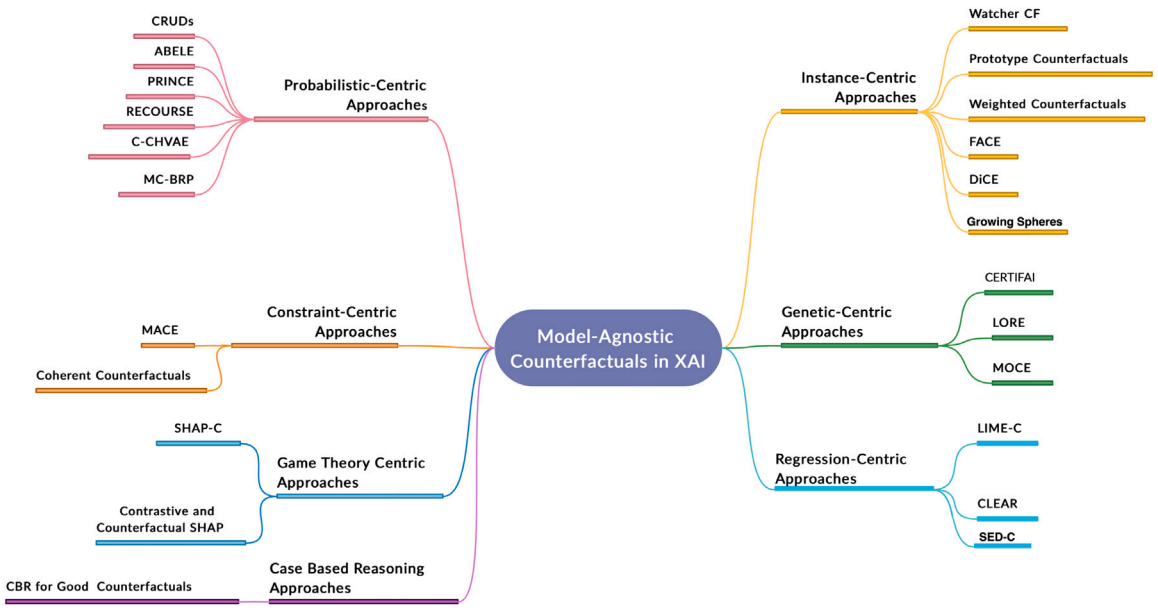

Fig. 10. Proposed taxonomy for model-agnostic counterfactual approaches for XAI.

- **Probabilistic-Centric.** Corresponds to approaches that model the counterfactual generation problem as a probabilistic problem. These approaches often consider random walks, Markov sampling, variational autoencoders (to learn efficient data codings in an unsupervised manner), or using probabilistic graphical models (PGMs). Approaches based on PGMs have the potential to satisfy the causality framework proposed by Pearl [42], promote causability to the user and ensure plausibility and feasibility in the generated counterfactuals.

6. Counterfactual approaches to explainable AI: The algorithms

In this systematic review, we found 23 model-agnostic XAI counterfactual algorithms. We analysed each algorithm in-depth and classified them according to the different properties presented in Section 4. 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 following sections, each algorithm of Table 2 is analysed relative to its grounding *theoretical master algorithm*.

6.1. Instance-centric algorithms

In this section, we summarise 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 optimisation algorithm and the distance function specification.

- **WatcherCF by Wachter et al.** [53]. WatcherCF corresponds to one of the first algorithms in model-agnostic counterfactuals for XAI. They extend the notion of a minimum distance between datapoints that was proposed initially by Lewis [91]. The goal is to find a counterfactual x' as close as possible to the original point x_i as possible such that a new target y' (the counterfactual) is found.
  - **Loss function.** The loss function takes as input the data instance to be explained, x, the counterfactual candidate, x', and a parameter λ that balances the distance in the prediction (first term) against the distance in feature values (second term) [59]. The higher the value of λ, the closer the counterfactual candidate, x', is to the desired outcome, y'. Eq. (10) presents the loss function and respective optimisation problem proposed by Wachter et al. [53]. The authors argue that the type of optimiser is relatively unimportant since most optimisers used to train classifiers are efficient in this approach.

$$\mathcal{L}(x, x', y', \lambda) = \lambda \left(f(x') - y'\right)^2 + d(x, x') \qquad arg \min_{x'} \max_{\lambda} \mathcal{L}(x, x', y', \lambda) \tag{10}$$

  - **Distance function.** Although the choice of optimiser does not impact the search for counterfactuals, the choice of the distance function does. Wachter et al. [53] argue that the L_1-norm normalised 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. Eq. (11) presents the distance function used in their loss function.

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

  - **Optimisation algorithm:** The Adam Gradient descent algorithm is used to minimise Eq. (10).

- **Prototype Counterfactuals by Looveren and Klaise** [100]. Prototype Guided Explanations consist of adding a prototype loss term in the objective result to generate more interpretable counterfactuals. The authors performed experiments with two types of prototypes: an encoder and k-d trees, which resulted in a significant speed-up in the counterfactual search and generation process [100].
  - **Loss function**: The loss function consists in two different steps: (1) guide the perturbations δ towards an interpretable counterfactual x_{cf} which falls in the distribution of

counterfactual class i, and (2) accelerate the counterfactual searching process. This is achieved through Eq. (12),

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

L_{pred} measures the divergence between the class prediction probabilities, L_1 and L_2 correspond to the elastic net regulariser, L_{AE} represents an autoencoder loss term that penalises out-of-distribution counterfactual candidate instances (which can lead to uninterpretable counterfactuals). 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 [100] use the L_2-norm to find the closest encoding of perturbed instances, $ENC(x+\delta)$ of a data instance, x, to its prototype class, $proto_i$. This is given by Eq. (13).

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

- **Optimisation function**: Looveren and Klaise [100] adopted a fast integrative threshold algorithm (FISTA) which helps the perturbation parameter δ to reach momentum for N optimisation steps. The optimisation function uses the L_1 regularisation.

- **Weighted Counterfactuals by Grath et al.** [101]. Weighted counterfactuals extend the WatcherCF approach in two dimensions by proposing: (1) the concepts of positive and weighted counterfactuals, and (2) two weighting strategies to generate more interpretable counterfactuals, one based on global feature importance, the other based on nearest neighbours.
  Traditional counterfactuals address the question *why my load was not granted?* through a hypothetical *what-if* scenario. On the other hand, when the desired outcome is reached, positive counterfactuals address the question *by how much was my loan accepted?*

  - **Loss function.** The weighted counterfactuals are computed in the same way as in WatcherCF [53] as expressed in Eq. (10).
  - **Distance function.** The distance function used to compute weighted counterfactuals is the same as in WatcherCF [53] with the addition of a weighting parameter θ_j,

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

  - **Optimisation algorithm.** While Wachter et al. [53] used gradient descent to minimise the loss function, Grath et al. [101] used the Nelder–Mead algorithm, which is used to find the minimum of a function in a multidimensional space. The Nelder–Mead algorithm is better for dealing with the L_1-norm since it works well with nonlinear optimisation problems for which derivatives may not be known.

  Experiments conducted by Grath et al. [101] showed that weights generated from feature importance lead to more compact counterfactuals and consequently offered more human-understandable interpretable features than the ones generated by nearest neighbours.
- **Feasible and Actionable Counterfactual Explanations (FACE) by Poyiadzi et al.** [54].
  FACE 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 data points and the data instance to be explained [55].
  Under this approach, feasibility refers to the search for a counterfactual that does not lead to *paradoxical interpretations*. For instance, low-skilled unsuccessful mortgage applicants may be told to double their salary, which may be challenging without increasing their skill level. This may render counterfactual explanations that are impractical and sometimes outright offencive [54].

  - **Main Function.** The primary function of FACE's algorithm is given by Eq. (15), where f corresponds to a positive scalar function and γ is a function that connects the path between a data instance x_i and a counterfactual candidate instance x_j.

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

    When the partition $\hat{D}_{f,\gamma}$ converges, Poyiadzi et al. [54] suggest, for a given threshold ϵ, using weights of the form

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

    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 minimised using Dijkstra's shortest path algorithm by approximating the f-distance using a finite graph over the dataset.
  - **Distance Function.** Poyiadzi et al. [54] 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.
  - **Optimisation Function.** Poyiadzi et al. [54] suggested three approaches that can be used to estimate the weights in Eq. (16):

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

    The first equation requires Kernel Density Estimators (KDE) to allow convergence, the second requires a k-NN graph construct, and the third equation requires ϵ-graphs. In their experiments, Poyiadzi et al. [54] found that the third weight equation together with ϵ-graphs generated the most feasible counterfactuals.

- **Diverse Counterfactual Explanations (DiCE) by Mothilal et al.** [48]. DICE is an extension and improvement of WatcherCF [53] 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 more understandable and interpretable counterfactuals. Diversity is formalised as 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 presented in Eq. (18), and is given by a linear combination of three components: (1) a hinge loss function that is a metric that minimises the distance between the model prediction $f(.)'$ for $c_i s$ and an ideal outcome y, $loss(f(c_i), y)$, (2) a proximity factor, which is given by a distance function, and (3) a diversity factor $dpp\_diversity(c_1, \ldots, c_k)$.

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

– **Distance Function.** DiCE uses the L_1-norm normalised by the inverse of the median absolute deviation of feature j over the dataset just like in WatcherCF [53].
– **Optimisation Function.** Gradient descent is used to minimise Eq. (18).

- **Unjustified Counterfactual Explanations (Growing Spheres) by Laugel et al.** [102–104] Growing Spheres consists in approaching the problem of determining the minimal changes to alter a prediction by proposing an inverse classification approach [102]. The authors present the *Growing Spheres* algorithm, which consists of identifying a close neighbour classified differently through the specification of sparsity constraints that define the notion of *closeness*.

  – **Loss Function.** [103] presented a formalisation for binary classifications that simplifies the process for reaching the closest desirable feature. They achieve this by finding an observed value e and classifying it to a class different from x. For instance, $f(e) = f(x)$ indicates that the observation is classified as the same class as x. If it is classified to the opposite class, then the desirable feature is found. The authors define a function $c(x, e)$, corresponding to the cost of moving from x to the observed value e.

$$e^* = \underset{e\in X}{argmin}\{c(x, e) \mid f(e) \neq f(x)\} \quad \text{with} \quad c(x, e) = \|x - e\|_2 + \gamma\,\|x - e\|_0 \tag{19}$$

  – **Distance Function.** Eq. (19) consists in the minimisation of a cost function under the constraint that the observation e is classified into the same class as x. This cost function is defined as a weighted linear equation consisting of L_2-norm and L_0-norm between e and x. The L_2-norm computes the proximity between x and e, while the L_0-norm is used as a weighted average to guarantee that the explanation is human interpretable.
  – **Optimisation Function.** The authors used the proposed growing sphere algorithm as an optimiser. The algorithm applies a greedy method to find the closest feature in all possible directions until the decision boundary is reached.
  In a later work, the authors proposed to distinguish between justified and unjustified counterfactual explanations [103, 104]. In this sense, unjustified explanations refer to counterfactuals resulting from artefacts learned by an interpretable post-hoc model and do not represent the ground truth data while justified explanations refer to counterfactual explanations according to the ground truth data.

6.2. Constraint-centric approaches

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

- **Model-Agnostic Counterfactual Explanations for Consequential Decisions (MACE) by Karimi et al.** [93]
  MACE maps the problem of counterfactual explanation search into a satisfiability modulo theory (SMT) model. MACE receives as input a sequence of satisfiability problems expressing the predictive model, f, the distance function, d, and the constraint functions. The algorithm aims to map these sequences into logical formulae and verify if there is a counterfactual explanation that satisfies a distance smaller than some given threshold. The constraints that are taken into consideration in this approach are *plausibility*, *proximity* and *sparcity*. 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 ϵ, 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 ϵ. MACE is a general algorithm that supports any L_p-norm as a distance function, as well as any number of constraints.
  MACE was able to not only achieve high plausibility (what the authors define as *coverage*) but was also able to generate counterfactuals at more favourable distances than existing optimisation-based approaches [93].

- **Coherent Counterfactuals by Russell** [98]
  This approach focuses on generating diverse counterfactuals based on "mixed polytope" methods to handle complex data with a contiguous range or an additional set of discrete states. Russell [98] created a novel set of criteria for generating diverse counterfactuals to integrate them with the mixed polytope method and to map them back into the original space. Before achieving the two targets (coherence and diversity), Russell [98] firstly offers a solution on generating a counterfactual which is similar to Wachter et al. [53]'s counterfactuals by finding the minimum change in the counterfactual candidate that would lead to a change in the prediction (using the L_1-norm). Then, the author proposes the mixed polytope as a novel set of constraints. The program uses an integer programming solver and receives a set of constraints to find coherent and diverse counterfactuals efficiently.

6.3. Genetic-centric approaches

In this section, we summarise 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.** [105]
  This approach provides interpretable and faithful explanations by learning a local interpretable predictor on a synthetic neighbourhood 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 neighbour instances of x using an ad-hoc genetic algorithm. The interpretable model used to fit the data corresponds to a decision tree from which counterfactual rules can be extracted as explanations.
  The distance function used in this algorithm is given by Eq. (21). 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 using the L_2-norm to normalise the continuous features. Assuming h

corresponds to categorical features and $m-h$ to continuous ones, then the distance function is given by

$$d(x,z)=\frac{h}{m}.SimpleMatch(x,z)+\frac{m-h}{m}.NormEuclid(x,z). \quad (21)$$

- **Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models (CERTIFAI) by Sharma et al.** [106]
  CERTIFAI is a custom genetic algorithm based explanation 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: creating an original genetic framework, selecting a distance function, and improving counterfactuals with constraints. Consider c as the counterfactual candidate instance of x and $d(x,c)$ the distance between them. The distance function used is the L_1-norm normalised by the median absolute deviation (MAD) as proposed by Wachter et al. [53]. The goal is to minimise the distance between x and c by applying a genetic algorithm.
  Given a variable W, we define the space from which individuals can be generated, to ensure feasible solutions. By taking n dimensions as input, multiple constraints need to be created to match with continuous, categorical, and immutable features. For instance, $W_1, W_2 ..., W_n$ is defined for continuous feature constraints as $W_i \in W_{imin,W_{imax}}$, and $W_i \in W_1, W_2, \ldots, W_j$ is for categorical variables. Finally, a feature i for an input x should be immutable by setting $W_i = X_i$.
  The robustness and fairness of the population of the generated counterfactuals are given by

$$CERScore(model)=\frac{E}{x}[d(x,c^*)]. \quad (22)$$

  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.** [92]. This approach consists of a multi-objective counterfactual explanation algorithm which translates the counterfactual search into a multi-objective optimisation problem based on genetic algorithms. This approach brings the benefit of providing a diverse set of counterfactuals with a variety of trade-offs between the proposed objectives and maintains diversity in feature space at the same time.
  Dandl et al. [92] proposed a four-objective loss equation to generate an explanation:

$$L(x,x',y',X^{obs})=(o_1(f\hat{}(x',y'),o_2(x,x'),o_3(x,x'),o_4(x',X^{obs}))). \quad (23)$$

  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 closest possible result from the prediction of counterfactual x' to the desired prediction y'. It minimises the distance between $f(x')$ and y', and calculates it through the $L_1$1-norm. Objective 2 states that the ideal counterfactual should be as similar as possible to instance x. It quantifies the distance between x' and x using Grower's distance. Objective 3, o_3 is used to calculate the sparse feature changes through L_0-norm. This norm is necessary because Grower distance can handle numerical and categorical features but cannot count how many features were changed. Finally, Objective 4 states that the ideal counterfactual should have similar feature value combinations as the original data point. The solution is to measure how "likely" a data point uses the training data, $Xobs$, which promotes *plausability*.
  Dandl et al. [92] 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.

6.4. Regression-centric approaches

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

- **Counterfactual Local Explanations via Regression (CLEAR) by White and d'Avila Garcez** [107]: This method aims to provide a counterfactual explained by regression coefficients, including interaction terms, and significantly improve the fidelity of the regression. [107] firstly generates a boundary for counterfactual explanations, which state the minimum changes necessary to flip a prediction's classification. Then, it builds local regression models using the boundary of the counterfactuals to measure and improve the fidelity of its regressions.
  CLEAR proposes a *counterfactual fidelity error*, CFE, which is based on the concept of b-perturbation. It compares each b-perturbation with an estimation of that value, $estimated b$-perturbation, calculated by a local regression. CFE is given by

$$CFE=|estimated\ b-perturbation-b-perturbation|. \quad (24)$$

  The generation of counterfactuals in CLEAR has the following steps: (1) determine x's boundary perturbations; (2) generate labelled synthetic observations; (3) create a balanced neighbourhood dataset; (4) perform a stepwise regression on the neighbourhood dataset, under the constraint that the regression curve should go through x; (5) Estimate the b-perturbations; (6) Measure the fidelity of the regression coefficients; (7) Iterate until the best explanation is found.
  [107] compared the performance in terms of fidelity between CLEAR and LIME. The result showed that CLEAR has significantly higher fidelity than LIME in five case studies.
- **Local Interpretable Model-Agnostic Explanations-Counterfactual (LIME)-C by Ramon et al.** [108]. LIME-C is a hybrid solution that connects additive feature attribution explanations (like in LIME) with counterfactuals. The motivation for this hybrid solution starts from an assumption which states that if the importance-rankings of the features are sufficiently accurate, it may be possible to compute counterfactuals from them.
  Additive feature attribution methods use an explanation model g as an interpretable approximation of the trained classification model C, which can be represented by a linear model [108]:

$$g(x')=\phi_0+\sum_{j=1}^{m}\phi_j x'_j \quad (25)$$

  In Eq. (25), $x'_j \in 0,1$ is the binary representation of x'_j (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 $\phi_0, \phi_j \in \mathcal{R}$. To generate a ranked list of counterfactuals, the authors used SEDC, a linear algorithm for finding counterfactuals proposed by [95].
  [108] points out that this method is stable and effective 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.
- **Search for Explanations for Document Classification (SEDC) by Martens and Provost** [95]
  This approach was proposed in the domain of information retrieval back in 2014. It consists of generating explanations for the user's understanding of the system and also for model inspection.

SEDC was one of the first works that used Lewis [91] definition of counterfactual in an algorithm that provides minimal sets of words as explanations, such as removing all words within this set from the document that changes the predicted class from the class of interest.
SEDC outputs minimum-size explanations for linear models by ranking all words appearing in the document through the product $\beta_{jX_{ij}}$, where β_j is the linear model coefficient. The explanation with the top-ranked words is an explanation of the smallest size, and therefore a counterfactual. Cosine-similarity is used to measure the proximity between a document and the counterfactual document candidate.

6.5. Game theory centric approaches

This section summarises the algorithms that we classified as Game Theory Centric using the proposed taxonomy.

- **SHAP Counterfactual (SHAP-C) by Ramon et al.** [108]
  SHAP-C is a hybrid algorithm that combines Kernel SHAP [35] with SEDC [95].
  SHAP-C works as follows. Given a data point and a black-box predictive model, first, we compute the respective Shapley values using kernel SHAP. The algorithm ranks the most important features by their respective SHAP values and adds these features to a set called the *evidence counterfactual*. The algorithm then proceeds similarly as in SEDC: the most important features are perturbed so that a minimum set of perturbations is found to flip the prediction of the datapoint. This *evidence counterfactual* is returned as the explanation.
- **Contrastive and Counterfactual SHAP (SHAP-CC) by Rathi** [109].
  SHAP-CC attempts to generate partial post-hoc contrastive explanations with a corresponding counterfactual. Rathi [109] used a P-contrastive methodology for generating contrastive explanations that would allow the user to understand why a particular feature is important and why another specific feature is not.
  The main idea of this explanation is to consider a P-contrast question which is equivalent to the format *"Why [desired class]"* instead of *[predicted class]?*. To answer these questions, Rathi [109] computed the Shapely values for each of the possible target classes. A negative Shapely value indicates the features that have negatively contributed to the specific class classification. Rathi [109] generate a "Why P not Q" explanation by breaking it down into two questions: "why P?" and "Why not Q". The positive and negative Shapley values are given as an answer to these questions, respectively. The final contrastive counterfactual explanation is presented to the user in terms of natural language.

6.6. Case-based reasoning approaches

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

- **Case-based reasoning (CBR) for Good counterfactual by Keane and Smyth** [96]
  CBR for good counterfactuals uses a case-based system where examples of good counterfactuals are stored. By good counterfactuals, Keane and Smyth [96] understands as counterfactuals that are sparse, plausible, diverse, and feasible. The authors also introduce the *explanatory coverage* and *counterfactual potential* as properties of CBR systems that can promote good counterfactuals. In their algorithm, Keane and Smyth [96] refers to the pairing of a case and its corresponding good counterfactual as an *explanation case* or XC. The goal is to generate good counterfactuals by retrieving, reusing, and revising a nearby explanation case by taking the following steps: (1) identify the XC that is most similar to the query datapoint p. In other words, XC corresponds to the explanatory coverage between two data points, $xc(x, x')$; (2) for each of the features matched in $xc(x, x')$, we map these features from p to the new generated counterfactual p' in the same way, we add to p' the most different features in $xc(x, x')$. This procedure guarantees the diversity of the counterfactuals; (3) through the definition of a counterfactual, p' needs to provide a prediction contrary to p. This implies that p' needs to go through the predictive black-box and be returned to the user if the prediction flips; (4) otherwise, an adaptation step is applied to revise the values of the different features in p' until there is a change of class.

6.7. Probabilistic-centric approaches

In this section, we summarise the algorithms that we classified as Probabilistic-Centric using the proposed taxonomy:

- **Provider-side Interpretability with Counterfactual Explanations in Recommender Systems (PRINCE) proposed by Ghazimatin et al.** [110]
  This approach aims to detect the actual cause of a recommendation by a heterogeneous information network (HIN) with users, items, reviews, and categories. It identifies a small set of a user's actions by removing actions that would replace the recommended item with a different item.
  This approach provides user explanations by displaying what they can do to receive more relevant recommendations to the users. Personalised PageRank (PPR) scores were chosen as the recommender model to create a heterogeneous knowledge network. PRINCE is based on a polynomial-time algorithm for searching the space for subsets of user behaviours that could lead to a recommendation. By adopting the reverse local push algorithm to a dynamic graph environment, the algorithm efficiently computes PPR contributions for groups of actions about an object.
  Experiments performed by Ghazimatin et al. [110] using data from Amazon and Goodreads showed that simpler heuristics fail to find the best explanations. On the other hand, PRINCE can guarantee optimality because it outperformed baselines in terms of interpretability in user studies.
- **Counterfactual Conditional Heterogeneous Autoencoder (C-CHVAE) by Pawelczyk et al.** [111,112]
  C–CHVAE uses Variational Autoencoders (VAE) to search for faithful counterfactuals, which consist of counterfactuals that are not local outliers and that are *connected* to regions with significant data densities (similar to the notion of feasibility introduced by Poyiadzi et al. [54]). Given the original dataset, the data is converted into an *encoding vector* where the it is represented in a lower dimension space, and each dimension represents some learned probability distribution. It is this encoder that specifies which low-dimensional neighbourhood one should look for potential counterfactuals. The next steps consist of perturbing the low dimensional data and passing it through a decoder, reconstructing the lower dimensional potential counterfactuals into their original space. Finally, the newly generated counterfactuals are given to a pre-trained black-box to assess whether the prediction has been altered.
- **Monte Carlo Bounds for Reasonable Predictions (MC-BRP) by Lucic et al.** [113]. MC-BRP is an algorithm that focuses on predictions where Monte Carlo sampling methods generate a set of permutations of an original data point instance that result in reasonable predictions.
  Given a local instance, x_i, a set of important features $\Phi^{(x_i)}$, a black-box model f, and an error threshold ϵ, MC-BRP uses Tuckey's fence to determine outliers (predictions with high errors) for each feature of the set of important features,

  $$\epsilon > Q3(E) + 1.5(Q3(E) - Q1(E)),$$

where $Q_1(E)$, $Q_3(E)$ are the first and third quartiles of the set of errors, E, of each feature, respectively. Tukey's fence will return a set of boundaries for which reasonable predictions would be expected for each of the most important features. Using these boundaries, MC-BRP generates a set of permutations using Monte Carlo sampling methods, which will be passed into the black-box f to obtain a new prediction. Finally, a trend is computed based on Pearson correlation over the reasonable new predictions. The reasonable bounds for each feature are recomputed and presented to the user in a table.

- **Adversarial black-box Explainer generating Latent Exemplars (ABELE) by Guidotti et al.** [114]
  ABELE is a local model-agnostic explainer for image data that uses Adversarial AutoEncoders (AAEs) that aim at generating new counterfactuals that are highly similar to the training data.
  ABELE generates counterfactuals in four steps: (1) by generating a neighbourhood in the latent feature space using the AAEs, (2) by learning a decision tree on the generated latent neighbourhood by providing local decision and counterfactuals rules, (3) by selecting and decoding exemplars and counter-examples satisfying these rules, and (4) by extracting the saliency maps out of them. Guidotti et al. [114] found that ABELE outperforms current state-of-the-art algorithms, such as LIME, in terms of coherence, stability, and fidelity.
- **CRUDS: Counterfactual Recourse Using Disentangled Subspaces by Downs et al.** [115]
  CRUDs is a probabilistic model that uses conditional subspace variational autoencoders (CSVAEs) that extracts latent features relevant for a prediction. CSVAE partitions the latent space into two parts: one to learn representations that predict the labels, and another one to learn the remaining latent representations required to generate data.
  In CRUDs, counterfactuals that target desirable outcomes are generated using CSVAEs in four major steps: (1) disentangling latent features relevant for classification from those that are not, (2) generating counterfactuals by changing only relevant latent features, filtering counterfactuals given constraints, and (4) summarise counterfactuals for interpretability.
  Downs et al. [115] evaluated CRUDS on seven synthetically generated and three real datasets. The result indicates that CRUDS counterfactuals preserve the true dependencies between the covariates in all datasets except one.
- **Recourse: Algorithmic Recourse Under Imperfect Causal Knowledge by Karimi et al.** [116]
  Traditional works on counterfactuals for XAI focus on finding the nearest counterfactual that promotes a change in the prediction to a favourable outcome. Recourse generates actions that an individual can perform to obtain a favourable outcome. This implies a shift from minimising a distance function to optimising a personalised cost function [116,117].
  To the best of our knowledge, recourse is the only model-agnostic algorithm in the literature that attempts to use a causal probabilistic framework as proposed in Pearl [42] grounded on structural causal models to generate counterfactuals, and where *recourse actions* are defined as interventions of the form of $do-calculus$ operations.
  Karimi et al. [116] highlight that it is very challenging to extract a structured causal model (SCM) from observed data [118], and in their work, they assume an imperfect and partial SCM to apply recourse, which can lead to incorrect counterfactuals. To overcome this challenge, Karimi et al. [116] proposed two probabilistic approaches that relax the strong assumptions of an SCM: the first one consists of using additive Gaussian noise and Bayesian model averaging to estimate the counterfactual distribution; the second removes any assumptions on the structural causal equations by computing the average effect of recourse actions on individuals. Experiment results on synthetic data showed that *Recourse* was able to make more reliable recommendations under a partial SCM than other non-probabilistic approaches.

6.8. Summary

We conducted a thorough systematic literature review guided by the argument that for an XAI system to have causability, then the system needs to be underpinned by a probabilistic causal framework such as the one proposed in Pearl [42]. We believe that counterfactual reasoning could provide human causal understandings of explanations, although some authors challenge this notion. Zheng et al. [119] conducted studies to investigate whether presenting causal explanations to a user would lead to better decision-making. Their results were mixed. They found that if the user has prior and existing domain knowledge, then presenting causal information did not improve the decision-making quality of the user. On the other hand, if the user did not have any prior knowledge or beliefs about a specific task, then causal information enabled the user to make better decisions.

We also found that the majority of the counterfactual generation approaches are not grounded on a formal and structured theory of causality as proposed by Pearl [42]. Current counterfactual explanation generation approaches for XAI 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. [37] highlighted in his work about medical decision making.

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

7. Counterfactual approaches to explainable AI: Applications

This work is motivated by the 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 [20,45,46,82,139–141]. One way to achieve this causal understanding is through counterfactuals.

One of the main areas that showed a need for counterfactual explanations is in medical decision support systems. As pointed by Holzinger et al. [142], medical decision-making faces several challenges ranging from small labelled datasets to the integration, fusion, and mapping of heterogeneous data to appropriate visualisations [143]. Structured causal models that provide explanatory factors of the data could be used to support medical experts. However, learning causal relationships from observational data is a very difficult problem [144,145].

XAI is very relevant for industries [146]. Another application of counterfactuals in XAI that is highly mentioned in the literature is loan credit evaluation. Grath et al. [101] developed a model-agnostic counterfactual explainer with an interface that uses weights generated from feature importance to generate more compact and intelligible counterfactual explanations for end users. Lucic et al. [113] also developed a model-agnostic counterfactual explanation in the context of a challenge from the finance industry's interest in exploring algorithmic explanations [147].

Recently, interfaces that generate counterfactuals as explanations have been proposed in the literature. The "What-IF tool" [148] is an open-source application that allows practitioners to probe, visualise, and analyse machine learning systems with minimal coding. It also enables the user to investigate decision boundaries and explore how general data points affect the prediction.

ViCE (Visual Counterfactual Explanations for Machine Learning Models) [149] is an interactive visual analytics tool that generates

Table 2
Classification of collected model-agnostic counterfactual algorithms for XAI based on different properties, theoretical backgrounds and applications.

					Properties						
Theory/Approach	Algorithms	Ref.	Applications	Code?	Proximity	Plausibility	Sparsity	Diversity	Feasibility	Optimisation	Causal?
Instance-Centric	WatcherCF	[53]	C [Tab/Img]	Yes [120] [Algo: CF]	✓ [L_1/L_2-norm]	✗	✓	✗	✗	Gradient Descent	✗
	Prototype Counterfactuals	[100]	C [Tab/Img]	Yes [120] [Algo: CFProto]	✓ [L_1/L_2-norm]	✗	✓ [kd-trees/auto-encoders]	✗	✗	FISTA	✗
	FACE	[54]	C [Tab/Img]	Yes [121]	✓ [L_2-norm]	✗	✓	✓	✓	KDR/K-NN [graph/ϵ-graph]	✗
	Weighted Counterfactual	[101]	C [Tab]	No	✓ [L_1-norm]	✗	✓	✗	✗	Gradient Descent	✗
	Growing Spheres	[102–104]	C [Tab/Txt/Img]	Yes [122]	✓ [L_0-norm]	✗	✓	✗	✗	Growing Spheres	✗
	DICE	[48]	C [Tab]	Yes [123]	✓ [L_1/L_2-norm]	✓	✓	✓	✓	Gradient Descent	✗
Probabilistic-Centric	CRUDS	[115]	C [Tab]	No	✓	✗	✓ [Variation Autoencoders]	✓	✓	Gradient Descent	✗
	PRINCE	[110]	C/R [Tab/Txt]	Yes [124]	✓	✗	✓ [Random Walk]	✗	✗	PageRank	✗
	C-CHVAE	[111,112]	C [Tab]	Yes [125]	✓	✗	✓ [Variation Autoencoders]	✗	✗	Integer Programming Optimisation	✗
	ABELE	[114]	C [Img]	Yes [126]	✓	✗	✓ [Variational Autoencoders]	✗	✗	-	✗
	RECOURSE	[116]	C [Tab]	Yes [127]	✓	✓	✓ [Variation Autoencoders]	✓	✓	Gradient Descent	✓
	MC-BRP	[113]	R [Tab]	Yes [128]	✓	✗	✓	✓	✗	Monte Carlo	✗
Constraint-Centric	MACE	[93]	C [Tab]	Yes [129]	✓ [$L_0/L_1/L_\infty$-norm]	✓	✓ [constraint satisfaction]	✓	✓	SMT	✗
	Coherent Counterfactuals	[98]	C/R [Tab/Txt]	Yes [130]	✓ [L_1-norm]	✓	✓ [mixed polytopes]	✓	✓	Gurobi Optimisation	✗

(*continued on next page*)

Table 2 (*continued*).

Theory/Approach	Algorithms	Ref.	Applications	Code?	Properties						
					Proximity	Plausibility	Sparsity	Diversity	Feasibility	Optimisation	Causal?
Genetic-Centric	MOCE	[92]	C [Tab]	Yes [131]	✓ [L_0/L_1-norm]	✗	✓ [min feature changes]	✓ [mutations]	✗	NSGA-II	✗
	CERTIFAI	[106]	C [Tab/Img]	Yes [132]	✓ [L_1-norm/SSIM]	✗	✗	✓ [mutations]	✓	Fitness	✗
	LORE	[105]	C [Tab]	Yes [133]	✓ [L_2-norm/Match]	✗	✓	✓ [mutations]	✗	Decision Tree Model	✗
Regression-Centric	LIME-C	[108,134]	C/R [Tab/Txt/Img]	Yes [135]	✓	✗	✗	✗	✗	Additive Feature Attribution	✗
	SED-C	[95,108,134]	C [Txt]	Yes [136]	✓ [cosine similarity]	✗	✓	✗	✗	-	✗
	CLEAR	[107]	C [Tab]	Yes [137]	✓ [L_2-norm]	✗	✓ [min feature changes]	✗	✗	Regression	✗
Game Theory Centric	SHAP-C	[108,134]	C/R [Tab/Txt/Img]	Yes [138]	✓	✗	✗	✗	✗	Shapley Values	✗
	SHAP-CC	[109]	C/R [Tab]	No	✓	✗	✗	✗	✗	Shapley Values	✗
Case Based Reasoning	CBR for Good Counterfactuals	[96]	C [Tab/Txt]	No	✓ [L_1-norm]	✓	✓ [counterfactual potential]	✓	✓	Nearest Unlikely Neighbour	✗

instance-centric counterfactual explanations to contextualise and evaluate model decisions in a home equity line of credit scenario. ViCE highlights counterfactual explanations and equips users with interactive methods to explore both the data and the model.

DECE (Decision Explorer with Counterfactual Explanations for Machine Learning Models) [150] is another example of an interactive visual analytics tool that generates counterfactuals at an instance and subgroup levels. The main difference to ViCE is that DECE allows users to interact with the counterfactuals to find more actionable counterfactuals that suit their needs. DECE showed effectiveness in supporting decision exploration tasks and instance explanations.

8. Towards causability: Opportunities for research

Although the demand for providing XAI systems that promote causability, the literature is very scarce in this aspect. We only found one recent article that proposed an explanation framework (FATE) based on causability [151]. This framework focuses on human interaction, and the authors used the system causability scale proposed by [20] to validate the effectiveness of their system's explanations.

Shin [151] highlight that causability represents the quality of explanations and emphasise that it is an antecedent role to explainability. Furthermore, they found that transparency, fairness, and accuracy are critical in improving user trust in the explanations. In general, this framework is a guideline for developing user-centred interface design from 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, the FATE causability system is not underpinned by any formal theory of causality. The causability metrics applied in this work focused on the interaction of the human with the system.

Holzinger et al. [20] 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 visualisation techniques that can be trainable by medical experts, as the specialists can survey the underlying explanatory factors of the data. Another point is to formalise a structural causal model of human decision-making and delineating features in the model. Holzinger [45] 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. [82] proposes a system causability scale framework as an evaluation tool for causability in XAI systems.

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. [105] stated, explainability is the ability to present interpretations in a meaningful and effective way to a human user. We argue that for a system to be both explainable and promote causability, 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 [152].

8.1. The main characteristics of a causability system

In this paper, we argue that the following properties should be satisfied for a system to promote causability.

- **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 make 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, although some authors challenge this notion. Zheng et al. [119] found that providing causal information to human users in some tasks resulted in poor decision-making. Zheng et al. [119] conducted studies to investigate whether presenting causal explanations to a user would lead to better decision-making. Their results were mixed. They found that if the user has prior and existing domain knowledge, then presenting causal information did not improve the decision-making quality of the user. On the other hand, if the user did not have any prior knowledge or beliefs about a specific task, then causal information enabled the user to make better decisions. More work is needed on whether causal information leads to better decisions. Many unstudied factors may contribute to this diverse literature on the topic ranging from human cognitive bias to how the explanations are presented in the interface (supporting interactivity or not).
- **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 [86]. Counterfactual reasoning induces mental representations of an event that happened and representations of some other event alternative to it [56]. It has been recognised in the literature that counterfactuals tend to help humans make causal judgements [153]. Additionally, humans tend to think in a cause/effect way, but not in a strict probabilistic sense [154]. It follows that for a machine to achieve a certain degree of human intelligence, 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 [20,82]. 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 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 [142,155] containing contextual knowledge and domain-specific information. This interactive process can 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 to 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. One way to achieve this is by estimating causal effects

using observational data. Causal analysis with observational data is one research direction that can promote causal analysis in explainable AI by matching units with similar covariate distributions for treatment-effect estimations. Recently, some authors proposed to create treatment-control matches for categorical data in the potential outcomes framework that enable the creation of a hierarchy of covariate combinations that can optimally match for treatment effect estimations [156]. These cause/effect estimations promote interpretability and causability, since they can explain what features caused a specific outcome to occur. Extensions of the potential outcome framework with almost matching covariate estimations can be found in [157,158].

- **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 [159] and consequently achieve causability.

9. Answers to research questions

This section summarises the key points presented throughout this work by answering the research questions that guided our research.

9.1. RQ1 & RQ2: What are the main theoretical approaches and algorithms for counterfactuals in XAI?

Our systematic literature review revealed many different counterfactual algorithms proposed in the literature. We were able to identify key elements shared by these algorithms based on how the optimisation problem was framed and by considering the counterfactual generation process. We classified the existing model-agnostic-XAI by their *"master theoretical algorithm"* from which each algorithm derived:

- **Instance-Centric.** These approaches are based on random feature permutations and finding counterfactuals closed to the original instance by some distance function. These approaches are relatively straightforward to implement. However, the generated counterfactuals are susceptible to fail the plausibility and the diversity properties, although some algorithms incorporate mechanisms to overcome this issue (for instance, FACE and DiCE). Examples of algorithms that fall in this category are WatcherCF [53], prototype counterfactuals [100], weighted counterfactuals [101],FACE [54], DiCE [48], and Growing Spheres [103].
- **Constraint-Centric.** These approaches are modelled as a constraint satisfaction problem. The primary advantage of these approaches is that they are general and can easily meet most of counterfactuals properties including plausibility, diversity, and feasibility. Examples of algorithms that fall in this category are MACE [93], and Coherent Counterfactuals [98].
- **Genetic-Centric.** These approaches generate counterfactuals using the principles of genetic algorithms. Due to genetic principles such as mutation or crossover, these approaches often satisfy properties such as proximity and diversity. Examples of algorithms that fall in this category are CERTIFAI [106], MOCE [92], and LORE [105].
- **Regression-Centric.** These approaches have LIME as their underlying framework, and they use linear regression to fit a set of permutated features. Counterfactuals based on these approaches have difficulties satisfying several properties such as plausibility and diversity. Examples of algorithms that fall in this category are LIME-C [108], SED-C [95], and CLEAR [107].
- **Game Theory Centric.** These approaches have SHAP as their underlying framework, and they use Shapley values to determine the local feature relevance. Counterfactuals based on these approaches also have difficulties satisfying several properties such as plausibility and diversity. Examples of algorithms that fall in this category are SHAP-C [108], and SHAP-CC [109].
- **Case-Based Reasoning.** These approaches are inspired by the case-based reasoning paradigm of artificial intelligence and cognitive science. Since they store in-memory examples of good counterfactuals, these approaches tend to satisfy different counterfactual properties, such as plausibility and diversity. An example of an algorithm that falls in this category is CBR Explanations by Keane and Smyth [96].
- **Probabilistic-Centric.** These approaches mainly use probabilistic models to find the nearest counterfactuals. Approaches such as *recourse* [116] have the potential to generate causal counterfactuals based on the causality framework proposed by Pearl [42]. However, as the authors acknowledge, it is challenging to learn causal relationships from observational data without introducing assumptions in the causal model.

This research suggests that current model-agnostic counterfactual algorithms for explainable AI are not grounded on a causal theoretical formalism and, consequently, might not promote causability to a human decision-maker. Our findings show that the explanations derived from most of the model-agnostic algorithms in the literature provide spurious correlations rather than cause/effects relationships, leading to sub-optimal, erroneous, or even biased explanations. This opens the door to new research directions on incorporating formal causal theories of causation in XAI. The closest work that we found that meets this goal is the *Recourse* algorithm [116]. However, research is still needed to investigate the extraction of structured causal models from observational data.

There are also novel model-agnostic approaches proposed in the literature of XAI based on probabilistic graphical models. For instance, Moreira et al. [160] proposed to learn a local Bayesian network that enables the user to see which features are correlated (or conditional independent from) the class variable. They found four different rules that can measure the degree of confidence of the interpretable model over the explanations and provide specific recommendations for the user. However, this model is not causal, and further research is needed to understand if such structures can be mapped into structured causal models.

9.2. RQ3: What are the sufficient and necessary conditions for a system to promote causability (applications)?

The primary purpose of this research work is to highlight some properties that find relevant and necessary for causability systems. We proposed the following properties.

- Explanations need to be grounded on a structured and formal theory of Causality. This will enable the usage of a framework of algorithms for causal discovery that have been proposed throughout the years [44].
- Explanation algorithms need to be computed in the form of counterfactuals. Due to the evidence from cognitive science and social sciences, counterfactuals are among the best approaches to promote human understandability and interpretability [22] although some authors challenge this [119].
- Explanations need to be Human-Centric. Explanations need to be specific to the user's needs: a medical doctor will be interested in different explanations from a standard user.
- The user should be able to interact with the generated explanations. The interaction with explanations can help the user increase the levels of understandability and interpretability of the

internal workings of the XAI algorithm. Probabilistic inference is a promising tool to provide answers to users' questions regarding explanations.
- Explainable AI systems need to be complemented with semantic annotations of features and domain knowledge. To achieve explainability, contextual knowledge and domain-specific information need to be included in the system.

9.3. RQ4: What are the pressing challenges and research opportunities in XAI systems that promote causability?

This literature review enabled us to understand the current pressing challenges and opportunities involved in creating XAI models that promote causability. We identified the following research opportunities that can be used for future research in the area.

- **Causal Theories for XAI.** Pearl [88] argues that causal reasoning is indispensable for machine learning to reach the human-level artificial intelligence since it is the primary mechanism of humans to be aware of the world. As a result, the causal methodology gradually becomes a vitally important component in explainable and interpretable machine learning. However, most current interpretability techniques pay attention to solving the correlation statistic rather than the causation. Therefore, causal approaches should be emphasised to achieve a higher degree of interpretability. The reason why causal approaches for XAI are scarce is that finding causal relationships from observational data is very hard and still an open research question [144].
- **Standardised Evaluation Metrics for XAI.** The field of metrics for XAI is also a topic that needs development. Measures such as stability or fidelity [161] are not very clear for counterfactuals [29]. Ultimately, XAI metrics should be able to answer the following question: *how does one know whether the explanation works and the user has achieved a pragmatic understanding of the AI?* [30] We highlight that one research concern in XAI should be to develop generalised and standardised evaluation protocols for XAI in different levels: Objective Level (user-free), Functional Level (functionality-oriented), and User Level (human-centric). The main challenges consist in deriving standardised protocols that could fit so many algorithms underpinned by different *master theoretical approaches* and at so many different levels. However, some interesting works have already been proposed in terms of causability [45,46,82].
- **Intelligent Interfaces for Causability in XAI.** XAI's basilar applications lie at the core of Intelligent User Interfaces (IUIs). Rather than generating explanations as linear symbolic sequences, graphical interfaces enable people to visually explore ML systems to understand how they perform over different stimuli. The What-If tool [148] provides an excellent example, enabling people to visualise model behaviour across multiple models and subsets of input data and for different ML fairness metrics. Such visual techniques leverage the human visual channel's high bandwidth to explore probabilistic inference, allowing humans to interact with explanations while recommending different descriptions. Taking advantage of the innate human ability to spot patterns, these methods can provide better answers than purely automatic approaches.

  In a related direction van der Waa et al. [162] propose a framework that considers users' experience and reactions to explanations and evaluates these effects in terms of understanding, persuasive power, and task performance. This user-centric approach is crucial to assess assumptions and intuitions to yield more effective explanations effectively.

  Recent Intelligent Exploration Interfaces focus on making explanations accessible to non-expert users to interpret the underlying models better. Hoque and Mueller [163] argues that predictive and interactive models based on causality are inherently interpretable and self-contained. They developed Outcome Explorer, a causality-guided interactive interface that allows experts and non-experts to acquire a comprehensive understanding of the models.

10. Conclusion

We conducted a systematic literature review to determine the modern theories underpinning model-agnostic counterfactual algorithms for XAI and analyse whether existing algorithms can promote causability. 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, case-based reasoning centric, and probabilistic-centric. Our research also showed that model-agnostic counterfactuals are not based on a formal and structured theory of causality as proposed by [42]. 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.

We conclude this survey by highlighting new key points to promote causability in XAI systems, which derive from formal theories of causality such as inference, counterfactuals, and probabilistic graphical models. Causal models are a new research area, bursting with exciting new research challenges and opportunities for XAI approaches grounded on probabilistic theories of causality and graphical models. Indeed this field is highly relevant to Intelligent User Interfaces (IUIs) [164] by its very nature, both in terms of content generation engines and user interface architecture. Therefore, more than a contraption powered by robust and effective causal models, XAI can be seen as a cornerstone for next-generation IUIs. This can only be achieved by marrying sound explanations delivered by fluid storytelling to persuasive and articulate argumentation and a harmonious combination of different interaction modalities. These will usher in powerful engines of persuasion, ultimately leading to the rhetoric of causability.

CRediT authorship contribution statement

Yu-Liang Chou: Conceptualization, Methodology, Software, Visualization, Writing – review & editing. **Catarina Moreira:** Conceptualization, Visualization, Writing – review & editing, Supervision. **Peter Bruza:** Reviewing , Supervision. **Chun Ouyang:** Reviewing and editing. **Joaquim Jorge:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was partially supported by Portuguese government national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020.

This work was also partially supported by Queensland University of Technology (QUT) Centre, Australia for Data Science First Byte Funding Program and by QUT's Women in Research Grant Scheme.

References

[1] W. Tan, P. Tiwari, H.M. Pandey, C. Moreira, A.K. Jaiswal, Multi-modal medical image fusion algorithm in the era of big data, Neural Comput. Appl. (2020).

[2] Z.C. Lipton, The mythos of model interpretability, Communications ACM 61 (10) (2018) 36–43.

[3] D. Doran, S. Schulz, T.R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, in: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 Co-Located with 16th International Conference of the Italian Association for Artificial Intelligence, 2017, arXiv:1710.00794.

[4] C.T. Ramaravind K. Mothilal, Examples are not enough, learn to criticize! Criticism for Interpretability, in: Proceedings of the 2020 Conference on Fairness, Accountability, and TransparencyJanuary, 2020.

[5] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "Right to explanation", AI Mag. 38 (2017) 50–57.

[6] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books, 2017.

[7] 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.

[8] A. Lau, E. Coiera, Do people experience cognitive biases while searching for information? J. Am. Med. Inf. Assoc. 14 (2007) 599–608.

[9] G. Saposnik, D. Redelmeier, C.C. Ruff, P.N. Tobler, Cognitive biases associated with medical decisions: a systematic review, BMC Med. Inform. Decis. Mak. 16 (2016) 138.

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

[11] 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.

[12] 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.

[13] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proc. Natl. Acad. Sci. USA 115 (2018) 3635–3644.

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

[15] M. Kosinski, Y. Wang, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, J. Personal. Soc. Psychol. 114 (2018) 246–257.

[16] 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.

[17] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arxiv:1702.08608.

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

[19] J.W. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. 116 (2019) 22071–22080.

[20] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (2019) e1312.

[21] A. Páez, The pragmatic turn in explainable artificial intelligence (XAI), Minds Mach. (Dordrecht) 29 (2019) 441–459.

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

[23] 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 (5) (2018) 93:1–93:42.

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

[25] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[26] 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, Inf. Fusion 58 (2020) 82–115.

[27] S. Mohseni, N. Zarei, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, 2020, pp. 1–45, CoRR cs.HC/1811.11839.

[28] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (2021) 593.

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

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

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

[32] 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.

[33] 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.

[34] 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 International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[35] 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, pp. 4765–4774.

[36] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

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

[38] 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.

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

[40] 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.

[41] B. Lake, T. Ullman, J. Tanenbaum, S. Gershman, Building machines that learn and think like humans, Brain Behav. Sci. 40 (2017) e253.

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

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

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

[45] A. Holzinger, Explainable AI and multi-modal causability in medicine, I-Com 19 (2020) 171–179.

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

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

[48] 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.

[49] J. Halpern, J. Pearl, Causes and explanations: A structural-model approach. Part I: Causes, British J. Philos. Sci. 56 (2005) 889–911.

[50] S. Psillos, Causation and Explanation, MPG Books Group, 2002.

[51] D. Hume, A Treatise of Human Nature, John Noon, London, 1739.

[52] D. Lewis, Causation, J. Phil. 70 (1973) 113–126.

[53] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black-box: Automated decisions and the GDPR, Harv. J. Law & Technol. 31 (2018).

[54] 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.

[55] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, 2020, arxiv:2010.10596.

[56] 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.

[57] A. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: definitions, formulations, solutions, and prospects, 2021, arxiv:2010.04050.

[58] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, 2020, arXiv:2009.11698.

[59] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Leanpub, 2018.

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

[61] B. Kim, J. Park, J. Suh, Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information, Decis. Support Syst. 134 (2020) 113302.

[62] 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.

[63] 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.

[64] 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.

[65] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision – ECCV 2014, 2014, pp. 818–833.

[66] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, The pragmatic turn in explainable artificial intelligence (XAI), Nature Commun. 10 (2019) 1096.

[67] H.F. Tan, K. Song, M. Udell, Y. Sun, Y. Zhang, Why should you trust my interpretation? Understanding uncertainty in LIME predictions, 2019, arxiv:1904.12991.

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

[69] B. Osbert, K. Carolyn, B. Hamsa, Interpretability via model extraction, 2017, arxiv:1705.08504.

[70] J. Thiagarajan, B. Kailkhura, P. Sattigeri, K.N. Ramamurthy, TreeView: Peeking into deep neural networks via feature-space partitioning, Nature Commun. (2019).

[71] 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, 2020.

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

[73] 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.

[74] L.S. Shapley, A Value for n-Person Games, Rand Coporation, 1952, p. 15.

[75] E. Strumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst. 41 (3) (2013) 647–665.

[76] A. Shrikumar, P. Greenside, A. Kundaje, learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3145–3153.

[77] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 2522–5839.

[78] A.C. Miller Janny Ariza-Garzón, M.-J. Segovia-Vargas, Explainability of a machine learning granting scoring model in peer-to-peer lending, in: Proceedings of IEEE Access, 2020.

[79] 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, Accid. Anal. Prev. 136 (2020) 105405.

[80] J. Wang, J. Wiens, S. Lundberg, Shapley flow: A graph-based approach to interpreting model predictions, in: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, 2021, arXiv:2010.14592.

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

[82] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: The system causability scale (SCS), KI - Künstliche Intell. 34 (2020) 193–198.

[83] R. Byrne, Cognitive processes in counterfactual thinking about what might have been, Psychol. Learn. Motiv. Adv. Res. Theory 37 (1997) 105–154.

[84] D. Wesberg, A. Gopnik, Pretense, counterfactuals, and Bayesian causal models: Why what is not real really matters, Cogn. Sci. 37 (2013) 1368–1381.

[85] L.M. Pereira, A.B. Lopes, Cognitive prerequisites: The special case of counterfactual reasoning, Mach. Ethics Stud. Appl. Phil. Epistemol. Rational Ethics 53 (2020).

[86] J. Paik, Y. Zhang, P. Pirolli, Counterfactual reasoning as a key for explaining adaptive behavior in a changing environment, Biol. Inspir. Cogn. Archit. 10 (2014) 24–29.

[87] 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, Nat. Mach. Intell. 2 (2020) 369–375.

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

[89] 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.

[90] 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 Comput. Intell. Magazine 14 (1) (2019).

[91] D. Lewis, Counterfactuals, Blackwell, Oxford, 1973.

[92] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, Lecture Notes in Comput. Sci. (2020) 448–469.

[93] 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.

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

[95] D. Martens, F. Provost, Explaining data-driven document classifications, MIS Q. 38 (1) (2014).

[96] 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.

[97] M. Pawelczyk, K. Broelemann, G. Kasneci, On counterfactual explanations under predictive multiplicity, in: Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, 2020.

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

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

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

[101] 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.

[102] 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.

[103] 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.

[104] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, Unjustified classification regions and counterfactual explanations in machine learning, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, 2019.

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

[106] S. Sharma, J. Henderson, J. Ghosh, CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, 2019, arxiv:1905.07857.

[107] A. White, A. d'Avila Garcez, Measurable counterfactual local explanations for any classifier, 2019, arxiv:1908.03020.

[108] 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, Adv. Data Anal. Classif. 1 (1) (2020).

[109] S. Rathi, Generating counterfactual and contrastive explanations using SHAP, 2019, arXiv:1906.09293.

[110] A. Ghazimatin, O. Balalau, R. Saha Roy, G. Weikum, PRINCE: provider-side interpretability with counterfactual explanations in recommender systems, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 196–204.

[111] M. Pawelczyk, K. Broelemann, G. Kasneci, Learning model-agnostic counterfactual explanations for tabular data, in: Proceedings of the World Wide Web Conference 2020, 2020.

[112] M. Pawelczyk, J. Haug, K. Broelemann, G. Kasneci, Towards user empowerment, in: Proceedings of the Thirty-Third Annual Conference on Neural Information Processing Systems, Workshop on Human-Centric Machine Learning, 2019.

[113] A. Lucic, H. Haned, M. de Rijke, Why does my model fail? contrastive local explanations for retail forecasting, in: FAT* '20: Conference on Fairness, Accountability, and Transparency, 2020.

[114] R. Guidotti, A. Monreale, S. Matwin, D. Pedreschi, Black box explanation by learning image exemplars in the latent feature space, in: Proccedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2020.

[115] M. Downs, J.L. Chu, Y. Yacoby, F. Doshi-Velez, W. Pan, Cruds: Counterfactual recourse using disentangled subspaces, ICML WHI 2020 (2020) 1–23.

[116] A. Karimi, B.J. von Kügelgen, B. Schölkopf, I. Valera, Algorithmic recourse under imperfect causal knowledge: a probabilistic approach, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, 2020.

[117] K. Rawal, Himabindu, Beyond individualized recourse: Interpretable and interactive summaries of actionable recourse, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020.

[118] S. Barocas, A.D. Selbst, M. Raghavan, The hidden assumptions behind counterfactual explanations and principal reasons, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.
[119] M. Zheng, J.K. Marsh, J.V. Nickerson, S. Kleinberg, How causal information affects decisions, Cogn. Res. Princ. Implic. 5 (2020).
[120] ALIBI, 2019, URL: https://github.com/SeldonIO/alibi.
[121] FACE, 2020, URL: https://github.com/sharmapulkit.
[122] Growing spheres, 2019, URL: https://github.com/thibaultlaugel/truce.
[123] DICE, 2020, URL: https://github.com/interpretml/DiCE.
[124] PRINCE, 2019, URL: https://github.com/azinmatin/prince/.
[125] C-CHVAE, 2020, URL: https://github.com/MartinPawel/c-chvae.
[126] ABELE, 2020, URL: https://github.com/riccotti/ABELE.
[127] RECOURSE, 2020, URL: https://github.com/amirhk/recourse.
[128] MC-BRP, 2019, URL: https://github.com/a-lucic/mc-brp.
[129] MACE, 2019, URL: https://github.com/amirhk/mace.
[130] COHERENT Counterfactuals, 2019, URL: https://bitbucket.org/ChrisRussell/diverse-coherent-explanations/src/master/.
[131] MOCE, 2020, URL: https://github.com/susanne-207/moc.
[132] CERTIFAI, 2020, URL: https://github.com/Ighina/CERTIFAI.
[133] LORE, 2018, URL: https://github.com/riccotti/LORE.
[134] Y. Ramon, D. Martens, F. Provost, T. Evgeniou, Counterfactual explanation algorithms for behavioral and textual data, 2019, arXiv:1912.01819.
[135] LIME counterfactual, 2020, URL: https://github.com/yramon/LimeCounterfactual.
[136] SEDcounterfactual, 2020, URL: https://github.com/yramon/edc.
[137] CLEAR, 2020, URL: https://github.com/ClearExplanationsAI/CLEAR.
[138] SHAP counterfactual, 2020, URL: https://github.com/yramon/ShapCounterfactual.
[139] 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.
[140] A. Holzinger, From machine learning to explainable AI, in: Proceedings of the 2018 World Symposium on Digital Intelligence for Systems and Machines, 2018.
[141] G. Xu, T.D. Duong, Q. Li, S. Liu, X. Wang, Causality learning: A new perspective for interpretable machine learning, 2020, arXiv:2006.16789.
[142] 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, Appl. Intell. 49 (2019) 2401–2414.
[143] A. Holzinger, Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning, IEEE Intell. Inf. Bull. 15 (2014) 6–14.
[144] Q. Zhao, T. Hastie, Causal interpretations of black-box models, J. Bus. Econom. Statist. (2019) 1–10.
[145] O. Peters, The ergodicity problem in economics, Nat. Phys. 15 (2019) 1216–1221.
[146] J. Rehse, N. Mehdiyev, P. Fettke, Towards explainable process predictions for industry 4.0 in the DFKI-smart-lego-factory, Künstliche Intell. 33 (2) (2019) 181–187.
[147] FICO, 2017, URL: https://community.fico.com/s/explainable-machine-learning-challenge.
[148] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Trans. Vis. Comput. Graphics (2019) 1.
[149] O. Gomez, S. Holter, J. Yuan, E. Bertini, ViCE: Visual counterfactual explanations for machine learning models, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 531–535.
[150] F. Cheng, Y. Ming, H. Qu, DECE: Decision explorer with counterfactual explanations for machine learning models, in: Proceedings of the IEEE VIS 2020, 2020.
[151] D. Shin, The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI, Int. J. Hum.-Comput. Stud. 146 (2021) 102551.
[152] R. Confalonieri, L. Coba, B. Wagner, T.R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Min. Knowl. Discov. 11 (2021) e1391.
[153] T. Gerstenberg, M.F. Peterson, N.D. Goodman, D.A. Lagnado, J.B. Tenenbaum, Eye-tracking causality, Psychol. Sci. 28 (2017) 1731–1744.
[154] E. Goldvarg, P. Johnson-Laird, Naive causality: a mental model theory of causal meaning and reasoning, Cogn. Sci. 25 (4) (2001) 565–610.
[155] A. Holzinger, Interactive machine learning for health informatics: When do we need the human-in-the-loop? Brain Informatics 3 (2016) 119–131.
[156] A. Dieng, Y. Liu, S. Roy, C. Rudin, A. Volfovsky, Interpretable almost-exact matching for causal inference, in: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, AISTATS, in: Proceedings of Machine Learning Research, vol. 89, 2019, pp. 2445–2453.
[157] T. Wang, M. Morucci, M.U. Awan, Y. Liu, S. Roy, C. Rudin, A. Volfovsky, FLAME: A fast large-scale almost matching exactly approach to causal inference, J. Mach. Learn. Res. 22 (2021) 1–41.
[158] M.U. Awan, M. Morucci, V. Orlandi, S. Roy, C. Rudin, A. Volfovsky, Almost-matching-exactly for treatment effect estimation under network interference, in: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, AISTATS, Vol. 108, PMLR, 2020, pp. 3252–3262.
[159] R.N. Yale, Measuring narrative believability: Development and validation of the narrative believability scale (NBS-12), J. Commun. 63 (2013) 578–599.
[160] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, P. Bruza, LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models, Decis. Support Syst. (2021) 113561.
[161] M. Velmurugan, C. Ouyang, C. Moreira, R. Sindhgatta, Evaluating explainable methods for predictive process analytics: a functionally-grounded approach, in: Proceedings of the 33rd International Conference on Advanced Information Systems Engineering Forum, 2020.
[162] 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.
[163] M.N. Hoque, K. Mueller, Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making, 2021, arXiv:2101.00633.
[164] S.T. 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, pp. 477–487.