Counterfactual Explanations for Machine Learning: A Review

Sahil Verma
Arthur AI, University of Washington
Washington D.C., USA
vsahil@cs.washington.edu

John Dickerson
Arthur AI
Washington D.C., USA
john@arthur.ai

Keegan Hines
Arthur AI
Washington D.C., USA
keegan@arthur.ai

Abstract

Machine learning plays a role in many deployed decision systems, often in ways that are difficult or impossible to understand by human stakeholders. Explaining, in a human-understandable way, the relationship between the input and output of machine learning models is essential to the development of trustworthy machine-learning-based systems. A burgeoning body of research seeks to define the goals and methods of *explainability* in machine learning. In this paper, we seek to review and categorize research on *counterfactual explanations*, a specific class of explanation that provides a link between what could have happened had input to a model been changed in a particular way. Modern approaches to counterfactual explainability in machine learning draw connections to the established legal doctrine in many countries, making them appealing to fielded systems in high-impact areas such as finance and healthcare. Thus, we design a rubric with desirable properties of counterfactual explanation algorithms and comprehensively evaluate all currently-proposed algorithms against that rubric. Our rubric provides easy comparison and comprehension of the advantages and disadvantages of different approaches and serves as an introduction to major research themes in this field. We also identify gaps and discuss promising research directions in the space of counterfactual explainability.

1 Introduction

Machine learning is increasingly accepted as an effective tool to enable large-scale automation in many domains. In lieu of hand-designed rules, algorithms are able to learn from data to discover patterns and support decisions. Those decisions can, and do, directly or indirectly impact humans; high-profile cases include applications in credit lending [99], talent sourcing [97], parole [102], and medical treatment [46]. The nascent Fairness, Accountability, Transparency, and Ethics (FATE) in machine learning community has emerged as a multi-disciplinary group of researchers and industry practitioners interested in developing techniques to detect bias in machine learning models, develop algorithms to counteract that bias, generate human-comprehensible explanations for the machine decisions, hold organizations responsible for unfair decisions, etc.

Human-understandable explanations for machine-produced decisions are advantageous in several ways. For example, focusing on a use case of applicants applying for loans, the benefits would include:

- An explanation can be beneficial to the applicant whose life is impacted by the decision. For example, it helps an applicant understand which of their attributes were strong drivers in determining a decision.

NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA).

- Further, it can help an applicant challenge a decision if they feel an unfair treatment has been meted, e.g., if one's race was crucial in determining the outcome. This can also be useful for organizations to check for bias in their algorithms.
- In some instances, an explanation provides the applicant with feedback that they can act upon to receive the desired outcome at a future time.
- Explanations can help the machine learning model developers identify, detect, and fix bugs and other performance issues.
- Explanations help in adhering to laws surrounding machine-produced decisions, e.g., GDPR [10].

Explainability in machine learning is broadly about using inherently interpretable and transparent models or generating post-hoc explanations for opaque models. Examples of the former include linear/logistic regression, decision trees, rule sets, etc. Examples of the latter include random forest, support vector machines (SVMs), and neural networks.

Post-hoc explanation approaches can either be model-specific or model-agnostic. Explanations by feature importance and model simplification are two broad kinds of model-specific approaches. Model-agnostic approaches can be categorized into visual explanations, local explanations, feature importance, and model simplification.

Feature importance finds the most influential features in contributing to the model's overall accuracy or for a particular decision, e.g., SHAP [80], QII [27]. Model simplification finds an interpretable model that imitates the opaque model closely. Dependency plots are a popular kind of visual explanation, e.g., Partial Dependence Plots [51], Accumulated Local Effects Plot [14], Individual Conditional Expectation [53]. They plot the change in the model's prediction as a feature, or multiple features are changed. Local explanations differ from other explanation methods because they only explain a single prediction. Local explanations can be further categorized into approximation and example-based approaches. Approximation approaches sample new datapoints in the vicinity of the datapoint whose prediction from the model needs to be explained (hereafter called the explainee datapoint), and then fit a linear model (e.g., LIME [92]) or extracts a rule set from them (e.g., Anchors [93]). Example-based approaches seek to find datapoints in the vicinity of the explainee datapoint. They either offer explanations in the form of datapoints that have the same prediction as the explainee datapoint or the datapoints whose prediction is different from the explainee datapoint. Note that the latter kind of datapoints are still close to the explainee datapoint and are termed as "counterfactual explanations".

Recall the use case of applicants applying for a loan. For an individual whose loan request has been denied, counterfactual explanations provide them feedback, to help them make changes to their features in order to transition to the desirable side of the decision boundary, i.e., get the loan. Such feedback is termed *actionable*. Unlike several other explainability techniques, counterfactual explanations do not explicitly answer the "why" part of a decision; instead, they provide suggestions in order to achieve the desired outcome. Counterfactual explanations are also applicable to black-box models (only the `predict` function of the model is accessible), and therefore place no restrictions on model complexity and do not require model disclosure. They also do not necessarily approximate the underlying model, producing accurate feedback. Owing to their intuitive nature, counterfactual explanations are also amenable to legal frameworks (see appendix C).

In this work, we collect, review and categorize 39 recent papers that propose algorithms to generate counterfactual explanations for machine learning models. Many of these methods have focused on datasets that are either tabular or image-based. We describe our methodology for collecting papers for this survey in appendix B. We describe recent research themes in this field and categorize the collected papers among a fixed set of desiderata for effective counterfactual explanations (see table 1).

The contributions of this review paper are:

1. We examine a set of 39 recent papers on the same set of parameters to allow for an easy comparison of the techniques these papers propose and the assumptions they work under.
2. The categorization of the papers achieved by this evaluation helps a researcher or a developer choose the most appropriate algorithm given the set of assumptions they have and the speed and quality of the generation they want to achieve.

3. Comprehensive and lucid introduction for beginners in the area of counterfactual explanations for machine learning.

2 Background

This section gives the background about the social implications of machine learning, explainability research in machine learning, and some prior studies about counterfactual explanations.

2.1 Social Implications of Machine Learning

Establishing fairness and making an automated tool's decision explainable are two broad ways in which we can ensure equitable social implication of machine learning. Fairness research aims at developing algorithms that can ensure that the decisions produced by the system are not biased against a particular demographic group of individuals, which are defined with respect to sensitive features, for e.g., race, sex, religion. Anti-discrimination laws make it illegal to use the sensitive features as the basis of any decision (see Appendix C). Biased decisions can also attract widespread criticism and are therefore important to avoid [55, 69]. Fairness has been captured in several notions, based on a demographic grouping or individual capacity. Verma and Rubin [109] have enumerated, and intuitively explained many fairness definitions using a unifying dataset. Dunkelau and Leuschel [45] provide an extensive overview of the major categorization of research efforts in ensuring fair machine learning and enlists important works in all categories. Explainable machine learning has also seen interest from other communities, specifically healthcare [103], having huge social implications. Several works have summarized and reviewed other research in explainable machine learning [11, 22, 58].

2.2 Explainability in Machine Learning

This section gives some concrete examples that emphasize the importance of explainability and give further details of the research in this area. In a real example, the military trained a classifier to distinguish enemy tanks from friendly tanks. Although the classifier performed well on the training and test dataset, its performance was abysmal on the battlefield. Later, it was found that the photos of friendly tanks were taken on sunny days, while for enemy tanks, photos clicked only on overcast days were available [58]. The classifier found it much easier to use the difference between the background as the distinguishing feature. In a similar case, a husky was classified as a wolf because of the presence of snow in the background, which the classifier had learned as a feature associated with wolves [92]. The use of an explainability technique helped discover these issues.

The explainability problem can be divided into model explanation and outcome explanation problems [58].

Model explanation searches for an interpretable and transparent global explanation of the original model. Various papers have developed techniques to explain neural networks and tree ensembles using single decision tree [25, 70, 34] and rule sets [28, 13]. Some approaches are model-agnostic, e.g. Golden Eye, PALM [59, 71, 116].

Outcome explanation needs to provide an explanation for a specific prediction from the model. This explanation need not be a global explanation or explain the internal logic of the model. Model-specific approaches for deep neural networks (CAM, Grad-CAM [115, 96]), and model agnostic approaches (LIME, MES [92, 106]) have been proposed. These are either feature attribution or model simplification methods. Example based approaches are another kind of explainability techniques used to explain a particular outcome. In this work, we focus on counterfactual explanations which is an example-based approach.

By definition, counterfactual explanations are applicable to supervised machine learning setup where the desired prediction has not been obtained for a datapoint. The majority of research in this area has applied counterfactual explanations to classification setting, which consists of several labeled datapoints that are given as input to the model, and the goal is to learn a function mapping from the input datapoints (with say m features) to labels. In classification, the labels are discrete values. $\mathcal{X}^m$ is used to denote the input space of the features, and $\mathcal{Y}$ is used to denote the output space of the

labels. The learned function is the mapping $f : \mathcal{X}^m \rightarrow \mathcal{Y}$, which is used to predict labels for unseen datapoints in the future.

2.3 History of Counterfactual Explanations

Counterfactual explanations have a long history in other fields like philosophy, psychology, and the social sciences. Philosophers like David Lewis, published articles on the ideas of counterfactuals back in 1973 [78]. Woodward [114] said that a satisfactory explanation must follow patterns of counterfactual dependence. Psychologists have demonstrated that counterfactuals elicit causal reasoning in humans [20, 21, 62]. Philosophers have also validated the concept of causal thinking due to counterfactuals [17, 114].

There have been studies which compared the likeability of counterfactual explanations with other explanation approaches. Binns et al. [18] and Dodge et al. [32] performed user-studies which showed that users prefer counterfactual explanations over case-based reasoning, which is another example-based approach. Fernández-Loría et al. [48] give examples where counterfactual explanations are better than feature importance methods.

3 Counterfactual Explanations

This section illustrates counterfactual explanations by giving an example and then outlines the major aspects of the problem.

3.1 An Example

Suppose Alice walks into a bank and seeks a home mortgage loan. The decision is impacted in large part by a machine learning classifier which considers Alice's feature vector of {*Income*, *CreditScore*, *Education*, *Age*}. Unfortunately, Alice is denied for the loan she seeks and is left wondering (1) why was the loan denied? and (2) what can she do differently so that the loan will be approved in the future? The former question might be answered with explanations like: "CreditScore was too low", and is similar to the majority of traditional explainability methods. The latter question forms the basis of a *counterfactual explanation*: what small changes could be made to Alice's feature vector in order to end up on the other side of the classifier's decision boundary. Let's suppose the bank provides Alice with exactly this advice (through the form of a counterfactual explanation) of what she might change in order to be approved next time. A possible counterfactual recommended by the system might be to increase her *Income* by $\$10$K or get a new master's degree or a combination of both. The answer to the former question does not tell Alice what action to take, while the counterfactual explanation explicitly helps her. Figure 1 illustrates how the datapoint representing an individual, which originally got classified in the negative class, can take two paths to cross the decision boundary into the positive class region.

The assumption in a counterfactual explanation is that the underlying classifier would not change when the applicant applies in the future. And if the assumption holds, the counterfactual guarantees the desired outcome in the future time.

3.2 Desiderata and Major Themes of Research

The previous example alludes to many of the desirable properties of an effective counterfactual explanation. For Alice, the counterfactual should quantify a relatively small change, which will lead to the desired alternative outcome. Alice might need to increase her income by $\$10$K to get approved for a loan, and even though an increase of $\$50$K would do the job, it is most pragmatic for her if she can make the smallest possible change. Additionally, Alice might care about a simpler explanation - it is easier for her to focus on changing a few things (such as only *Income*) instead of trying to change many features. Alice certainly also cares that the counterfactual she receives is giving her advice, which is realistic and actionable. It would be of little use if the recommendation were to decrease her age by ten years.

These desiderata, among others, have set the stage for recent developments in the field of counterfactual explainability. As we describe in this section, major themes of research have sought to incorporate increasingly complex constraints on counterfactuals, all in the spirit of ensuring the result-

ing explanation is truly actionable and useful. Development in this field has focused on addressing these desiderata in a way that is generalizable across algorithms and is computationally efficient.

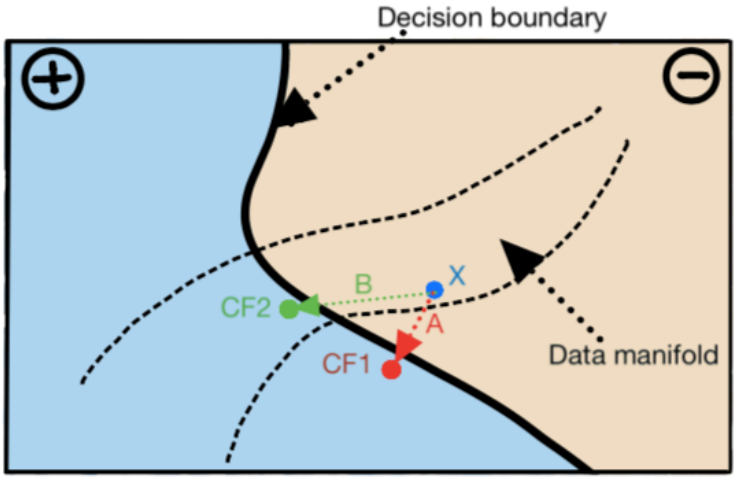

Figure 1: Two possible paths for a datapoint (shown in blue), originally classified in the negative class, to cross the decision boundary. The end points of both the paths (shown in red and green) are valid counterfactuals for the original point. Note that the red path is the shortest, whereas the green path adheres closely to the manifold of the training data, but is longer.

1. *Validity*: Wachter et al. [111] first proposed counterfactual explanations in 2017. They posed counterfactual explanation as an optimization problem. Equation (1) states the optimization objective, which is to minimize the distance between the counterfactual (x') and the original datapoint (x) subject to the contraint that the output of the classifier on the counterfactual is the desired label ($y' \in \mathcal{Y}$). Converting the objective into a differentiable, unconstrained form yields two terms (see Equation (2)). The first term encourages the output of the classifier on the counterfactual to be close to the desired class and the second term forces the counterfactual to be close the original datapoint. A metric d is used to measure the distance between two datapoints $x, x' \in \mathcal{X}$, which can be the L1/L2 distance, or quadratic distance, or distance functions which take as input the CDF of the features [107]. Thus, this original definition already emphasized that an effective conterfactual must be *small change* relative to the starting point.

$$\arg\min_{x'} d(x, x') \text{ subject to } f(x') = y' \tag{1}$$

$$\arg\min_{x'} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{2}$$

   A counterfactual which indeed is classified in the desired class is a valid counterfactual. As illustrated in fig. 1, the points shown in red and green are valid counterfactuals, as they are indeed in the positive class region, and the distance to the red counterfactual is smaller than the distance to the green counterfactual.

2. *Actionability*: An important consideration while making recommendation is about which features are mutable (for e.g. income, age) and which aren't (for e.g. race, country of origin). A recommended counterfactual should never change the immutable features. In fact, if change to a legally sensitive feature produces a change in prediction, it shows inherent bias in the model. Several papers have also mentioned that an applicant might have a preference order amongst the mutable features (which can also be hidden.) The optimization problem is modified to take this into account. We might call the set of actionable features $\mathcal{A}$, and update our loss function to be,

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') \tag{3}$$

3. *Sparsity*: There can be a trade-off between the number of features changed and the total amount of change made to obtain the counterfactual. A counterfactual ideally should change smaller number of features in order to be most effective. It has been argued that people find it easier to

understand shorter explanations [84], making sparsity an important consideration. We update our loss function to include a penalty function which encourages sparsity in the difference between the modified and the original datapoint, $g(x' - x)$, e.g. L0/L1 norm.

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') + g(x' - x) \tag{4}$$

4. *Data Manifold closeness*: It would be hard to trust a counterfactual if it resulted in a combination of features which were utterly unlike any observations the classifier has seen before. In this sense, the counterfactual would be "unrealistic" and not easy to realize. Therefore, it is desirable that a generated counterfactual is realistic in the sense that it is near the training data and adheres to observed correlations among the features. Many papers have proposed various ways of quantifying this. We might update our loss function to include a penalty for adhering to the data manifold defined by the training set $\mathcal{X}$, denoted by $l(x'; \mathcal{X})$

$$\arg\min_{x' \in \mathcal{A}} \max_{\lambda} \lambda(f(x') - y')^2 + d(x, x') + g(x' - x) + l(x'; \mathcal{X}) \tag{5}$$

   In fig. 1, the region between the dashed lines shows the data manifold. For the blue datapoint, there are two possible paths to cross the decision boundary. The shorter, red path takes it to a counterfactual that is outside the data manifold, whereas a bit longer, green path takes it to a counterfactual that follows the data manifold. Adding the data manifold loss term, encourages the algorithm to choose the green path over red path, even if it is slightly longer.

5. *Causality*: Features in a dataset are rarely independent, therefore changing one feature in the real world affects other features. For example, getting a new educational degree necessitates increasing to the age of the individual by at least some amount. In order to be realistic and actionable, a counterfactual should maintain any known causal relations between features. Generally, our loss function now accounts for (1) counterfactual validity, (2) sparsity in feature vector (and/or actionability of features), (3) similarity to the training data, (4) causal relations.

Following research themes are not added as terms in the optimization objective; they are properties of the counterfactual algorithm.

6. *Amortized inference*: Generating a counterfactual is expensive, which involves solving an optimization process for each datapoint. Mahajan et al. [82] focused on "amortized inference" using generative techniques. Thus learning to predict the counterfactual allows the algorithm to quickly compute a counterfactual (or several) for any new input x, without requiring to solve an optimization problem.
7. *Alternative methods*: Finally, several papers solve the counterfactual generation problem using linear programming, mixed-integer programming, or SMT solvers. These approaches give guarantees and optimize fast, but are limited to classifiers with linear (or piece-wise linear) structure.
8. **[[Add about model-agnosticity and black-boxness]]**.

3.3 Relationship to other related terms

Out of the papers collected, different terminology often captures the basic idea of counterfactual explanations, although subtle differences exist between the terms. Several terms worth noting include:

- *Recourse* - Ustun et al. [107] point out that counterfactuals do not take into account the actionability of the prescribed changes, which recourse does. The difference they point out was from the original work of Wachter et al. [111]. Recent papers in counterfactual generation take actionability and feasibility of the prescribed changes, and therefore the difference with recourse has blurred.
- *Inverse classification* - Inverse classification aims to perturb an input in a meaningful way in order to classify it into its desired class [12, 72]. Such an approach prescribes the actions to be taken in order to get the desired classification. Therefore inverse classification has the same goals as counterfactual explanations.
- *Contrastive explanation* - Contrastive explanations generate explanations of the form "an input x is classified as y because features $f_1, f_2, \ldots, f_k$ are present and $f_n, \ldots, f_r$ are absent". The features which are minimally sufficient for a classification are called pertinent positives, and the

features whose absence is necessary for the final classification are termed as pertinent negatives. To generate both pertinent positives and pertinent negatives, one needs to solve the optimization problem to find the minimum perturbations needed to maintain the same class label or change it, respectively. Therefore contrastive explanations (specifically pertinent negatives) are related to counterfactual explanations.

- *Adversarial learning* - Adversarial learning is a closely-related field, but the terms are not interchangeable. Adversarial learning aims to generate the least amount of change in a given input in order to classify it differently, often with the goal of far-exceeding the decision boundary and resulting in a highly-confident misclassification. While the optimization problem is similar to the one posed in counterfactual-generation, the desiderata are different. For example, in adversarial learning (often applied to images), the goal is an imperceptible change in the input image. This is often at odds with the counterfactual's goal of sparsity and parsimony (though single-pixel attacks are an exception). Further, notions of data manifold and actionability/causality are rarely considerations in adversarial learning.

4 Assessment of the approaches on counterfactual properties

For easy comprehension and comparison, we identify several properties that are important for a counterfactual generation algorithm to be assessed on. For all the collected papers which propose an algorithm to generate counterfactual explanation, we assess the algorithm they propose against these properties. The results are presented in table 1. For papers that do not propose new algorithms, but discuss related aspects of counterfactual explanations are mentioned in section 4.2. The methodology which we used to collect the papers is given in appendix B.

4.1 Properties of counterfactual algorithms

This section expounds on the key properties of a counterfactual explanation generation algorithm. The properties form the columns of table 1.

1. *Model access* - The counterfactual generation algorithms require different levels of access to the underlying model for which they generate counterfactuals. We identify three distinct access levels - access to complete model internals, access to gradients, and access to only the prediction function (*black-box*). Access to the complete model internals are required when the algorithm uses a solver based method like, mixed integer programming [95, 107, 64, 65, 63] or if they operate on decision trees [104, 47, 79] which requires access to all internal nodes of the tree. A majority of the methods use a gradient-based algorithm to solve the optimization objective, modifying the loss function proposed by Wachter et al. [111], but this is restricted to differentiable models only. Black-box approaches use gradient-free optimization algorithms such as Nelder-Mead [56], growing spheres [74], FISTA [30, 108], or genetic algorithms [72, 98, 26] to solve the optimization problem. Finally, some approaches do not cast the goal into an optimization problem and solve it using heuristics [57, 91, 113, 67]. Poyiadzi et al. [89] propose FACE, which uses Dijkstra's algorithm [31] to find the shortest path between existing training datapoints to find counterfactual for a given input. Hence, this method does not generate new datapoints.
2. *Model agnostic* - This column describes the domain of models a given algorithm can operate on. As examples, gradient based algorithms can only handle differentiable models, the algorithms based on solvers require linear or piece-wise linear models [95, 107, 64, 65, 63], some algorithms are model-specific and only work for those models like tree ensembles [104, 63, 47, 79]. Black-box methods have no restriction on the underlying model and are therefore model-agnostic.
3. *Optimization amortization* - Among the collected papers, the proposed algorithm mostly returned a single counterfactual for a given input datapoint. Therefore these algorithms require to solve the optimization problem for each counterfactual that was generated, that too, for every input datapoint. A smaller number of the methods are able to generate multiple counterfactuals (generally diverse by some metric of diversity) for a single input datapoint, therefore they require to be run once per input to get several counterfactuals [57, 95, 98, 85, 82, 64, 26, 47]. Mahajan et al. [82]'s approach learns the mapping of datapoints to counterfactuals using a variational auto-encoder (VAE) [33]. Therefore, once the VAE is trained, it can generate multiple counterfactuals for all input datapoints, without solving the optimization problem separately, and is thus very fast. We report two aspects of optimization amortization in the table.

- *Amortized Inference* - This column is marked `Yes` if the algorithm can generate counterfactuals for multiple input datapoints without optimizing separately for them, otherwise it is marked `No`.
- *Multiple counterfactual (CF)* - This column is marked `Yes` if the algorithm can generate multiple counterfactual for a single input datapoint, otherwise it is marked `No`.

4. *Counterfactual (CF) attributes* - These columns evaluate algorithms on sparsity, data manifold adherence, and causality.

   Among the collected papers, methods using solvers explicitly constrain sparsity [107, 64], black-box methods constrain L0 norm of counterfactual and the input datapoint [74, 26]. Gradient based methods typically use the L1 norm of counterfactual and the input datapoint. Some of the methods change only a fixed number of features [113, 67], change features iteratively [76], or flip the minimum possible split nodes in the decision tree [57] to induce sparsity. Some methods also induce sparsity post-hoc [74, 85]. This is done by sorting the features in ascending order of relative change and greedily restoring their values to match the values in the input datapoint until the prediction for the CF is still different from the input datapoint.

   Adherence to the data manifold has been addressed using several different approaches, like training VAEs on the data distribution [29, 61, 108, 82], constraining the distance of a counterfactual from the k nearest training datapoints [26, 63], directly sampling points from the latent space of a VAE trained on the data, and then passing the points through the decoder [87], mapping back to the data domain [76], using a combination of existing datapoints [67], or by simply not generating any new datapoint [89].

   The relation between different features is represented by a directed graph between them, which is termed as a causal graph [88]. Out of the papers that have addressed this concern, most require access to the complete causal graph [65, 66] (which is rarely available in the real world), while Mahajan et al. [82] can work with partial causal graphs. These three properties are reported in the table.

   - *Sparsity* - This column is marked `No` if the algorithm does not consider sparsity, else it specifies the sparsity constraint.
   - *Data manifold* - This column is marked `Yes` if the algorithm forces the generated counterfactuals to be close to the data manifold by some mechanism. Otherwise it is marked `No`.
   - *Causal relation* - This column is marked `Yes` if the algorithm considers the causal relations between features when generating counterfactuals. Otherwise it is marked `No`.

5. *Counterfactual (CF) optimization (opt.) problem attributes* - These are a few attributes of the optimization problem.

   Out of the papers that consider feature actionability, most classify the features into immutable and mutable types. Karimi et al. [65] and Lash et al. [72] categorize the features into immutable, mutable, and actionable types. Actionable features are a subset of mutable features. They point out that certain features are mutable but not directly actionable by the individual, e.g., *CreditScore* cannot be directly changed; it changes as an effect of changes in other features like income, credit amount. Mahajan et al. [82] uses an oracle to learn the user preferences for changing features (among mutable features) and can learn hidden preferences as well.

   Most tabular datasets have both continuous and categorical features. Performing arithmetic over continuous feature is natural, but handling categorical variables in gradient-based algorithms can be complicated. Some of the algorithms cannot handle categorical variables and filter them out [74, 79]. Wachter et al. [111] proposed clamping all categorical features to each of their values, thus spawning many processes (one for each value of each categorical feature), leading to scalability issues. Some approaches convert categorical features to one-hot encoding and then treat them as numerical features. In this case, maintaining one-hotness can be challenging. Some use a different distance function for categorical features, which is generally an indicator function (1 if a different value, else 0). Genetic algorithms and SMT solvers can naturally handle categorical features. We report these properties in the table.

   - *Feature preference* - This column is marked `Yes` if the algorithm considers feature actionability, otherwise marked `No`.
   - *Categorical distance function* - This column is marked – if the algorithm does not use a separate distance function for categorical variables, else it specifies the distance function.

Table 1: Assessment of the collected papers on the key properties, which are important for readily comparing and comprehending the differences and limitations of different counterfactual algorithms. Papers are sorted chronologically. Details about the full table is given in appendix A.

	Assumptions		Optimization amortization		CF attributes			CF opt. problem attributes	
Paper	Model access	Model domain	Amortized Inference	Multiple CF	Sparsity	Data manifold	Causal relation	Feature preference	Categorical dist. func
[72]	Black-box	Agnostic	No	No	Changes iteratively	No	No	Yes	-
[111]	Gradients	Differentiable	No	No	L1	No	No	No	-
[104]	Complete	Tree ensemble	No	No	No	No	No	No	-
[74]	Black-box	Agnostic	No	No	L0 and post-hoc	No	No	No	-
[57]	Black-box	Agnostic	No	Yes	Flips min. split nodes	No	No	No	Indicator
[29]	Gradients	Differentiable	No	No	L1	Yes	No	No	-
[56]	Black-box	Agnostic	No	No	No	No	No	No[1]	-
[95]	Complete	Linear	No	Yes	L1	No	No	No	N.A.[2]
[107]	Complete	Linear	No	No	Hard constraint	No	No	Yes	-
[98]	Black-box	Agnostic	No	Yes	No	No	No	Yes	Indicator
[30]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	-
[91]	Black-box	Agnostic	No	No	No	No	No	No	-
[61]	Gradients	Differentiable	No	No	No	Yes	No	No	-
[90]	Gradients	Differentiable	No	No	No	No	No	No	-
[113]	Black-box	Agnostic	No	No	Changes one feature	No	No	No	-
[85]	Gradients	Differentiable	No	Yes	L1 and post-hoc	No	No	No	Indicator
[89]	Black-box	Agnostic	No	No	No	Yes[3]	No	No	-
[108]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	Embedding
[82]	Gradients	Differentiable	Yes	Yes	No	Yes	Yes	Yes	-
[64]	Complete	Linear	No	Yes	Hard constraint	No	No	Yes	Indicator
[87]	Gradients	Differentiable	No	No	No	Yes	No	Yes	N.A.[4]
[67]	Black-box	Agnostic	No	No	Yes	Yes	No	No	-
[65]	Complete	Linear and causal graph	No	No	L1	No	Yes	Yes	-
[66]	Gradients	Differentiable	No	No	No	No	Yes	Yes	-
[76]	Gradients	Differentiable	No	No	Changes iteratively	Yes	No	No[5]	-
[26]	Black-box	Agnostic	No	Yes	L0	Yes	No	Yes	Indicator
[63]	Complete	Linear and tree ensemble	No	No	No	Yes	No	Yes	-
[47]	Complete	Random Forest	No	Yes	L1	No	No	No	-
[79]	Complete	Tree ensemble	No	No	L1	No	No	No	-

4.2 Other works

There exist papers which do not propose novel algorithms to generate counterfactuals, but explore other aspects about it.

Sokol and Flach [101] list several desirable properties of counterfactuals inspired from Miller [84] and state how the method of flipping logical conditions in a decision tree satisfies most of them. Fernández-Loría et al. [48] point out at the insufficiency of feature importance methods for explaining a model's predictions, and substantiate it with a synthetic example. They generate counterfactuals by removing features instead of modifying feature values. Laugel et al. [75] says that if the explanation is not based on training data, but the artifacts of non-robustness of the classifier, it is unjustified. They define justified explanations to be connected to training data by a continuous set of datapoints, termed $\mathcal{E}$-chainability. Laugel et al. [73] enlist *proximity*, *connectedness*, and *stability* as three desirable properties of a counterfactual, along with the metrics to measure them. Barocas et al. [16] state five reasons which have lead to the success of counterfactual explanations and also point out at the overlooked assumptions. They mention the unavoidable conflicts which arises due to the need for privacy-invasion in order to generate helpful explanations. Pawelczyk et al. [86] provide a general upper bound on the cost of counterfactual explanations under the phenomenon of predictive multiplicity, wherein more than one trained models have the same test accuracy and there is no clear winner among them. Artelt and Hammer [15] enlists the counterfactual optimization problem formulation for several model-specific cases, like generalized linear model, gaussian naive bayes, and mention the general algorithm to solve them. Wexler et al. [112] developed a model-agnostic interactive visual tool for letting developers and practitioners visually examine the effect of changes in various features. Tsirtsis and Gomez-Rodriguez [105] cast the counterfactual generation problem as a Stackelberg game between the decision maker and the person receiving the prediction. Given a ground set of counterfactuals, the proposed algorithm returns the top-k counterfactuals, which maximizes the utility of both the involved parties. Downs et al. [36] propose to use conditional subspace VAEs, which is a variant of VAEs, to generate counterfactuals that obey correlations between features, causal relations between features, and personal preferences. This method directly uses the training data and is not based on the trained model. Therefore it is unclear whether the counterfactual generated by this method would also get the desired label by the model.

5 Evaluation of counterfactual generation algorithms

This section lists the common datasets used to evaluate counterfactual generation algorithms and the metrics they are typically evaluated and compared on.

5.1 Commonly used datasets for evaluation

The datasets used in evaluation in the papers we review can be categorized into tabular and image datasets. Not all methods support image datasets. Some of the papers also used synthetic datasets for evaluating their algorithms, but we skip those in this review since they were generated for a specific paper and also might not be available. Common datasets in the literature include:

- *Image* - MNIST [77].
- *Tabular* - Adult income [37], German credit [40], Compas recidivism [60], Student Performance [43], LSAT [19], Pima diabetes [100], Breast cancer [38], Default of credit [39], FICO [50], Fannie Mae [81], Iris [41], Wine [44], Shopping [42].

5.2 Metrics for evaluation of counterfactual generation algorithms

Most of the counterfactual generation algorithms are evaluated on the desirable properties of counterfactuals. Counterfactuals are thought of as actionable feedback to individuals who have received

[1] It considers global and local feature importance, not preference.
[2] All features are converted to polytope type.
[3] Does not generate new datapoints
[4] The distance is calculated in latent space.
[5] It considers feature importance not user preference.

undesirable outcome from automated decision makers, and therefore a user study can be considered a gold standard. However, none of the collected papers perform a user study. The ease of acting on a recommended counterfactual is thus measured by using quantifiable proxies:

1. *Validity* - Validity measures the ratio of the counterfactuals that actually have the desired class label to the total number of counterfactuals generated. Higher validity is preferable. Most papers report it.
2. *Proximity* - Proximity measures the distance of a counterfactual from the input datapoint. For counterfactuals to be easy to act upon, they should be close to the input datapoint. Distance metrics like L1 norm, L2 norm, Mahalanobis distance are common. To handle the variability of range among different features, some papers standardize them in pre-processing, or divide L1 norm by median absolute deviation of respective features [111, 85, 95], or divide L1 norm by the range of the respective features [65, 64, 26]. Some papers term proximity as the average distance of the generated counterfactuals from the input. Lower values of average distance are preferable.
3. *Sparsity* - Shorter explanations are more comprehensible to humans [84], therefore counterfactuals ideally should prescribe a change in a small number of features. Although a consensus on hard cap on the number of modified features has not been reached, Keane and Smyth [67] cap a sparse counterfactual to at most two feature changes.
4. *Counterfactual generation time* - Intuitively, this measures the time required to generate counterfactuals. This metric can be averaged over the generation of a counterfactual for a batch of input datapoints or for the generation of multiple counterfactuals for a single input datapoint.
5. *Diversity* - Some algorithms support the generation of multiple counterfactuals for a single input datapoint. The purpose of providing multiple counterfactuals is to increase the ease for applicants to reach at least one counterfactual state. Therefore the recommended counterfactuals should be diverse, giving applicants the choice to choose the easiest one. If an algorithm is strongly enforcing sparsity, there could be many different sparse subsets of the features that could be changed. Therefore, having a diverse set of counterfactuals is useful. Diversity is encouraged by maximizing the distance between the multiple counterfactuals by adding it as a term in the optimization objective [85, 26] or as a hard constraint [107, 64], or by minimizing the mutual information between all pairs of modified features [76]. Mothilal et al. [85] reported diversity as the feature-wise distance between each pair of counterfactuals. A higher value of diversity is preferable.
6. *Closeness to the training data* - Recent papers have considered the actionability and realisticness of the modified features by grounding them in the training data distribution. This has been captured by measuring the average distance to the k-nearest datapoints [26], or measuring the local outlier factor [63], or measuring the reconstruction error from a VAE trained on the training data [82, 108]. A lower value of the distance and reconstruction error is preferable.
7. *Causal constraint satisfaction (feasibility)* - This metric captures how realistic the modifications in the counterfactual are by measuring if they satisfy the causal relation between features. Mahajan et al. [82] evaluated their algorithm on this metric.
8. *IM1 and IM2* - Van Looveren and Klaise [108] proposed two interpretability metrics specifically for algorithms that use auto-encoders. Let the counterfactual class be t, and the original class be o. AE_t is the auto-encoder trained on training instances of class t, and AE_o is the auto-encoder trained on training instances of class o. Let AE be the auto-encoder trained on the full training dataset (all classes.)

$$IM1 = \frac{\|x_{cf} - AE_t(x_{cf})\|_2^2}{\|x_{cf} - AE_o(x_{cf})\|_2^2 + \epsilon} \tag{6}$$

$$IM2 = \frac{\|AE_t(x_{cf}) - AE(x_{cf})\|_2^2}{\|x_{cf}\|_1 + \epsilon} \tag{7}$$

A lower value of *IM1* implies that the counterfactual (x_{cf}) can be better reconstructed by the auto-encoder trained on the counterfactual class (AE_t) compared to the auto-encoder trained on the original class (AE_o). Thus implying that the counterfactual is closer to the data manifold of the counterfactual class. A lower value of *IM2* implies that the reconstruction from the auto-encoder trained on counterfactual class and the auto-encoder trained on all classes is similar. Therefore, a lower value of *IM1* and *IM2* means a more interpretable counterfactual.

Some of the reviewed papers did not evaluate their algorithm on any of the above metrics. They only showed a couple of example input and respective counterfactual datapoints, details about which are available in the full table (see appendix A).

6 Open Questions

In this section, we delineate the open questions and challenges yet to be tackled by future work in counterfactual explanations.

Research Challenge 1 *Unify counterfactual explanations with traditional "explainable AI."*

Although counterfactual explanations have been credited to elicit causal thinking and provide actionable feedback to users, they do not tell which feature(s) was the principal reason for the original decision, and why. It would be nice if, along with giving actionable feedback, counterfactual explanations also gave the reason for the original decision, which can help applicants understand the model's logic. This is addressed by traditional "explainable AI" methods like LIME [92], Anchors [93], Grad-CAM [96]. Guidotti et al. [57] have attempted this unification, as they first learn a local decision tree and then interpret the inversion of decision nodes of the tree as counterfactual explanations. However, they do not show the counterfactual explanations they generate, and their technique also misses other desiderata of counterfactuals (see section 3.2.)

Research Challenge 2 *Provide counterfactual explanations as discrete and sequential steps of actions.*

Current counterfactual generation approaches return the modified datapoint, which would receive the desired classification. The modified datapoint (state) reflects the idea of instantaneous and continuous actions, but in the real world, actions are discrete and often sequential. Therefore the counterfactual generation process must take the discreteness of actions into account and provide a series of actions that would take the individual from the current state to the modified state, which has the desired class label.

Research Challenge 3 *Counterfactual explanations as an interactive service to the applicants.*

Counterfactual explanations should be provided as an interactive interface, where an individual can come at regular intervals, inform the system of the modified state, and get updated instructions to achieve the counterfactual state. This can help when the individual could not precisely follow the earlier advice due to various reasons.

Research Challenge 4 *Ability of counterfactual explanations to work with incomplete—or missing—causal graphs.*

Incorporating causality in the process of counterfactual generation is essential for the counterfactuals to be grounded in reality. Complete causal graphs and structural equations are rarely available in the real world, and therefore the algorithm should be able to work with incomplete causal graphs. Mahajan et al. [82]'s approach works with incomplete causal graphs, but this challenge has been scarcely incorporated into other methods.

Research Challenge 5 *The ability of counterfactual explanations to work with missing feature values.*

Along the lines of an incomplete causal graph, counterfactual explanation algorithms should also be able to handle missing feature values, which often happens in the real world [52].

Research Challenge 6 *Scalability and throughput of counterfactual explanations generation.*

As we see in table 1, most approaches need to solve an optimization problem to generate one counterfactual explanation. Some papers generate multiple counterfactuals while solving optimizing once, but they still need to optimize for different input datapoints. Counterfactual generating algorithms should, therefore, be more scalable. Mahajan et al. [82] learn a VAE which can generate multiple counterfactuals for any given input datapoint after training. Therefore, their approach is highly scalable.

Research Challenge 7 *Counterfactual explanations should account for bias in the classifier.*

Counterfactuals potentially capture and reflect the bias in the models. To underscore this as a possibility, Ustun et al. [107] experimented on the difference in the difficulty of attaining the provided counterfactual state across genders, which clearly showed a significant difference in the difficulty. More work requires to be done to find how equally easy counterfactual explanations can be provided across different demographic groups, or how adjustments should be made in the prescribed changes in order to account for the bias.

Research Challenge 8 *Generate robust counterfactual explanations.*

Counterfactual explanation optimization problems force the modified datapoint to obtain the desired class label. However, the modified datapoint could be labeled either in a robust manner or due to the classifier's non-robustness, e.g., an overfitted classifier. This can generate counterfactuals that might be non-sensical and have the desired class label only because of the classifier's artifact. Laugel et al. [73] term this as the *stability* property of a counterfactual. This is specifically a challenge for approaches that solve an optimization problem each time they generate a counterfactual (see **RC6**.) We see potential overlap between this nascent literature and the certifiability literature from the adversarial machine learning community.

Research Challenge 9 *Counterfactual explanations should handle dynamics (data drift, classifier update, applicant's utility function changing, etc.)*

All counterfactual explanation papers we review, assume that the underlying black box does not change over time and is monotonic. However, this might not be true; credit card companies and banks update their models as frequently as 12-18 months [7]. Therefore counterfactual explanation algorithms should take data drift, the dynamism and non-monotonicity of the classifier into account.

Research Challenge 10 *Counterfactual explanations should capture applicant's preferences.*

Along with the distinction between mutable and immutable features (finely classified into actionable, mutable, and immutable), counterfactual explanations should also capture preferences specific to an applicant. This is important because the ease of changing different features can differ across applicants. Mahajan et al. [82] captures the applicant's preferences using an oracle, but that is expensive and is still a challenge.

Research Challenge 11 *Counterfactual explanations should also inform the applicants about what must not change*

If a counterfactual explanation advises someone to increase their *income* but does not tell that their *length of last employment* should not decrease. And the applicant, in order to increase their income, switches to a higher-paying job may find themselves in a worse position than earlier. Thus by failing to disclose what must not change, an explanation may lead the applicant to an unsuccessful state [16]. This corroborates **RC3**, whereby an applicant might be able to interact with an interactive platform to see the effect of a potential real-world action they are considering to take to achieve the counterfactual state.

Research Challenge 12 *Handling of categorical features in counterfactual explanations*

Different papers have come up with various methods to handle categorical features, like converting them to one-hot encoding and then enforcing the sum of those columns to be 1 using regularization or hard-constraint, or clamping an optimization problem to a specific categorical value, or leave them to be automatically handled by genetic approaches and SMT solvers. Measuring distance in categorical features is also not obvious. Some papers use indicator function, which equates to 1 for unequal values and 0 if the same; other papers convert to one-hot encoding and use standard distance metric like L1/L2 norm. Therefore none of the methods developed to handle categorical features are obvious; future research must consider this and develop appropriate methods.

Research Challenge 13 *Evaluate counterfactual explanations using a user study.*

The evaluation for counterfactual explanations must be done using a user study because evaluation proxies (see section 5) might not be able to precisely capture the psychological and other intricacies of human cognition on the ease of actionability of a counterfactual.

Research Challenge 14 *Counterfactual explanations should be integrated with visualization features.*

Counterfactual explanations will be directly interacting with consumers who can have varying technical knowledge levels, and therefore, counterfactual generation algorithms should be integrated with visualizations. We already know that visualization can influence behavior [24]. This could involve collaboration between machine learning and HCI communities.

Research Challenge 15 *Strengthen the ties between machine learning and regulatory communities.*

A joint statement between the machine learning community and regulatory community (OCC, Federal Reserve, FTC, CFPB) acknowledging successes and limitations of where counterfactual explanations will be adequate for legal and consumer-facing needs would improve the adoption and use of counterfactual explanations in critical software.

7 Conclusions

In this paper, we collected and reviewed 39 papers, which proposed various algorithmic solutions to finding counterfactual explanations to the decisions produced by automated systems, specifically automated by machine learning. The evaluation of all the papers on the same rubric helps in quickly understanding the peculiarities of different approaches, the advantages, and disadvantages of each of them, which can also help organizations choose the algorithm best suited to their application constraints. This has also helped us identify the gaps readily, which will be beneficial to researchers scouring for open problems in this space and for quickly sifting the large body of literature. We hope this paper can also be the starting point for people wanting to get an introduction to the broad area of counterfactual explanations and guide them to proper resources for things they might be interested in.

References

[1] [n. d.]. Adverse Action Notice Requirements Under the ECOA and the FCRA. https://consumercomplianceoutlook.org/2013/second-quarter/adverse-action-notice-requirements-under-ecoa-fcra/. Accessed: 2020-10-15.

[2] [n. d.]. Algorithms in decision making. https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/351.pdf. Accessed: 2020-10-15.

[3] [n. d.]. Artificial Intelligence. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/ict-26-2018-2020. Accessed: 2020-10-15.

[4] [n. d.]. Broad Agency Announcement: Explainable Artificial Intelligence (XAI). https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf. Accessed: 2020-10-15.

[5] [n. d.]. The European Commission offers significant support to Europe's AI excellence. https://www.eurekalert.org/pub_releases/2020-03/au-tec031820.php. Accessed: 2020-10-15.

[6] [n. d.]. FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. Accessed: 2020-10-15.

[7] [n. d.]. MODEL LIFECYCLE TRANSFORMATION: HOW BANKS ARE UNLOCKING EFFICIENCIES. https://financeandriskblog.accenture.com/risk/model-lifecycle-transformation-how-banks-are-unlocking-efficiencies. Accessed: 2020-10-15.

[8] [n. d.]. Notification of action taken, ECOA notice, and statement of specific reasons. https://www.consumerfinance.gov/policy-compliance/rulemaking/regulations/1002/9/. Accessed: 2020-10-15.

[9] [n. d.]. RAPPORT DE SYNTHESE FRANCE INTELLIGENCE ARTIFICIELLE. https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf. Accessed: 2020-10-15.

[10] [n. d.]. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed: 2020-10-15.

[11] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* PP (09 2018), 1–1. https://doi.org/10.1109/ACCESS.2018.2870052

[12] Charu C. Aggarwal, Chen Chen, and Jiawei Han. 2010. The Inverse Classification Problem. *J. Comput. Sci. Technol.* 25, 3 (May 2010), 458–468. https://doi.org/10.1007/s11390-010-9337-x

[13] Robert Andrews, Joachim Diederich, and Alan B. Tickle. 1995. Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. *Know.-Based Syst.* 8, 6 (Dec. 1995), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4

[14] Daniel Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4) (06 2020), 1059–1086. https://doi.org/10.1111/rssb.12377

[15] André Artelt and Barbara Hammer. 2019. On the computation of counterfactual explanations – A survey. http://arxiv.org/abs/1911.07749

[16] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89. https://doi.org/10.1145/3351095.3372830

[17] C. Van Fraassen Bas. 1980. *The Scientific Image*. Oxford University Press.

[18] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[19] R. D. Boch and M. Lieberman. 1970. Fitting a response model for n dichotomously scored items. *Psychometrika* 35 (1970), 179–97.

[20] Ruth Byrne. 2008. The Rational Imagination: How People Create Alternatives to Reality. *The Behavioral and brain sciences* 30 (12 2008), 439–53; discussion 453. https://doi.org/10.1017/S0140525X07002579

[21] Ruth M. J. Byrne. 2019. 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, California, USA, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

[22] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[23] Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. 2018. Metamorphic Testing: A Review of Challenges and Opportunities. *ACM Comput. Surv.* 51, 1 (Jan. 2018), 27. https://doi.org/10.1145/3143561

[24] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300418

[25] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-Structured Representations of Trained Networks. In *Conference on Neural Information Processing Systems (NeurIPS) (NIPS'95)*. MIT Press, Cambridge, MA, USA, 24–30.

[26] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. http://arxiv.org/abs/2004.11165

[27] A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, USA, 598–617.

[28] Houtao Deng. 2014. Interpreting Tree Ensembles with inTrees. *arXiv:1408.5456* (08 2014). https://doi.org/10.1007/s41060-018-0144-8

[29] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 590–601.

[30] Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. 2019. Model Agnostic Contrastive Explanations for Structured Data. http://arxiv.org/abs/1906.00117

[31] Edsger W Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische mathematik* 1, 1 (1959), 269–271.

[32] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

[33] Carl Doersch. 2016. Tutorial on Variational Autoencoders. arXiv:stat.ML/1606.05908

[34] Pedro Domingos. 1998. Knowledge Discovery Via Multiple Models. *Intell. Data Anal.* 2, 3 (May 1998), 187–202.

[35] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, D. O'Brien, Stuart Schieber, J. Waldo, D. Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation.

[36] Michael Downs, Jonathan Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei. Pan. 2020. CRUDS: Counterfactual Recourse Using Disentangled Subspaces. In *Workshop on Human Interpretability in Machine Learning (WHI)*. https://finale.seas.harvard.edu/files/finale/files/cruds-_counterfactual_recourse_using_disentangled_subspaces.pdf

[37] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Adult Income. http://archive.ics.uci.edu/ml/datasets/Adult

[38] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Breast Cancer. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[39] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Default Prediction. http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[40] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - German Credit. http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[41] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Iris. https://archive.ics.uci.edu/ml/datasets/iris

[42] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Shopping. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[43] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Student Performance. http://archive.ics.uci.edu/ml/datasets/Student%2BPerformance

[44] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - Wine. https://archive.ics.uci.edu/ml/datasets/wine

[45] Jannik Dunkelau and Michael Leuschel. 2019. Fairness-Aware Machine Learning. , 60 pages. https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft/KMW_I/Working_Paper/Dunkelau___Leuschel__2019__Fairness-Aware_Machine_Learning.pdf

[46] Daniel Faggella. 2020. Machine Learning for Medical Diagnostics – 4 Current Applications. https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/. Accessed: 2020-10-15.

[47] Rubén R. Fernández, Isaac Martín de Diego, Víctor Aceña, Alberto Fernández-Isabel, and Javier M. Moguerza. 2020. Random forest explainability using counterfactual sets. *Information Fusion* 63 (Nov. 2020), 196–207. https://doi.org/10.1016/j.inffus.2020.07.001

[48] Carlos Fernández-Loría, Foster Provost, and Xintian Han. 2020. Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach. http://arxiv.org/abs/2001.07417

[49] Carlos Fernández-Loría, Foster Provost, and Xintian Han. 2020. Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach.

[50] FICO. 2018. FICO (HELOC) dataset. https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2

[51] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. http://www.jstor.org/stable/2699986

[52] P. J. García-Laencina, J. Sancho-Gómez, and A. R. Figueiras-Vidal. 2009. Pattern classification with missing data: a review. *Neural Computing and Applications* 19 (2009), 263–282.

[53] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2013. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 24 (09 2013). https://doi.org/10.1080/10618600.2014.907095

[54] Bryce Goodman and S. Flaxman. 2016. EU regulations on algorithmic decision-making and a "right to explanation". *ArXiv* abs/1606.08813 (2016).

[55] Preston Gralla. 2016. Amazon Prime and the racist algorithms. https://www.computerworld.com/article/3068622/amazon-prime-and-the-racist-algorithms.html

[56] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable Credit Application Predictions With Counterfactual Explanations. http://arxiv.org/abs/1811.05245

[57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. http://arxiv.org/abs/1805.10820

[58] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[59] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A Peek into the Black Box: Exploring Classifiers by Randomization. *Data Min. Knowl. Discov.* 28, 5–6 (Sept. 2014), 1503–1529. https://doi.org/10.1007/s10618-014-0368-8

[60] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. 2016. UCI Machine Learning Repository. https://github.com/propublica/compas-analysis/

[61] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. http://arxiv.org/abs/1907.09615

[62] D. Kahneman and D. Miller. 1986. Norm Theory: Comparing Reality to Its Alternatives. *Psychological Review* 93 (1986), 136–153.

[63] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. 2020. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, California, USA, 2855–2862. https://doi.org/10.24963/ijcai.2020/395

[64] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. 2020. Model-Agnostic Counterfactual Explanations for Consequential Decisions. http://arxiv.org/abs/1905.11190

[65] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic Recourse: from Counterfactual Explanations to Interventions. http://arxiv.org/abs/2002.06278

[66] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. http://arxiv.org/abs/2006.06831

[67] Mark T. Keane and Barry Smyth. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). arXiv:cs.AI/2005.13997

[68] Boris Kment. 2006. Counterfactuals and Explanation. *Mind* 115 (04 2006). https://doi.org/10.1093/mind/fzl261

[69] Will Knight. 2019. The Apple Card Didn't 'See' Gender—and That's the Problem. https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

[70] R. Krishnan, G. Sivakumar, and P. Bhattacharya. 1999. Extracting decision trees from trained neural networks. *Pattern Recognition* 32, 12 (1999), 1999 – 2009. https://doi.org/10.1016/S0031-3203(98)00181-2

[71] Sanjay Krishnan and Eugene Wu. 2017. PALM: Machine Learning Explanations For Iterative Debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17)*. Association for Computing Machinery, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/3077257.3077271

[72] Michael T. Lash, Qihang Lin, William Nick Street, Jennifer G. Robinson, and Jeffrey W. Ohlmann. 2017. Generalized Inverse Classification. In *SDM*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 162–170.

[73] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. arXiv:1906.04774

[74] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2018. Comparison-Based Inverse Classification for Interpretability in Machine Learning. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems, Theory and Foundations (IPMU)*. Springer International Publishing. https://doi.org/10.1007/978-3-319-91473-2_9

[75] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. http://arxiv.org/abs/1907.09294

[76] Thai Le, Suhang Wang, and Dongwon Lee. 2019. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. arXiv:cs.LG/1911.02042

[77] Yann LeCun and Corinna Cortes. 2010. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/. (2010). http://yann.lecun.com/exdb/mnist/

[78] David Lewis. 1973. *Counterfactuals*. Blackwell Publishers, Oxford.

[79] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. 2020. Actionable Interpretability through Optimizable Counterfactual Explanations for Tree Ensembles. http://arxiv.org/abs/1911.12199

[80] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[81] Fannie Mae. 2020. Fannie Mae dataset. https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html

[82] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2020. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. http://arxiv.org/abs/1912.03277

[83] David Martens and Foster J. Provost. 2014. Explaining Data-Driven Document Classifications. *MIS Q.* 38 (2014), 73–99.

[84] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[85] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (FAT* '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3351095.3372850

[86] Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of Machine Learning Research*, Jonas Peters and David Sontag (Eds.). PMLR, Virtual, 9. http://proceedings.mlr.press/v124/pawelczyk20a.html

[87] Martin Pawelczyk, Johannes Haug, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data. , 3126–3132 pages. https://doi.org/10.1145/3366423.3380087 arXiv: 1910.09398.

[88] Judea Pearl. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, USA.

[89] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. , 344–350 pages. https://doi.org/10.1145/3375627.3375850 arXiv: 1909.09369.

[90] Goutham Ramakrishnan, Y. C. Lee, and Aws Albarghouthi. 2020. Synthesizing Action Sequences for Modifying Model Decisions. In *Conference on Artificial Intelligence (AAAI)*. AAAI press, California, USA, 16. http://arxiv.org/abs/1910.00057

[91] Shubham Rathi. 2019. Generating Counterfactual and Contrastive Explanations using SHAP. http://arxiv.org/abs/1906.09293 arXiv: 1906.09293.

[92] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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 (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Conference on Artificial Intelligence (AAAI)*. AAAI press, California, USA, 9. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982

[94] David-Hillel Ruben. 1992. *Counterfactuals*. Routledge Publishers. https://philarchive.org/archive/RUBEE-3

[95] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 20–28. https://doi.org/10.1145/3287560.3287569

[96] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *IEEE International Conference on Computer Vision*. 618–626.

[97] Kumba Sennaar. 2019. Machine Learning for Recruiting and Hiring – 6 Current Applications. https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring/. Accessed: 2020-10-15.

[98] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. http://arxiv.org/abs/1905.07857

[99] Saurav Singla. 2020. Machine Learning to Predict Credit Risk in Lending Industry. https://www.aitimejournal.com/@saurav.singla/machine-learning-to-predict-credit-risk-in-lending-industry. Accessed: 2020-10-15.

[100] J. W. Smith, J. Everhart, W. C. Dickson, W. Knowler, and R. Johannes. 1988. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, Washington,D.C., 261–265.

[101] Kacper Sokol and Peter Flach. 2019. Desiderata for Interpretability: Explaining Decision Tree Predictions with Counterfactuals. *Conference on Artificial Intelligence (AAAI)* 33 (July 2019). https://doi.org/10.1609/aaai.v33i01.330110035

[102] Jason Tashea. 2017. Courts Are Using AI to Sentence Criminals. That Must Stop Now. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/. Accessed: 2020-10-15.

[103] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. arXiv:cs.LG/1907.07374

[104] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-Based Ensembles via Actionable Feature Tweaking. In *International Conference on Knowledge Discovery and Data Mining (KDD) (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 465–474. https://doi.org/10.1145/3097983.3098039

[105] Stratis Tsirtsis and Manuel Gomez-Rodriguez. 2020. Decisions, Counterfactual Explanations and Strategic Behavior. arXiv:cs.LG/2002.04333

[106] Ryan Turner. 2016. A Model Explanation System: Latest Updates and Extensions. *ArXiv* abs/1606.09517 (2016).

[107] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT) (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19. https://doi.org/10.1145/3287560.3287566

[108] Arnaud Van Looveren and Janis Klaise. 2020. Interpretable Counterfactual Explanations Guided by Prototypes. http://arxiv.org/abs/1907.02584 arXiv: 1907.02584.

[109] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[110] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (06 2017). https://doi.org/10.1093/idpl/ipx005

[111] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal* 31, 2 (2017), 842–887. https://doi.org/10.2139/ssrn.3063289

[112] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65.

[113] Adam White and Artur d'Avila Garcez. 2019. Measurable Counterfactual Local Explanations for Any Classifier. http://arxiv.org/abs/1908.03020

[114] James Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

[115] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*. IEEE, New York, USA, 2921–2929.

[116] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. 2009. The Feature Importance Ranking Measure. In *Machine Learning and Knowledge Discovery in Databases*, Vol. 5782. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04174-7_45

A Full Table

Initially, we categorized the set of papers with more columns and in a much larger table. We selected the most critical columns and put them in table 1. We will provide the full table in the final version of the paper.

B Methodology

B.1 How we collected the paper to review?

We collected a set of 39 papers. In this section, we provide the exact procedure used to arrive at this set of papers. We started from a seed set of papers recommended by other people [82, 85, 90, 107, 111], followed by snowballing their references.

For an even complete search, we searched for "counterfactual explanations", "recourse", and "inverse classification" on two popular search engines for scholarly articles, Semantic Scholar and Google scholar. On both the search engines, we looked for papers published in the last five years. This is a sensible time-frame since the paper that started the discussion of counterfactual explanations in the context of machine learning (specifically for tabular data) was published in 2017 [111]. We collect papers that were published before 31st July 2020. The papers we collected were published at conferences like KDD, IJCAI, FAccT, AAAI, WWW, NeurIPS, WHI, or uploaded to Arxiv.

B.2 Scope of the review

Even though the first paper we reviewed was published online in 2017, and most other papers we review cite it Wachter et al. [111] as the seminal paper that started the discussion around counterfactual explanations, we do not claim that this is an entirely new idea. Communities from data mining [49, 83], causal inference [88], and even software engineering [23] have explored similar ideas to identify the principal cause of a prediction, an effect, and a bug, respectively. Even before the emergence of counterfactual explanations in applied fields, they have been the topic of discussion in fields like social sciences [84], philosophy [68, 78, 94], psychology [20, 21, 62]. In this review paper, we restrict our discussion to the recent set of papers that discuss counterfactual explanations in machine learning, specifically classification settings. These papers have been inspired by the emerging trend of FATE and the legal requirements pertaining to explainability in tasks automated by machine learning algorithms.

C Burgeoning legal frameworks around explanations in AI

To increase the accountability of automated decision systems—specifically, AI systems—laws and regulations regarding the decisions produced by such systems have been proposed and implemented across the globe [35]. The most recent version of the European Union's General Data Protection Regulation (GDPR), enforced starting on May 25, 2018, offered a right to information about the existence, logic, and envisaged consequences of such a system [54]. This also includes the right to not being a subject of an automated decision making system. Although the closeness of this law to "right to explanation" is debatable and ambiguous [110], the official interpretation by Working Party for Article 29 has concluded that the GDPR requires explanations of specific decisions, and therefore counterfactual explanations are apt. In the US, the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) require the creditor to inform the reasons for an adverse action, such as rejection of a loan request [8, 1]. They generally compare the applicant's feature to the average value in the population to arrive at the principal reasons. Government reports from the United Kingdom [2] and France [6, 9] also touched on the issue of explainability in AI systems. In the US, Defense Advanced Research Projects Agency (DARPA) launched the Explainable AI (XAI) program in 2016 to encourage research into designing explainable models, understanding the psychological requirements of explanations, and the design of explanation interfaces [4]. The European Union has taken similar initiatives as well [3, 5]. While many techniques have been proposed for explainable machine learning, it is yet unclear if and how these specific techniques can help address the letter of the law. Future collaboration between AI researchers, regulators, the legal community, and consumer watchdog groups will help ensure the development of trustworthy AI.