Achieving Diversity in Counterfactual Explanations: a Review and Discussion

Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, Marcin Detyniecki

Achieving Diversity in Counterfactual Explanations: a Review and Discussion

Thibault Laugel*
AXA
Paris, France
thibault.laugel@axa.com

Adulam Jeyasothy*
Sorbonne Université, CNRS, LIP6,
F-75005, Paris, France
adulam.jeyasothy@lip6.fr

Marie-Jeanne Lesot
Sorbonne Université, CNRS, LIP6,
F-75005, Paris, France

Christophe Marsala
Sorbonne Université, CNRS, LIP6,
F-75005, Paris, France

Marcin Detyniecki
AXA, Paris, France
Sorbonne Université, CNRS, LIP6,
F-75005, Paris, France
Polish Academy of Science, IBS PAN
Warsaw, Poland

ABSTRACT

In the field of Explainable Artificial Intelligence (XAI), counterfactual examples explain to a user the predictions of a trained decision model by indicating the modifications to be made to the instance so as to change its associated prediction. These counterfactual examples are generally defined as solutions to an optimization problem whose cost function combines several criteria that quantify desiderata for a good explanation meeting user needs. A large variety of such appropriate properties can be considered, as the user needs are generally unknown and differ from one user to another; their selection and formalization is difficult. To circumvent this issue, several approaches propose to generate, rather than a single one, a set of diverse counterfactual examples to explain a prediction. This paper proposes a review of the numerous, sometimes conflicting, definitions that have been proposed for this notion of diversity. It discusses their underlying principles as well as the hypotheses on the user needs they rely on and proposes to categorize them along several dimensions (explicit vs implicit, universe in which they are defined, level at which they apply), leading to the identification of further research challenges on this topic.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; • **Human-centered computing**;

KEYWORDS

XAI, counterfactual explanations, actionable recourse, explainability, interpretability, transparency, survey, review, diversity.

*Both authors contributed equally to this research.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594122

ACM Reference Format:
Thibault Laugel, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2023. Achieving Diversity in Counterfactual Explanations: a Review and Discussion. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594122

1 INTRODUCTION

Over the last years, the need for a better understanding, and accountability, of Machine Learning systems has led to the soaring of domains around the topic of Responsible Artificial Intelligence. Among these, the eXplainable Artificial Intelligence (XAI) domain [9, 37] focuses on the generation of explanations for the decisions of AI and Machine Learning models. In particular, local post-hoc methods [23] aim at generating explanations regarding the prediction performed by a given trained classifier (post-hoc property) for a given data instance of interest (local property). They come in different formats such as feature importance (e.g LIME [59] and SHAP [38]) or counterfactual examples [79] (e.g. Growing Spheres [34] and FACE [54]).

However, generating explanations has been proven to be a difficult task, due to the subjective and vague nature of the concept of interpretability [69], and thus the difficulty to define what a good explanation is. This topic has been explored from the point of view of cognitive sciences as well as educational sciences among others, as for instance summarised in the rich survey proposed by Miller [41] that underlines the wide range of possibly desirable properties. As a result, numerous explainability approaches have been proposed over the last years, mirroring the absence of consensus regarding the properties explanations should satisfy. This issue is even more prevalent, and visible in the case of counterfactual examples [21, 79]. Indeed, this type of instance-based explanations relies on solving an optimization problem, and therefore explicitly depends on the selection of considered quality criteria. Although some of these seem to be consensual, such as the closeness to the instance of interest and the sparsity of the explanation [76], there is generally no global agreement over numerous additional possibly desirable properties, nor on how to formalize them in measurable numerical criteria. Moreover, once selected, these criteria most often need to be combined or aggregated to define a multi-criteria optimization problem, generally resulting in the generation a single

final explanation (see e.g. [2, 34, 39, 54, 79]). Consequently, in addition to the issue of identifying the most relevant criteria assessing the quality of a candidate counterfactual example, the choice of this aggregation operator is also obviously crucial and plays a major role in the implementation of the definition of a good explanation. The possibly subjective and personal characteristic of the latter can be considered as advocating for making it dependent on the targeted user, his/her specific needs and prior knowledge. Yet, the selection of this aggregation is rarely motivated and often implicitly relies on the principle that all of the selected criteria should be optimized at the same time. However, this is often impossible, due to their mutual dependencies and the trade-offs existing between them.

This paper proposes to discuss the importance of this aggregation operator and shows that failing to motivate its choice, as it is often the case in the literature, may lead to unsatisfactory explanations. This provides additional arguments to questioning the relevance of generating a single counterfactual example and arguing that the generation of multiple counterfactual examples may be more suited to meet (sometimes unformulated) user needs, a commonly identified shortcoming of interpretability. Turning to existing approaches that make it possible to generate such multiple explanations, the paper then proposes to discuss the notion of diversity they usually integrate, so that the output explanations are not redundant one with another. It offers to categorize these approaches based on the strategies they consider for this multiple explanation generation problem, and discusses how these help overcoming one pitfall of interpretability methods: matching explanations to unobserved user needs. The contributions of the paper are thus both to discuss the generally poorly covered topic of combining quality criteria for XAI methods, and to propose a review of the current literature on diverse counterfactual methods.

The paper is organized as follows. Section 2 first reminds the formal definition of explanations based on counterfactual examples as well as the most common quality criteria proposed in existing works. Section 3 then discusses the importance of the operator used to aggregate these criteria and its potential overlooked undesired consequences. After discussing how providing multiple explanations to users might help circumventing these issues, we propose in Section 4 a survey of different strategies to achieve this goal, discussing the various notions of diversity they rely on and how they integrate it in the explanation generation process. This study opens the way to identifying research challenges, as discussed in Section 5. Section 6 concludes the paper.

2 BACKGROUND: COUNTERFACTUAL EXPLANATIONS AND COMMON QUALITY CRITERIA

Within the wide domain of XAI [43, 62, 71], counterfactual example explanations (see [1, 21, 40, 76] for dedicated surveys) focus on the case where a user wants to understand the reason for a given prediction: given a data instance of interest, denoted x in the following, and a trained machine learning model, denoted f, counterfactual examples aim at providing insights to understand the generated prediction $f(x)$. More precisely, counterfactual examples aim at answering the question "Why $f(x)$ and not c'?", where c' denotes a possible alternative output the model f may have given. The answer to this question is expressed through a set of modifications that can be applied to x to obtain the different prediction by the model, which amounts to answering the question: "What changes would be required to modify this prediction?".

This type of explanations, directly meeting the desirable *contrastive* property for a good explanation proposed by Miller [41], has been praised for their higher transparency [79] and actionability [29] as compared to other types of explanations. Indeed, they have the benefit of directly describing actions that can be performed by the user to have a recourse on the prediction. On the contrary, other forms of explanations such as local feature importance vectors rely on debatable, sometimes opaque definitions of importance, and therefore have been shown to be often misunderstood (see for instance [31] for misuses of the explanations provided by SHAP [38]).

There exists numerous approaches to build counterfactual example explanations (e.g. see [21] for one of the latest surveys). This section does not aim at providing such an exhaustive overview, but rather to summarize the required background on which the later discussion is built: after providing a reminder about the general formulation problem to generate counterfactual examples, it discusses some desirable properties they have been required to offer.

2.1 General Formulation of the Counterfactual Example Generation Problem

This section introduces the general counterfactual example problem, focusing on the case when the machine learning model to be explained is a binary classifier, which is the most classical one. In this case, the alternative output c' given by this classifier can only be the other class. Denoting $\mathcal{X}$ the data feature space and $\mathcal{Y}$ the (binary) label space, the classifier is $f : \mathcal{X} \longrightarrow \mathcal{Y}$ and $x \in \mathcal{X}$ is the data instance about which the user requests an explanation regarding its associated prediction $f(x)$. A counterfactual example explaining this prediction is then formally defined as:

$$e^* = \underset{e \in \mathcal{E}}{\operatorname{argmin}}\, pen_x(e) \tag{1}$$

that depends on the search space $\mathcal{E}$ and the penalty function pen_x . The former, $\mathcal{E}$, defines the space in which the final explanation, the counterfactual example, is allowed to evolve. In its most general form, it is defined as the set of all instances predicted to belong to a different class than x, formally:

$$\mathcal{E} = \{e \in \mathcal{X}, f(e) \neq f(x)\} \tag{2}$$

This formulation assumes that all instances from the opposite class from x are equivalent, which is not always the case. For instance, some works, e.g. [79], take into account, when available, the classification score given as output by the classifier.

The penalty function pen_x to be minimized, measures the cost of a candidate explanation, which corresponds to a decreasing function of its quality. The most commonly accepted definition imposes that the counterfactual example must be as close as possible to the instance of interest, so as to minimize the amount of changes needed to alter the prediction, i.e. to lower the efforts required from the user to meet his/her objective. A common choice to capture numerically this closeness requirement is to define the penalty function as a distance between the candidate counterfactual example

Type	General	Data-context.	User-context.
Property	closeness, sparsity	local density proximity path density justification	actionability causality, personalization
Dependency	x	x, X	$x, U(opt. : X)$

Table 1: Categorization of the most frequently considered desirable properties for a candidate counterfactual example e. The last row indicates the parameters they depend on: x is the data instance of interest, X a set of data points, U the user who receives the explanation.

and the instance of interest:

$$pen_x(e) = \| x - e \| \tag{3}$$

Existing works often consider l_1 [79] or l_2 [32, 34] distances. However, other distances are sometimes used, such as weighted Manhattan distance [2], or elastic net loss [75].

Although alternative formulations of the counterfactual problem can be found in some works (e.g. [79]) framing it as a weighted sum of the penalty function and a classification confidence score, the general counterfactual problem formulation presented in Equation 1 can be used to understand most existing counterfactual approaches. However, it is important to keep in mind that a lot of approaches to generate counterfactual examples actually rely on heuristic and do not make explicit the underlying cost function they optimize.

2.2 Some Desirable Properties for Counterfactual Examples

Beside being close to the instance whose prediction is to be explained, other desirable properties have been identified for counterfactual examples. These further constrain the optimization problem through associated criteria, aiming to lead to more understandable, useful or relevant explanations. Some existing literature reviews (see e.g. [76]) have proposed to categorize counterfactual methods depending on these formulated objectives. In this section, we list some of them, depending on whether they are general for any candidate or aim at taking into account some enriched information about their context, if available. For the latter, we distinguish between data-contextualisation and user-contextualisation. Table 1 summarises this categorisation. For each desirable property, we also present the associated criteria generally used to represent it in the optimization problem.

2.2.1 Sparsity. One of the most frequent desiderata for counterfactual explanations is the sparsity of the explanation vector. Indeed, it ensures that the effort required to alter the prediction of x focuses on a low number of features, making it more understandable and actionable by the user. Sparsity can be numerically measured by the l_0 distance [14, 22, 34], and optimized directly or through the l_1 distance [2, 79].

2.2.2 Data-contextualisation Criteria. Blindly minimizing a distance function has been observed to lead to unrealistic explanations: the solution to the optimisation problem may for instance lie out of the data distribution [33, 35], making it difficult to understand or even fully absurd to the user receiving the explanation. To avoid this issue, some constraints on where the counterfactual example should lie are often proposed and aim at providing a contextual enrichment to its generation, depending on possibly available additional information about the data.

These constraints can for instance be directly formulated as hard convex constraints on the allowed feature range [28, 32, 55, 75], or by imposing that the generated counterfactual example belongs to a dense region of the distribution [2, 54]. Rather than density constraints, Laugel et al. [35] argue that counterfactual explanations should be "justified" by the ground-truth data, guaranteeing their realism by their connection to observed data points. On a different note, model-based counterfactual approaches, generally relying on Autoencoders Generative Adversarial networks, impose through a reconstruction loss that the generated counterfactual should be close (proximity criterion), for instance using Euclidean distance, to ground-truth instances [39, 45, 51, 75].

Instead of solely imposing constraint on the counterfactual example to generate, Poyiadzi et al. [54] argues that the entirety of the path connecting it to x should be in data-dense regions only, to ensure that the intermediary steps are feasible.

2.2.3 User-centered Contextualisation Criteria. Measuring the quality of a candidate counterfactual example in the context of its use should also be made dependent on the user it is generated for, allowing for the generation of personalised explanations [26, 39, 73]: beyond the context provided by the domain and the other data among which it is looked for, taking into account the user context makes it possible to get back to a human-in-the-loop paradigm that is crucial in the XAI domain. Indeed, as mentioned in the introduction, actionability is often one of the strongest arguments in favor of the use of counterfactual explanations. As such, it has been included by some works as an explicit objective to guarantee more useful explanations, under the assumption that such user information is available.

Some works [28, 73] for instance consider that a set of editable features is provided by the users, so that an actionable counterfactual explanation is one that requires modification along these features only. On a different note, actionability is also integrated through causal constraints (see [29] for a survey on actionable recourse). The underlying assumption is that the changes along different features proposed by a counterfactual explanation are not independent: in order to be actionable, an explanation should take into account the causal relationships between features. The latter can for instance be modelled within Pearl's framework [53] to represent a causal graph and possibly structural causal equations describing the causal interactions between the features. Actionability is then measured as the extent to which the explanation fits this graph (see for instance [30, 39], and more generally [29, 50] for recent surveys). Another user-centered constraint is proposed by Jeyasothy et al [26], who argue that explanations should be personalized and adapted to the user's knowledge for it to be understood. A similar notion is proposed in [81], where a personalized cost function is proposed to answer user's needs.

In addition to these properties, mostly centered around how easy to understand and use an explanation is, other criteria, out of the

scope of this work, could be mentioned. These include for instance constraints that the explanation provider may want to impose, such as information leakage risk [52] or robustness to manipulation [66].

3 THE IMPLICIT DIFFICULTY OF GENERATING ONE COUNTERFACTUAL EXPLANATION

As discussed in the previous section, the definition of interpretability objectives and numerical criteria to measure the extent to which they are achieved is a difficult and complex task. However, selecting which properties are the most relevant for a given problem is only one of the issues to be considered: once they have been expressed, either by the user or the machine learning practitioner, most existing counterfactual approaches then combine them into an aggregated cost function, to be minimised to generate a unique explanation. Although a rich literature on aggregation operators exists (see e.g. [10, 15, 19]), it is rarely leveraged in the field of XAI, leaving the topic of combining explainability criteria, to the best of our knowledge, rarely discussed. Yet, this way of combining different criteria obviously directly impacts the generated explanation. After detailing how this combination is usually done in the state of the art, we propose in this section a discussion to question the relevance of this step itself.

3.1 Criteria Combination Methods

Although the properties and their associated criteria presented in the previous section are by nature all desirable, it is usually impossible by design to maximize all of them. For instance, optimizing the sparsity of a counterfactual explanation is often at odds with maximizing its closeness to the instance of interest, as shown by [34] for instance. As a result, conjunctive aggregation operators, that require all criteria to be simultaneously satisfied, are rarely used by counterfactual methods. Recognizing this impossibility, some approaches thus define an objective function that constitutes an explicit trade-off between the various criteria. This is for instance how Mahajan et al. [39] propose to aggregate the penalty of the explanation (measured by the l_1 distance) and the degree to which it satisfies the considered causal constraints (causal distance). Similarly, Jeyasothy et al. [26] balance the penalty with the incompatibility to the user knowledge using a weighted average, to propose a personalized explanation.

However, specifying the right balance between criteria can be difficult for the user. This is even truer as the various presented criteria are often not commensurable. To circumvent this issue, instead of combining several criteria into a trade-off objective, numerous other approaches propose to impose a priority order between the criteria. By combining them into (explicit or implicit) constrained optimization problems, the objective becomes to generate the closest counterfactual example (i.e. minimizing the penalty function) in a subspace defined by constraints on other criteria. This optimization subspace may for instance be defined by constraints on density [2, 54], or actionability [73]. On the contrary, other approaches such as Growing Spheres [34] and LORE [22] optimize the sparsity of the counterfactual explanation after optimizing its penalty.

Specifying this priority order is also connected to the proposed optimization process: due to the general model-agnostic (and sometimes data-agnostic) paradigm considered by these approaches, imposing a priority order between criteria also helps with the optimization of the objective function. Indeed the subspace satisfying the higher-order constraint can then be identified in a preprocessing step (e.g. the construction of a graph for FACE [54] or the generation of a neighborhood for LORE [22]).

3.2 The Underestimated Consequences of the Aggregation Step

While much discussion is generally proposed to motivate the choice of the desired explanation properties, on the other hand, the definition of the aggregation operator used to combine these criteria is, to the best of our knowledge, rarely defended. As mentioned earlier, numerous counterfactual approaches rely on heuristics, and as such do not even have an explicit optimization problem, leaving this aggregation to be implicit. Yet, we show in this section that in both cases, the resulting aggregation of these criteria also has important, potentially undesirable, consequences on the final explanation. The importance of this choice is underlined in Section 3.2.1, while Section 3.2.2 discusses the potential undersirable consequences of the aggregation.

3.2.1 The aggregation operator: undiscussed, yet impactful. Although it may seem obvious in a multicriteria optimization perspective that the aggregation operator directly impacts the nature of the solution, to the best of our knowledge this point is rarely discussed in explanation approaches. This may seem surprising, especially as some existing approaches propose different heuristics for the same groups of criteria, therefore essentially differing from one another in terms of how these are combined. For instance, Growing Spheres [34] and LORE [22] both optimize the closeness of the explanation (measured by the Euclidean distance) and its sparsity (measured with the l_0 distance).

More generally, from an optimization perspective, given two criteria that cannot been optimized simultaneously (as it is often the case for counterfactual explanations), the set of possible solutions would be the Pareto front, highlighting the diversity of the possible solutions (as considered by [14]). To illustrate this point, we conduct a simple experiment by generating multiple counterfactual explanations optimizing the same criteria with different aggregation operators. More precisely, we consider the 2-dimensional half-moons dataset[1] and the Boston dataset[2], on which two classifiers are trained: a SVM classifier for the half-moons dataset (0.99 in accuracy), and a Random Forest classifier for the Boston dataset (0.86 in accuracy). Figure 1 displays illustrations of the experiment: on the left, the half-moons dataset; on the right, a 2D representation of the Boston dataset using t-SNE [74]. In both cases, an instance x whose prediction is to be explained, represented by the green point, is randomly picked. We then consider two desirable properties for counterfactual examples: the closeness of the explanation, as measured by the Euclidean distance, and its belonging to a dense region

[1] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html
[2] https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

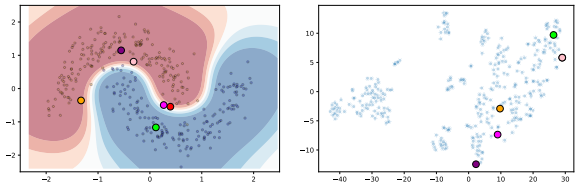

Figure 1: Illustration of the aggregation operator impact: the orange, red, pink and magenta points represent the counterfactual examples generated to explain the prediction associated to the instance of interest represented by the green point, when considering several types of aggregation (see details in the text). Left: half-moons dataset. Right: 2D t-SNE projection of the Boston dataset.

of its targeted class, as measured by the log-likelihood of the counterfactual example under a Gaussian Kernel Density Estimation (KDE) trained on the corresponding data. These two criteria are then combined using various aggregation operators found in the literature, and the resulting counterfactual examples shown on the figure in different colors: a weighted sum [26, 39, 45] (pink), the maximization of the closeness under density constraints [54] (magenta), the maximization of the density under closeness constraints [22, 34] (orange), and a maximisation of each criterion independently (red for the closeness, purple for the density). As expected, the resulting counterfactual examples are quite scattered across the dataset, covering regions of the feature space characterized by different decision boundaries. This illustrates the importance of the aggregation operator for counterfactual explanations.

3.2.2 Undesirable consequences. Beside potentially leading to drastically different results, the proposed choice of the aggregation operators sometimes raises questions in terms of relevance and user needs. For instance, optimizing for criteria which are mathematically at odds with one another may lead to trade-off solutions, that satisfy a bit of both, without satisfying fully any of them. This can be seen as especially problematic in the case when the properties involved belong to different property categories (see. Section 2.2), as they involve parameters that can hardly be compared (e.g. data vs. user preferences). It would then seem unlikely that users with non-technical background, who are often considered to be potential users of XAI methods, would understand that generated explanation does not satisfy all of the desired properties. For instance in the case of an aggregation performed with a weighted sum [39], users may not understand that the *causality* of the explanation is not guaranteed, let alone that it would come at the expense of more easiness of action (closeness of the explanation). The high number of possible desirable explanation objectives makes this issue all the more problematic.

As a conclusion, the criteria combination should be discussed just as much as which their individual selection. Considering the difficulty of aggregating numerous properties, forcing it through heuristics thus seems questionable. On the other hand, the variety of possible criteria, especially the ones where there is a tradeoff involved, pushes towards the generation of multiple explanations, that could then for instance focus on different properties.

4 GENERATING DIVERSE EXPLANATIONS

Contrary to the approaches discussed in the previous section, some other methods aim to provide users with multiple explanations at the same time to explain a single prediction. In this section, after providing some additional arguments for these approaches, we categorize existing works based on the diversity function they use to generate several explanations. We end the section with a discussion over these notions of diversity, showing how they help addressing some of the identified shortcomings of explainability approaches.

4.1 Additional Motivations and Discussion

Although a portion of existing works proposing diverse counterfactuals state some motivations for doing so, these efforts remain generally light and scattered. We seek in this section to provide a stronger and more in-depth case for multiple explanations, centered around two arguments that complement the one previously developed in Section 3.2: (i) results from social and cognitive sciences have proven the strong benefits of using multiple explanations to teach complex concepts; (ii) having multiple explanations can help overcoming one of the main shortcomings of machine learning interpretability, namely identifying and addressing unformulated user needs.

4.1.1 Insights from social sciences: More (carefully selected) explanations leads to better understanding. In various scientific fields, providing multiple explanations has long been identified as a key factor for a better understanding on complex concepts. For instance in a clinical context, Wang et al. [80] insist that multiple explanations are required for physicians to make better diagnostics. On the same note, stronger conclusions are drawn in the fields of education and psychology: when using analogies to teach complex concepts to medical students, providing a single explanation was shown to create a high risk of misconceptions [68]. On the other hand, providing multiple, carefully selected, analogies is presented as requirement for a good understanding. More generally, gathering and discussing some insights from several social sciences, Miller [41] insists that causes for an event must be seen as multiple, and that one important aspect of generating a good explanation is the selection by the user of his/her preferred explanation among a set of plausible ones. More recently, Bove et al. [6] empirically show the benefits of providing multiple explanations to users on a classification task: the latter leads to an increase both in terms of objective comprehension and subjective satisfaction.

4.1.2 Multiple explanations may help in overcoming a critical issue of interpretability: matching (unknown) user needs. One of the most crucial identified pain point of interpretability in general is the difficulty to determine user needs. Although it is commonly recognized that explanations should be adapted to those needs, as well as to the user's characteristics such as their knowledge and expertise (see e.g. [7] for explanations in general, and [18] for explanations for machine learning models), providing generic tools to do so is complex and constitutes a poorly covered task (some exceptions include for instance [16, 78]). In this regard, it seems illusory to hope for one explanation method to satisfy these undeclared user needs, especially as humans have been known to perceive feature

interactions and effects differently [20]. Generating several explanations and letting the user choose the most relevant one(s) to them seems, in this vein, a way to leave this 'extra-mile' task of mapping user needs to explanations to the user, i.e. have the user select the most suited explanation and discount others [25, 41, 70].

Although some of these reasons have also been identified by previous machine learning works, there still are few works in this direction. In the next section, we present them and discuss *how* they propose multiple explanations.

4.2 Existing Diverse Counterfactual Explanation Approaches

The arguments presented in Sections 3.2 and 4.1 have led several approaches to explain predictions through multiple counterfactual examples. For this purpose, most of them rely on the notion of *diversity* (see e.g. [45, 61]), imposing that the multiple explanations differ from one another, to avoid redundancies and to "propose various alternatives when user preferences are not known". Yet, mirroring the lack of consensus among desirable properties for explanations, this notion of diversity has been defined in various ways, generally with few discussions associated. In this section, we review these notions of diversity, discussing the existing literature of diverse counterfactual examples. These discussions are summarized in Table 2. The first three subsections detail in turn the three types of diversity we propose to distinguish, respectively named criteria, feature space and actions. Other criteria, related to the optimization procedure itself, are discussed in the fourth subsection.

4.2.1 Diversity in Criteria. A first type of diversity definition depends on the quality criteria the counterfactual examples are required to optimize: it proposes to use different means to combine them, often relying on different aggregation operators instead of a single one. For example, Dandl et al. [14] and Rasouli et al [56] focus on generating multiple counterfactual examples, which all optimise the same criteria but perform different trade-offs between them. The generated counterfactual examples thus correspond to different positions on the Pareto front defined by the considered quality criteria, then chosen according to different strategies. To do so, Rasouli et al. [56] consider that a hierarchy among the different criteria is provided by the user. This avoids the risk induced by a trade-off operator that may lead to a solution that actually has a medium value for all considered criteria. Indeed, a user-defined hierarchy allows to select the criterion to be optimized first and those to be optimized later on.

4.2.2 Diversity in the Feature Space. A second type of diversity focuses on the relative position of the generated counterfactual examples in the feature space. According to this definition, diverse counterfactual examples lie far away from each other the feature space space. Here, the quality criteria are therefore not only used to evaluate the counterfactual examples individually, but also to analyze the relations between them. In this category, two strategies can be identified: the former induces diversity by first explicitly defining different constraints, and then generating counterfactual examples for each of them; the latter relies on defining diversity as a distance to be maximized between the generated counterfactual examples in the feature space.

The constraints considered by the approaches relying on the first strategy can take various forms. Most commonly [11, 47, 60, 72], they are defined to partition the feature space into several subspaces. Counterfactual examples are then generated in each of the identified subspaces, which allows to obtain explanations that are diverse as they belong to different areas of the space. For instance, [72] relies on the user defining this partitioning along features of interest (e.g., age group) and generating one counterfactual per subspace.

Approaches relying on the second strategy define diversity as a similarity or distance between counterfactual examples. This allows them to directly integrate diversity in the counterfactual generation. This can be done either by modifying the optimization problem by integrating a diversity criterion into the cost function, enabling them to a set of counterfactual examples at once: Equation 1 is then modified to:

$$\{e_1^*, \dots, e_k^*\} = \underset{\{e_1,\dots,e_k\}\subset\mathcal{E}}{\operatorname{argmin}} \; agg\left(\sum_{i=1}^{k} pen_x(e_i), \varphi(div(\{e_1, \dots, e_k\}))\right) \tag{4}$$

where k denotes the number of desired counterfactual examples, div is a function assessing their diversity, seen as a new quality criterion to be maximized, φ a decreasing function and agg an aggregation operator to combine the average quality of the counterfactual candidates and their diversity. The penalty function can obviously be combined with some of the additional criteria discussed in Section 2.2, such as sparsity, data or user contextualization.

This diversity measure itself can take multiple forms. For example, several approaches focus on maximizing the diversity of features used in the final explanations [4, 60, 61], which can be translated as maximizing the l_0 distance between the proposed counterfactual examples. Other approaches, such as [45], define diversity as the distance between the generated counterfactual examples (norm l_1, l_2, or both). As a result, the obtained counterfactual examples are thus distant from one another in the input space, and may use different features.

Instead of defining a set of diverse counterfactual explanations as a solution to a single optimisation problem, other methods [24, 42, 61] rely on an iterative process to generate the multiple explanations, a counterfactual example being generated at each step. To ensure that a new explanation is different from the previous ones, these approaches then consider constraints on the distance between the new explanation and the ones already generated.

In the case of non-binary classification, Ley et al. [36] propose to take into account, in addition to the feature space, the prediction space: they generate counterfactual examples associated with various classes among the ones different from the one predicted for the instance of interest.

4.2.3 Diversity in Actions. A counterfactual explanation by design suggests actions, as modifications to the instance of interest, that allow to get a different prediction. A third type of diversity aims at proposing explanations which need/use different actions. They are related to the diversity in terms of features discussed above, with a slightly different interpretation, more related to the notion of algorithmic recourse. Beside relying on the l_0 distance, this type of diversity can be achieved in more specific settings: Guidotti et

Method	Diversity		Counterfactual Search		
	Diversity type	Diversity criterion	Number of CF	Explicit	Single run
LORE [22]	Actions	Diverse leaves of a decision tree	algo	Yes	Yes
Mahajan [39]	Feature values	Stochasticity in the generation	user	No	Yes
Russell [61]	Actions	Rerun & exclusion of the previous results	user+algo	Yes	No
CADEX [44]	Feature values	Rerun & exclusion of the previous features used	user	No	No
Ustun [73]	Actions	Rerun & exclusion of the previous results	user	Yes	No
CERTIFAI[65]	Feature values	Exploration by sampling	user	No	Yes
Tsirtsis [72]	Feature values	Partitioning of the data space	user	Yes	Yes
MOC [14]	Feature values & Criteria	Pareto front in objective space	algo	Yes	Yes
MACE1 [28]	Features values	Rerun & exclusion of the previous results	user	Yes	No
DICE [45]	Feature values	Diversity term in the optimization	user	Yes	Yes
DECE [12]	Feature values	Consideration of different constraints	user	Yes	Yes
CRUDS [17]	Feature values	Partitioning of the data space	user	Yes	Yes
DiVE [60]	Feature values	Diversity term in the optimization	user+algo	Yes	No
OCEAN[49]	Feature values	Rerun & exclusion of the previous results	user+algo	Yes	No
MCCE [58]	Feature values	Exploration by sampling	user	Yes	Yes
OrdCE[27]	Criteria	Pareto front in objective space	user	Yes	No
MCS [82]	Features values	Exploration by sampling	user	No	No
CSCF [46]	Actions	Sequential approach to obtain CF with different sequences	user	Yes	Yes
MIP-DIVERSE [42]	Feature values	Rerun & exclusion of the previous results	user	Yes	No
Hada [24]	Feature values	Consideration of different constraints	user + algo	No	No
Navas [47]	Feature values	Consideration of different constraints	user	Yes	Yes
Samoilescu [63]	Feature values	Partitioning of the data space	user	Yes	Yes
Becker [3]	Actions	Diverse leaves of a decision tree	user + algo	No	Yes
GeCo [64]	Feature values	Exploration by sampling	user	No	Yes
FastAR [77]	Actions	Stochasticity in the generation	user	No	Yes
Carreira [11]	Feature values	Consideration of different constraints	user	No	No
EMC [81]	Criteria	Initialisation of different cost functions	user	Yes	Yes
δ-CLUE [36]	Feature values	Diverse initialization for the optimization	user	No	No
CARE1 [56]	Criteria	Pareto front in objective space	user	Yes	Yes
MACE2 [83]	Feature values	Exploration by sampling	user	Yes	Yes
Smyth [67]	Feature values	k-NN model to delimit group of CF	user	Yes	Yes
COPA [8]	Feature values	Optimization using gradient descent	user	Yes	Yes
FRPD [48]	Feature values	Diversity term in the optimisation	user	Yes	Yes

Table 2: Summary of existing diverse counterfactual example (CF) generation methods, discussed in Section 4.2 that details the diversity type and criterion columns. The "Number of CF" column indicates whether the user can choose the number of desired counterfactual examples (user) or if the latter is automatically selected by the algorithm (algo). The "explicit" column indicates whether the diversity objective is explicitly included in the optimisation process or not. The "Single-step" column indicates whether the optimization applies a single-step (Yes) or an iterative procedure (No) to generate all the counterfactual examples.

al. [22] and Becker et al. [3] use decision trees to generate explanations. Imposing the latter to be located in different leaves of the tree implies they follow different paths from the root to the leaves, and as a consequence rely on different actions.

Instead of proposing explanations that modify different features, Russell et al. [61] propose explanations that go in different modification directions: a first counterfactual example may recommend increasing the value of a given feature, whereas another one would recommend decreasing it. The induced actions are completely different. The same principle applies to the proposition of Verma et al. [77], that relies on performing successive small actions, in different directions, until obtaining the final explanation.

4.2.4 Optimisation-related dimensions. Beside relying on different definitions of the notion of diversity, that apply at different levels, as discussed in the previous paragraphs, existing algorithms to generate multiple counterfactual examples also differ in the optimization procedure they apply, as we propose to discuss in this section. These comparison dimensions are summarised in the last three columns of Table 2.

Explicit vs non-explicit diversity. Independently from the discussions of the previous sections describing how diversity may being defined, another possibility to differentiating factor for methods is associated to how explicitly diversity is incorporated in the counterfactual generation. We thus make a distinction between *explicit* and *non-explicit* methods, and discuss them below.

Explicit methods are characterized by the fact that they actively take into account diversity in the counterfactual counterfactual optimisation problem defined in Equation 1. This can be achieved with all diversity definitions, in various ways. A straightforward way is to include the notion of diversity directly in the cost function, so

that the optimization of diversity is guaranteed between the counterfactuals like Mothilal et al. [45] or Dandl et al. [14]. The former propose to integrate it by defining the diversity of a set of solutions while the latter integrate it through the aggregation function that combines the different criteria. Other approaches focus on solving simultaneously different optimization problems with their own constraints, such as [11, 24]. In this case, the information considered as input is not the same for each optimization problem, allowing to obtain explanations that answer different contexts or motivations and are thus diverse. Finally, some approaches propose to integrate diversity by excluding the explanations already generated from the set of possible solutions for later iterations, thus ensuring that the new explanations are different from the previous ones [49, 61]. Thus, for explicit methods, there is a dedicated mechanism in the explanation generation process that ensures diversity (regardless of its definition).

On the other hand, non-explicit approaches are generally non-deterministic approaches, meaning that using the same approach twice in the same setting does not necessarily return the same explanation. Non-explicit counterfactual methods such as [39, 44, 65] thus propose to generate diverse explanations by using the stochastic aspect of the generation process. Unfortunately, this means that the resulting diversity is not maximized nor guaranteed, as the set of generated examples may be close from one another for instance. Although some mechanisms are proposed to encourage diversity, such as in CERTIFAI [65] where the authors propose to modify the algorithm initialization, most of these remain unreliable when it comes to diversity.

Number of counterfactual examples returned. Another dimension related to diversity is the number of counterfactual examples that these methods allow to generate. Although most approaches propose to let the user set this number (e.g., among others [39, 45]), this choice may in some cases be limited by the method itself. For instance, for methods proposing to generate diverse explanations as examples belonging to different leaves of a tree [3, 22], the number of counterfactual examples to be generated is bounded by the number of leaves. However, despite the fact that the approaches that let the user set the number of desired counterfactual explanations seem to be less limited, they generally do not acknowledge that this number naturally often comes at odds with how diverse the explanations are. For many them (e.g. [45, 65, 73]), increasing this number will thus lead to the generation of redundant explanations.

One run vs. several runs. To generate multiple explanations, two strategies exist: either all explanations are generated as solution of a single optimisation problem; or the optimisation problem only leads to the generation of a single counterfactual example, in which case several steps are required to generate iteratively additional explanations. Methods generating multiple explanations in a single step, referred to as "Single run" in Table 2, often optimize a cost function applying to sets of candidates, as given in Equation (4). The explanation diversity between explanations is then one of the criteria included in the considered cost function. Other methods focus on simultaneously exploring the data space in different directions (such as [22, 72]).

On the other hand, other approaches [36, 61] generate explanations iteratively, possibly using the explanations obtained in the previous steps to generate the new one. Other approaches proposed by Carreira et al. [11] or Samoilescu et al. [63] look at different constraints at each stage to address different contexts. For example, Carreira et al. increase the number of considered constraints at each stage. Thus, at each step the algorithm is defined with different constraints. The studied optimization problem is then different for each counterfactual example.

5 DISCUSSION AND RESEARCH CHALLENGES

The richness of the diverse counterfactual approaches presented in the previous section underlines once again the importance of properly motivating the choice of the explanability objectives, including the diversity. In this section, we propose a discussion on the link between the diversity notions identified, and how they may better help with addressing user needs.

5.1 Diversity as a Way to Match Unknown User Needs

Section 4.1 reminds one of the strongest arguments supporting the use of multiple counterfactual explanations: their ability to help match unobserved user needs. An implicit assumption for this purpose is the explanations had to be *diverse* [70], leading to concurrent definitions of diversity, described in the previous section. Yet, we argue that all diversity definitions do not all unequivocally fulfill this promise.

Intuitively, *Diversity in Criteria*, described in Section 4.2.1, directly addresses this objective: by proposing multiple ways of combining different quality criteria, they allow the user to choose his/her own order of preference between the explanations' properties. For example, asking a user to specify the minimum level of sparsity of the explanation he wants for a problem might be complicated for him. Offering explanations with different levels of sparsity and penalty could help him to understand the trade-off between these two notions in the explained prediction and to select his preferred aggregation. Yet, this requires the user to be able to understand, if not the property captured, at the very least that the proposed diverse explanations vary along these criteria. This therefore questions the use of the notion of diversity in criteria for data-contextualization criteria (cf. Section 2.2), which are arguably (i) much harder for a user to understand and (ii) not directly visible when looking at presented counterfactual explanations, at least without additional context. User-contextualization criteria, on the other hand, do not suffer from this issue. By nature, although the numerical quantification of these properties can be challenged, it is expected that the user directly understands the differences. This leads us to believe that to be relevant, the diversity of a set of counterfactual explanations should be observable. This naturally questions the utility of approaches integrating a *non-explicit* diversity, as they do not guarantee the resolution of the problem.

By proposing explanations that provide a set of actions that vary in terms of targeted features, approaches integrating *Diversity in Actions* fulfill this objective of observable differences. The user is provided several alternative recourses, among which he/she may choose their preferred according to their internal unobserved preferences. On the other hand, *Diversity in Feature values* do not explicitly address a formulated user need. Apart from being a proxy

for Diversity in Actions when user-contextualization criteria are not available, they thus do not seem to answer user expectations, despite being the most represented type of approach (see Table 2). A notable exception would be in the context of model debugging, where such diversity might be providing relevant information about the local decision boundary and feature importance weights.

5.2 Diversity Beyond Counterfactual Explanations

On the contrary to most of the presented quality criteria, the penalty of the explanations is rarely included as a property to balance. Due to the usual formulation of counterfactual explanations, penalty is indeed generally seen as a criterion to minimize, or eventually that it is possible to sacrifice a little bit, only when necessary, to satisfy other properties. Few works thus explore the possibility of explicitly combining local counterfactual explanations with more global ones (a concept proposed for instance by [57]). Yet, several works have showcased in applied contexts the benefits of combining local and global explanations on interpretability. Such works include for instance [13], mixing local and global feature importances to explain fraud detection models, or [5], proposing local explanations with global contextualization to help customers understand insurance pricing. Integrating penalty as one criteria to balance among others for counterfactual explanations thus represents an interesting perspective. As illustrated by the works mentioned above, this seems strongly related to another notion of diversity that could be formulated, also out of the scope of this paper, which is the diversity in explanations forms.

6 CONCLUSION

In this work, we show how the combination of the quality criteria considered to generate counterfactual explanations is not trivial and may inaccurately match the needs of the user, which are often unobserved and only implicitly defined. We argue this provides new arguments in favour of approaches that propose diverse counterfactual explanation as a solution to this issue, letting the users choose their preferred explanation. As discussed in the paper, various definitions of diversity have been proposed in the literature, the conducted analysis shows that they do not equally help addressing user needs. Future works will aim at conducting an empirical study to complement this analysis, in a full human-in-the-loop paradigm, so as to examine thoroughly the impacts these diversity definitions can have on users and the respective cases where they might appear more appropriate. This direction of research also calls for new tools to represent and model user needs as well as user knowledge, so as to establish a correspondence with the most relevant approaches.

ACKNOWLEDGMENTS

This research was supported by TRAIL, a joint laboratory between SORBONNE UNIVERSITE/CNRS (LIP6) and AXA.

REFERENCES

[1] André Artelt and Barbara Hammer. 2019. On the computation of counterfactual explanations–A survey. *arXiv:1911.07749* (2019).
[2] André Artelt and Barbara Hammer. 2020. Convex Density Constraints for Computing Plausible Counterfactual Explanations. In *Artificial Neural Networks and Machine Learning*. 353–365.
[3] Maximilian Becker, Nadia Burkart, Pascal Birnstill, and Jürgen Beyerer. 2021. A Step Towards Global Counterfactual Explanations: Approximating the Feature Space Through Hierarchical Division and Graph Search. *Adv. Artif. Intell. Mach. Learn.* 1, 2 (2021), 90–110.
[4] Umang Bhatt, Isabel Chien, Muhammad Bilal Zafar, and Adrian Weller. 2021. DIVINE: Diverse Influential Training Points for Data Visualization and Model Refinement. *arXiv preprint arXiv:2107.05978* (2021).
[5] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *27th International Conference on Intelligent User Interfaces, IUI*. 807–819.
[6] Clara Bove, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2023. Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study. In *28th International Conference on Intelligent User Interfaces, IUI*.
[7] Garvin Brod, Markus Werkle-Bergner, and Yee Lee Shing. 2013. The influence of prior knowledge on memory: a developmental cognitive neuroscience perspective. *Frontiers in behavioral neuroscience* 7 (2013), 139.
[8] Ngoc Bui, Duy Nguyen, and Viet Nguyen. 2022. Counterfactual Plans under Distributional Ambiguity. *International Conference on Learning Representations* (2022).
[9] Nadia Burkart and Marco F. Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.
[10] Thomas Calvo, Gaspar Mayor, and Radko Mesiar (Eds.). 2002. *Aggregation Operators: New Trends and Applications*. Vol. 97. Springer.
[11] Miguel Á Carreira-Perpiñán and Suryabhan Singh Hada. 2021. Counterfactual explanations for oblique decision trees: Exact, efficient algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6903–6911.
[12] Furui Cheng, Yao Ming, and Huamin Qu. 2020. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1438–1447.
[13] Dac Collaris, Leo M. Vink, and Jarke J. van Wijk. 2018. Instance-level explanations for fraud detection: a case study. In *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*.
[14] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI*. Springer International Publishing.
[15] Marcin Detyniecki. 2001. Fundamentals on aggregation operators. *Thesis manuscript, Université Pierre et Marie Curie* (2001).
[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608* (2017).
[17] Michael Downs, Jonathan L. Chu, Yaniv Yacoby, Finale Doshi-Velez, and Pan WeiWei. 2020. CRUDS: Counterfactual Recourse Using Disentangled Subspaces. *ICML Workshop on Human Interpretability in Machine Learning* (2020), 1–23.
[18] David Gentile, Greg A. Jamieson, and Birsen Domez. 2021. Evaluating human understanding in XAI systems. In *ACM CHI XCXAI Workshop*.
[19] Michel Grabisch, Jean-Luc Marichal, Radko Mesiar, and Endre Pap. 2009. *Aggregation Functions*. Number 127 in Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press.
[20] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. 903–912. https://doi.org/10.1145/3178876.3186138
[21] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (04 2022), 1–55.
[22] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.
[23] 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 (2018), 42 pages.
[24] Suryabhan Singh Hada and Miguel Á Carreira-Perpiñán. 2021. Exploring counterfactual explanations for classification and regression trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 489–504.
[25] Denis J. Hilton. 1996. Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308. https://doi.org/10.1080/135467896394447
[26] Adulam Jeyasothy, Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2022. Integrating Prior Knowledge in Post-hoc Explanations. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'2022)*.
[27] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike, Kento Uemura, and Hiroki Arimura. 2021. Ordered counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*,

Vol. 35. 11564–11574.
[28] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 895–905.
[29] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)* (2022).
[30] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 353–362.
[31] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219
[32] Michael T. Lash, Qihang Lin, Nick Street, Jennifer G. Robinson, and Jeffrey Ohlmann. 2017. Generalized Inverse Classification. In *Proc. of the SIAM Int. Conf. on Data Mining*. 162–170.
[33] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arXiv:1906.04774* (2019).
[34] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2018. Comparison-based Inverse Classification for Interpretability in Machine Learning. In *Proc. of Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 100–111.
[35] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2801–2807.
[36] Dan Ley, Umang Bhatt, and Adrian Weller. 2022. Diverse, global and amortised counterfactual explanations for uncertainty estimates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7390–7398.
[37] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (2021).
[38] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*. 4768–4777.
[39] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *NeurIPS workshop* (2019).
[40] Raphael Mazzine and David Martens. 2021. A Framework and Benchmarking Study for Counterfactual Generating Methods on Tabular Data. *Applied Science* (2021).
[41] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[42] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. 2021. Scaling guarantees for nearest counterfactual explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 177–187.
[43] Christoph Molnar. 2022. *Interpretable Machine Learning* (2 ed.). https://christophm.github.io/interpretable-ml-book
[44] Jonathan Moore, Nils Hammerla, and Chris Watkins. 2019. Explaining deep learning models with constrained adversarial examples. In *Pacific Rim international conference on artificial intelligence*. Springer, 43–56.
[45] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency*. ACM.
[46] Philip Naumann and Eirini Ntoutsi. 2021. Consequence-aware sequential counterfactual generation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 682–698.
[47] Guillermo Navas-Palencia. 2021. Optimal counterfactual explanations for scorecard modelling. *arXiv preprint arXiv:2104.08619* (2021).
[48] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. 2023. Feasible Recourse Plan via Diverse Interpolation. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4679–4698.
[49] Axel Parmentier and Thibaut Vidal. 2021. Optimal counterfactual explanations in tree ensembles. In *International Conference on Machine Learning*. PMLR, 8422–8431.
[50] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 2021. CARLA: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms. *NeuRIPS 2021 Track on Datasets and Benchmarks* (2021).
[51] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*. 3126–3132.
[52] Martin Pawelczyk, Tobias Leemann, Asia Biega, and Gjergji Kasneci. 2022. On the Trade-Off between Actionable Explanations and the Right to be Forgotten. *arXiv preprint arXiv:2208.14137* (2022).
[53] Judea Pearl. 2009. Causality. *Cambridge University Press* (2009).
[54] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society*.
[55] Goutham Ramakrishnan, Yun Chan Lee, and Aws Albarghouthi. 2020. Synthesizing action sequences for modifying model decisions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5462–5469.
[56] Peyman Rasouli and Ingrid Chieh Yu. 2022. CARE: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics* (2022), 1–26.
[57] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. *Advances in Neural Information Processing Systems* 33 (2020), 12187–12198.
[58] Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. 2021. MCCE: Monte Carlo sampling of realistic counterfactual explanations. *arXiv preprint arXiv:2111.09790* (2021).
[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining*. 1135–1144.
[60] Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. 2021. Beyond Trivial Counterfactual Explanations With Diverse Valuable Explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1056–1065.
[61] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.
[62] Waddah Saeed and Christian Omlin. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* 263 (2023), 110273.
[63] Robert-Florian Samoilescu, Arnaud Van Looveren, and Janis Klaise. 2021. Model-agnostic and Scalable Counterfactual Explanations via Reinforcement Learning. *arXiv preprint arXiv:2106.02597* (2021).
[64] Maximilian Schleich, Zixuan Geng, Yihong Zhang, and Dan Suciu. 2021. GeCo: Quality Counterfactual Explanations in Real Time. *Proc. VLDB Endow.* 14, 9 (oct 2021), 1681–1693.
[65] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, 166–172.
[66] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. 2021. Counterfactual explanations can be manipulated. *Advances in neural information processing systems* 34 (2021), 62–75.
[67] Barry Smyth and Mark T Keane. 2022. A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*. Springer, 18–32.
[68] Rand J Spiro, Paul J Feltovich, Richard L Coulson, and Daniel K Anderson. 1989. Multiple analogies for complex concepts: antidotes for analogy-induced misconception in advanced knowledge acquisition. In *Similarity and analogical reasoning*. 498–531.
[69] Ramya Srinivasan and Ajay Chander. 2020. Explanation Perspectives from the Cognitive Sciences—A Survey. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. 4812–4818.
[70] Emily Sullivan and Philippe Verreault-Julien. 2022. From Explanation to Recommendation: Ethical Standards for Algorithmic Recourse. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 712–722.
[71] Erico Tjoa and Cuntai Guan. 2020. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* (2020).
[72] Stratis Tsirtsis and Manuel Gomez Rodriguez. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems* 33 (2020), 16749–16760.
[73] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proc. of the Conf. on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 10–19.
[74] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[75] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable Counterfactual Explanations Guided by Prototypes. In *Proc. of European Conf. on Machine Learning*.
[76] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).
[77] Sahil Verma, Keegan Hines, and John P Dickerson. 2021. Amortized Generation of Sequential Counterfactual Explanations for Black-box Models. *arXiv preprint arXiv:2106.03962* (2021).
[78] Tom Vermeire, Thibault Laugel, Xavier Renard, David Martens, and Marcin Detyniecki. 2022. How to choose an explainability method? towards a methodical implementation of XAI in practice. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*. Springer, 521–533.

[79] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard journal of law & technology* 31 (2018), 841–887.

[80] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[81] Prateek Yadav, Peter Hase, and Mohit Bansal. 2021. Low-cost algorithmic recourse for users with uncertain cost functions. *arXiv preprint arXiv:2111.01235* (2021).

[82] Fan Yang, Sahan Suresh Alva, Jiahao Chen, and Xia Hu. 2021. Model-based counterfactual synthesizer for interpretation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1964–1974.

[83] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. 2022. MACE: An Efficient Model-Agnostic Framework for Counterfactual Explanation. *arXiv preprint arXiv:2205.15540* (2022).