TOP
https://doi.org/10.1007/s11750-024-00670-2

ORIGINAL PAPER

Disagreement amongst counterfactual explanations: how transparency can be misleading

Dieter Brughmans[1] · **Lissa Melis**[2,3] · **David Martens**[1]

Received: 18 April 2023 / Accepted: 23 February 2024

Abstract

Counterfactual explanations are increasingly used as an Explainable Artificial Intelligence (XAI) technique to provide stakeholders of complex machine learning algorithms with explanations for data-driven decisions. The popularity of counterfactual explanations resulted in a boom in the algorithms generating them. However, not every algorithm creates uniform explanations for the same instance. Even though in some contexts multiple possible explanations are beneficial, there are circumstances where diversity amongst counterfactual explanations results in a potential disagreement problem among stakeholders. Ethical issues arise when for example, malicious agents use this diversity to fairwash an unfair machine learning model by hiding sensitive features. As legislators worldwide tend to start including the right to explanations for data-driven, high-stakes decisions in their policies, these ethical issues should be understood and addressed. Our literature review on the disagreement problem in XAI reveals that this problem has never been empirically assessed for counterfactual explanations. Therefore, in this work, we conduct a large-scale empirical analysis, on 40 data sets, using 12 explanation-generating methods, for two black-box models, yielding over 192,000 explanations. Our study finds alarmingly high disagreement levels between the methods tested. A malicious user is able to both exclude and include desired features when multiple counterfactual explanations are available. This disagreement seems to be driven mainly by the data set characteristics and the type of counterfactual algorithm. XAI centers on the transparency of algorithmic decision-making, but our analysis advocates for transparency about this self-proclaimed transparency.

Keywords XAI · Counterfactual explanations · Machine learning · Disagreement problem

Mathematics Subject Classification 68T27

Dieter Brughmans and Lissa Melis have contributed equally to this work.

Extended author information available on the last page of the article

Published online: 08 May 2024

1 Introduction

Artificial Intelligence (AI) or Machine Learning (ML) is rapidly evolving and disrupting various sectors, such as finance, healthcare, business (e.g., logistics, the labor market), education, and urban development. Besides the many benefits AI can create, multiple negative implications can be identified for each sector (Păvăloaia and Necula 2023). One of the re-occurring challenges concerning AI is the need for transparency: many AI models are opaque and operate on a black-box basis, which makes it difficult—or sometimes impossible—to interpret and explain a decision that has been made. Therefore, Explainable Artificial Intelligence (XAI) has recently emerged as a much-needed research field. Next to an obvious focus on the predictability of AI models, model explainability is necessary for users, developers, and other stakeholders of real-life AI applications. Not only do people generally want to know an explanation for an algorithm-based decision, but also legislation is backing up this need. For example, in 2018 the European Union stated in the new General Data Protection Regulation (GDPR) that subjects of algorithmic decision-making are entitled to insights about the logic involved. Users can ask for explanations of data-driven decisions that significantly influence their lives (Goodman and Flaxman 2017). These are categorised under high-risk applications of AI, such as credit scoring and employment services. People want, and are entitled to, an answer to why their loan is denied or why they are not hired for a job.

Reaching a certain level of explainability in AI models is possible by either developing models that are inherently more interpretable - but sometimes have less predictive power—or by using post-hoc XAI techniques to generate explanations after predictions have been made with a black-box model. Even though seemingly good explanations for a model's decision can be generated by the use of a post-hoc XAI method, and consequently the model and its decisions are qualified as transparent, research on the *uniformity* of these explanations is rather scarce. Many different post-hoc XAI methods exist and each method can generate different explanations for the same predicted outcome. Ergo, different stakeholders might be more interested in the explanations of one specific XAI method over another one. This raises the question of whether the transparency objective of XAI is achieved. In the literature, this phenomenon has been recently called the *disagreement problem* (Krishna et al. 2022; Neely et al. 2021; Roy et al. 2022).

Miller (2019) takes knowledge from psychology, sociology, and cognitive sciences to identify what are "good" explanations. They argue that explanations are contrastive, selected, social and that probabilities most likely will not matter. The first means that people generally don't ask why a certain decision is made. People wonder why a certain decision is made *instead* of another one. The second points to the fact that even though multiple explanations are possible to justify a decision, people are used to selecting one or two causes as *the* explanation. The third means an explanation is always dependent on the beliefs of the user and the last refers to the preference of *causes* over a probability or statistical relationship. These insights stress the usefulness of *counterfactual (CF) explanations*,

a post-hoc example-based XAI method which underlines a set of features that, when changed, alter a decision made by a model (Arrieta et al. 2020).

Evaluating the quality of counterfactual explanations in varying contexts and for different users is complex, leading to diverse counterfactual algorithms (Verma et al. 2020; Guidotti 2022). Similar to other types of post-hoc XAI explanations, limited research on the consistency of counterfactual explanations reveals a risk: disagreeing counterfactual explanations could lead to ethical issues and transparency concerns in XAI, especially if one party controls explanation selection.

Our research focuses on the disagreement problem in popular counterfactual (CF) explanation methods, an area that remains less explored compared to feature importance explanations in Explainable AI (XAI). Two primary reasons highlight the unique nature of the disagreement problem in counterfactual explanations. First, counterfactuals explain *decisions*, while feature importance explanations explain *prediction scores* (Fernández-Loría et al. 2020). This difference directly relates to the contexts where the disagreement problem is most critical: those of the high-risk scenarios as outlined by the AI Act, where the emphasis is on understanding decisions—such as denial of credit, job rejections, or medical diagnoses—rather than on interpreting prediction scores. Second, unlike feature importance methods that include all features, counterfactual explanations often focus on a small, selective subset of features. This selectiveness could introduce bias in the explanations. For example, explanations with specific features could be chosen to hide the fact a model is based on unethical features (see an elaborated example in Sect. 3). Our study aims to analyze the disagreement problem in counterfactual explanation algorithms, considering their unique challenges and significant role in ensuring ethical and transparent use of AI.

It's noteworthy that, although the disagreement problem has received attention in the context of feature importance methodologies, a comprehensive quantification of this challenge within counterfactual algorithms remains unaddressed. Recognizing the considerable potential for misuse, our research aims to bridge this critical gap in the literature by conducting a comprehensive analysis of the disagreement problem amongst counterfactual explanation algorithms.

In Sect. 2.3.1, we will situate counterfactual explanations in the diverse landscape of post-hoc XAI techniques and express how a lack of consistent evaluation methods for these techniques can lead to ambiguity in their explanations and ethical consequences. Consequently, in Sect. 3, we will quantify the disagreement amongst ten different counterfactual explanation methods next to Anchor and SHAP. Section 4 discusses the disagreement problem for counterfactual explanations and addresses potential ways to deal with the problem. The paper ends with conclusions and future research in Sect. 5.

2 The diverse landscape of post-hoc explanations

Post-hoc explanation methods are a subcategory of XAI that is concerned with explaining decisions made by complex black-box models, after these models have been trained. In contrast to intrinsic explanation methods, they do not try to create

interpretable white-box models, but are focused on explaining existing complex models (Linardatos et al. 2020). These methods are particularly interesting because their explanations seem to bypass the accuracy-explainability trade-off (Huysmans et al. 2006). This is a paradox stating that model performance often comes at a cost of model interpretability. Nonetheless, these post-hoc methods are able to explain complex models and thus theoretically achieve both high performance and explainability at the same time. However, the quality of post-hoc explanation has often been a point of discussion (Fernández-Loría et al. 2020; Doshi-Velez and Kim 2017).

Because these explanation methods are applied to models that are not intrinsically explainable, it is difficult to assess the quality of such explanations. The field of XAI evaluation has come up with different metrics to quantify this quality, however, no consensus has currently been reached. Since we cannot strictly quantify the quality of a post-hoc explanation method, many methods are proposed and used. This has led to ambiguity amongst explanations: explanations for the same instance are different depending on the post-hoc explanation method used (the disagreement problem). The quantified lack of uniformity in explanations has already been investigated for several post-hoc explanation methods (Krishna et al. 2022; Neely et al. 2021; Roy et al. 2022), however, to the best of our knowledge, this problem has not yet been investigated for counterfactual explanations, which is the main contribution of this work.

We first give an overview and classification of the post-hoc explanation methods used for comparison in this work in Sect. 2.1. In Sect. 2.2, we discuss how post-hoc explanation methods are currently evaluated and address some core issues regarding this topic. Lastly, Sect. 2.3 elaborates on the existing research on the disagreement problem and the need to apply this research to counterfactual explanations.

2.1 Counterfactual explanations and recent post-hoc explanation methods

The most popular post-hoc explanation methods can be divided into two groups: feature-based techniques (also called attribution methods) and example-based (also called instance-based) techniques (Dwivedi et al. 2023; Molnar 2018). The first group contains methods like local interpretable model-agnostic explanations (LIME), Shapley additive explanations (SHAP), and other feature importance techniques. The second group, example-based post-hoc explanation methods, contains Anchors and counterfactuals. We will briefly explain the methods used in our experiments: SHAP, Anchors, and counterfactual explanations. Figure 1 provides a figurative example of the different XAI methods to explain why a person (the instance) is predicted not to get their loan approved. For a more detailed description and examples, we refer to Molnar (2018) and the works mentioned below.

(LIME and) SHAP

SHAP (Lundberg and Lee 2017) and LIME (Ribeiro et al. 2016) are similar in the sense that the impact of a certain feature is measured related to the predictive outcome. The basic idea of LIME is to sample instances in the neighborhood of the

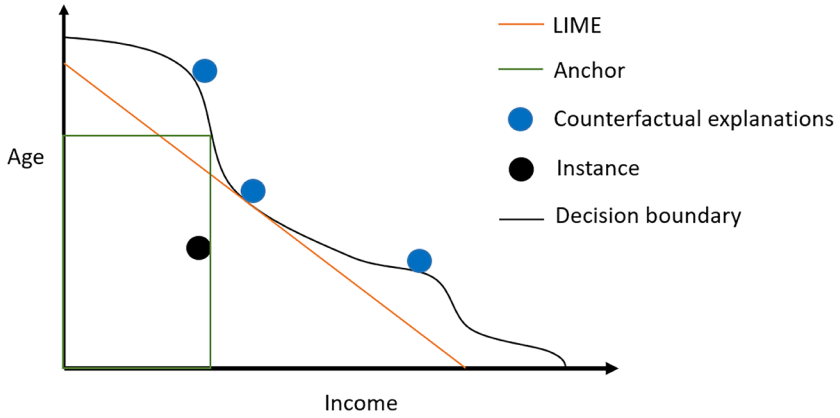


Fig. 1 Figurative toy example of LIME, Anchors and counterfactual explanations for a loan approval predictive model, based on Brughmans et al. (2023); Molnar (2018) and Ribeiro et al. (2016)

instance that is given to the prediction model and then train an interpretable model like linear regression or decision tree to explain this neighborhood. The interpretable model can consequently be used to explain the prediction that is made by the actual, black-box model. For tabular data, which are used in the experiments of this work, the issue is to define the neighborhood of an instance. If LIME would only sample closely around the given instance, chances are high all predictions will be exactly the same and LIME cannot comprehend how predictions change. Therefore, samples are taken broadly, e.g., by using a normal distribution. A major disadvantage of LIME is that the explanations differ depending on the samples used, which makes the explanations unstable and manipulable. Therefore, we do not include LIME in our comparisons.

A better feature-based technique can be found in SHAP, which combines the locality of LIME with the concept of Shapley values from coalitional or cooperative game theory. The contribution of each feature (player) to the prediction that is made by the model (outcome of the game) for a given instance is calculated. Moreover, the contribution of cooperation between players (multiple features) is examined. The average marginal contribution of a feature value across all cooperations is called the shapley value.[1] Because there are 2^k possible cooperations, for which models need to be trained, calculating all the Shapley values is computationally expensive. Therefore, by using the LIME-inspired sampling, the SHAP algorithm decreases the computation time.

[1] A common misinterpretation of the Shapley value is that it amounts to the difference in prediction after removing the feature from the model training.

Anchors

Anchors or scoped rules (Ribeiro et al. 2018) are high-precision easy-to-understand if-then rules. They portray feature conditions together with a predictive outcome. The rules are called Anchors because any changes to other features than the ones mentioned, will not result in another prediction. In contrast to e.g., LIME, Anchors will provide a region of instances to describe the model's behavior. They are consequently less instance-specific. For example, imagine if a person applies for a loan at a bank. This person is 50 years old, has a monthly income of $2000, his gender is male and he currently has $5000 in debt. A model has predicted that the loan application should be declined. The corresponding Anchor could then be: if the monthly income is lower than $5000 and the age is higher than 35, then predict that the loan application would be declined.

Counterfactual explanations

Counterfactual explanations describe a combination of feature changes that would alter the predicted class (Martens and Provost 2014). In other words, they determine what features should change to change the prediction and are consequently sometimes called what-if statements. As mentioned in Sect. 1, this type of explanation is especially human-friendly because they are contrastive and selective (Miller 2019). Counterfactual explanations are somewhat the opposite of Anchors. To revisit the same example: the person asking for a loan wants to know why he will not get one. A counterfactual explanation could then be: if your monthly income rises to $5000, you will get a loan.

Because of their many benefits and varying quality measures to optimize for, a *sprawl* of different counterfactual methods came into existence. This *possibly* leads to different respective explanations, which will be investigated in this work.

Guidotti (2022) and Verma et al. (2020) give an overview of counterfactual explanation techniques, however, to date, the state-of-the-art further unfolded with e.g., the introduction of NICE, a counterfactual generation algorithm which simultaneously achieves 100% coverage, model-agnosticism and fast counterfactual generation for different types of classification models.

2.2 Ambiguity due to a lack of consistent evaluation metrics for post-hoc explanations

As referred to in Sects. 1 and 2.1, evaluating XAI methods is a research field in its infancy today, even though a strong need for evaluation methods is identified by multiple authors such as Rosenfeld (2021). One reason for the limited amount of research done in this field can be the simple fact that evaluating XAI methods is *difficult*, especially for post-hoc explanation methods. Because of they explain black-box models, by definition, we don't know the logic involved in a decision made by such models.

Vilone and Longo (2021) divide XAI evaluation techniques into two groups: those that involve human-centered evaluations and those that evaluate with objective metrics. The first requires human participants to give qualitative or quantitative

feedback to XAI explanations, typically through surveys. For the second, to this day, more than 35 metrics have been proposed in the literature to evaluate XAI explanations. Examples of these metrics are, among others, actionability (knowledge is useful to the end-user), efficiency (computational speed of the algorithm), simplification (minimal features), stability (similar instances should provide similar explanations), etc. The authors conclude that the boom in the number of evaluation metrics calls for a general consensus among researchers on how an explanation should be evaluated.

Note that these objective metrics are sometimes hard to quantify. Qualitative quality properties are therefore often quantified in numbers. For counterfactual explanations, popular properties are proximity, sparsity, and plausibility (Verma et al. 2020). Proximity is a property that is somehow used in every counterfactual algorithm. It tries to measure the total change that is suggested by the counterfactual explanations with a distance metric (typically L1 or L2 distance) (Van Looveren and Klaise 2021; Mothilal et al. 2020; Wexler et al. 2019). It is intuitive that less change is better than more change in most situations. Sparsity is a special case of proximity. It refers to the number of features in the explanation (L0 distance) (Karimi et al. 2020; Dandl et al. 2020; Laugel et al. 2018). The argument is that shorter explanations are more comprehensible for humans than longer ones (Miller 1956). Finally, plausibility is a more conceptual property that refers to the closeness to the data manifold (Pawelczyk et al. 2020). For example, in a credit scoring context, advising someone to wait 200 years to get a loan, is not plausible.

Counterfactual explanations have an additional advantage in comparison to feature importance methods. The latter estimate the influence of each feature on the predicted score. These estimates potentially suffer from bias and features which have almost no influence on the model's decision might be labeled important (Fernández-Loría et al. 2020). Counterfactual explanations don't suffer from this bias, applying the suggested changes of a counterfactual explanation will always lead to a change in prediction. Consequently, a counterfactual explanation is always 'correct' (in the sense that it leads to a class change). However, counterfactual explanations are a simplification of all the information involved in the decision-making. Therefore, different explanations contain different bits of information. And while every counterfactual explanation is 'correct', it is not guaranteed to be useful.

The ambiguity of measuring the quality of counterfactual explanations has led to the development of many counterfactual algorithms and possibly as many different explanations (Verma et al. 2020; Guidotti 2022). As a result, when a stakeholder wants to use counterfactual explanations, he is presented with many options. This might be an advantage or can lead to the disagreement problem.

2.3 The disagreement "problem"

The disagreement problem in XAI arises when different interpretability methods, used to explain a given AI model, produce conflicting or contradictory explanations. Because of a lack of broadly used evaluation methods, this is often the case, resulting in explanations that are generally non-consistent and thus ambiguous. Neely

et al. (2021) raise the question of whether agreement as an evaluation method for XAI methods is suitable. When assuming agreement as an evaluation method, low agreement would mean only a few of the XAI methods are right, while the others are far from ideal. However, low agreement is not necessarily a bad thing.

Ambiguity can actually be valuable or result in possible ethical consequences (Martens 2022). It all depends on the context in which XAI methods are used (Bordt et al. 2022). Mothilal et al. (2020) argues that diversity among counterfactual explanations is beneficial. This for one increases the chance of generating usable explanations. For example, when someone is not allowed to get a loan according to a prediction model, and the only counterfactual explanation is to change their sex or lower their level of education, this explanation is argued not to be useful. Some people prefer to get an actionable explanation, such as 'increase your income with $X'. This actionability is not uniform over all decision subjects. Therefore, providing multiple explanations increases the chance that one explanation is useful for this specific user.

Bordt et al. (2022) examine when ambiguity in explanations is problematic. They differentiate between a cooperative and adversarial context. In a cooperative context, all stakeholders have the same interests. For example, in most medical applications of AI, both doctors and patients have the same goal: to improve or manage the patient's health. In adversarial contexts, this is not the case. Here, different parties have opposite interests. For example, when a student is denied admission to a prestigious university, the student is interested in challenging this decision. Another example is an autonomous car crashing into a wall to avoid a pedestrian. Insurance companies have other interests than the owner of the car or the developers of the software that steers the car's driving decisions. A final example is a denied bank loan: the bank and the client have different interests. In these cases, it might not be in the model user's best interest to look for the most correct or elaborate explanation of a decision that is made. The model user will most likely choose the explanation that fits their best interest, *if* diverse explanations are available. An adversarial context can lead to all kinds of ethical issues (Martens 2022). Aïvodji et al. (2019) examine the use of post-hoc explanations to fairwash or rationalize decisions made by an unfair prediction model, while Slack et al. (2020) and Lakkaraju and Bastani (2020) investigate the discriminatory characteristics of explanations. Imagine a model using a prohibited feature such as e.g., gender or race, or a feature that is linked to one of these e.g., zip code, when other more neat explanations are available. The model user could choose to ignore the discriminatory explanations and use another one instead. When considering the ethical consequences of disagreement, consensus amongst explanations might be desired. Therefore, consensus between explanations could be seen as a training objective to increase user trust (Schwarzschild et al. 2023; Hinns et al. 2021). Namely, if two explanations are consensual, the ethical consequences of choosing one XAI method over another one are less severe.

In fact, the scope of how explanation providers manipulate explanations extends beyond the selection of explanation algorithms. Goethals et al. (2023) identify a total of 6 stages in which explanations can be influenced. Besides algorithmic selection, users can also change the parameters of XAI algorithms to influence the explanations. Furthermore, some algorithms are non-deterministic and each run can result in a distinct explanation, which can also be exploited. Less obvious might be that

manipulation can happen in earlier stages as even changing the training data, predictive model or test data can also lead to different explanations. Our quantitative assessment is only limited to the algorithmic decision stage. However, our recommendations on how to move forward with disagreement in Sect. 4 and especially the call for transparency applies to all stages in the framework of Goethals et al. (2023).

2.3.1 Related work

The ethical issues related to the selection of model explanations can only arise if there actually is ambiguity. Neely et al. (2021) were the first to measure the disagreement problem in XAI. They compare LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, Deep-SHAP, and attention-based explanations with a rank correlation (Kendall's τ) metric. They conclude there is only low agreement in the explanations of these methods, between 0.19 and 0.27 depending on the data set used. Krishna et al. (2022) expand the previous study by comparing LIME, KernelSHAP, Vanilla Gradient, Gradient Input, Integrated Gradients, and SmoothGrad, once again finding disagreement amongst explanation of different methods, especially when the model complexity increases. Instead of only using a rank correlation metric, they use a feature agreement, (signed) rank agreement, sign agreement, and rank correlation. Depending on the type of data (tabular, text or image data), they use different of the above-mentioned evaluation metrics. For tabular data, which are used in this work, they found the rank and signed rank agreement to be significantly lower, compared to the feature agreement. They find the feature agreement to be between 79.1% and 100% agreement when looking at the top 5 features and 100% when looking at the top 7 features. Next to a quantitative comparison, the authors also perform a qualitative study on how practitioners handle the disagreement problem. 84% of practitioners interviewed by Krishna et al. (2022) mentioned encountering the disagreement problem on a day-to-day basis. They report there is no principle evaluation method to decide on which explanations to use, therefore, they simply choose to generate explanations with the XAI method they are most familiar with. Han et al. (2022) extend the study of Krishna et al. (2022) to investigate why the disagreement problem exists for these methods. They conclude that different XAI methods approximate a black-box model over different neighborhoods by applying other loss functions. If two explanations are trained to predict different sets of perturbations, then the explanations are each accurate in their own domain and may disagree. A more focused disagreement problem study can be found in Roy et al. (2022) where the explanations of LIME and SHAP are investigated for one single defect prediction model. They calculate the feature, rank, and sign agreement also proposed by Krishna et al. (2022). They conclude that LIME and SHAP disagree more on the ranking of important features compared to the feature agreement or the sign agreement of the features.

Table 1 gives an overview of the scarce literature on the quantitative evaluation of disagreement between XAI methods relative to our work. To the best of our knowledge, the disagreement problem has not yet been quantified for counterfactual XAI methods. This is remarkable because recently there has been a boom in the number of such algorithms, increasing the malicious user's potential for

Table 1 Literature overview of the quantitative evaluation of the disagreement between post-hoc XAI methods

Authors	Nb. of datasets	Type of datasets	Nb. of models	Nb. of XAI methods	XAI
Neely et al. (2021)	5	Text	2	6	LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, Deep-SHAP, and attention-based explanations
Krishna et al. (2022)	4	Tabular, Text and Image	8	6	LIME, Kernel SHAP and Integrated Gradients
Roy et al. (2022)	4	Tabular	1	2	LIME, SHAP
Brughmans et al. (2023)	40	Tabular	2	12	SHAP, Counterfactuals and Anchor

exploiting the disagreement problem in this category of XAI algorithms. Furthermore, as mentioned in Sect. 1, within counterfactual explanations, malicious users have a greater capacity to, for instance, circumvent specific biased attributes in the explanation, such as gender, when compared to feature importance algorithms. In the latter approach, every feature is considered, thus facilitating the detection of any inclusion of sensitive attributes in the model. A malicious user can deliberately select an explanation that omits gender, even if gender played a role in the model's decision. Consequently, the affected individual may never realize that their gender contributed to the decision. Because counterfactual explanations consistently yield 'a rule'-outcome, in the sense that following them alters the prediction, the individual affected by the decision is likely to readily accept the provided explanation as *the* reason behind the initial prediction. We therefore, argue it is important to quantify the disagreement amongst counterfactual explanation algorithms, which is elaborated upon in Sect. 3 with an example.

It should be noted that morally the disagreement problem for counterfactual explanations is similar to the Rashomon effect, introduced by Hasan and Talbert (2022). This effect concerns the diversity in multiple counterfactual explanations generated by *the same* counterfactual algorithm for the same instance and classifier. One explanation might say to change feature A (e.g., wait 5 years to get a loan), while another might say to change feature B while not adapting A (e.g., make sure your income increases with $500 to get a loan immediately). This initially seems like a contradiction as well. For example, the DiCE algorithm is focused on generating multiple explanations (Mothilal et al. 2020). In contrast to the Rashomon effect, the disagreement problem investigates diversity amongst *different* counterfactual explanation algorithms for the same instance and classifier. However, at the heart of the matter, the Rashomon effect and the disagreement problem face the same ethical issues and moral hazards: who chooses which explanation will be used?

3 The quantified disagreement amongst counterfactual explanation methods

In this section, we aim to quantify the disagreement between counterfactual explanation methods. We first illustrate the problem and research questions with an example in Sect. 3.1. Section 3.2 clarifies the large-scale experimental setup. Section 3.3 answers the research questions by providing metrics to quantify the disagreement amongst counterfactual algorithms and using these metrics in our large-scale experimental setup. Note that measuring the disagreement between counterfactual explanations comes with some new challenges. First of all, counterfactual explanations provide a set of features without ranking them. This makes measures such as (signed) rank agreement useless. Second, counterfactual explanations consist of variable sizes. Some algorithms might suggest six feature changes while others might only suggest two.

3.1 Example

Table 2 illustrates the disagreement problem for an example instance retrieved from the Adult data set (Dua and Graff 2017). This data set can be used to predict if a person would have an annual income higher or lower than $50,000. The person depicted in the instances is predicted to have an income lower than $50,000. Consequently, ten different counterfactual explanation algorithms generated counterfactual instances to tell which features should change in the original instance in order to change the prediction.

Firstly, it should be noted that one of the counterfactual explanation methods, CBR, is not able to find a counterfactual instance for the given original instance, while the others do find one. When having the need to explain a certain instance, it is consequently useful that other explanation methods are able to find explanations and thus, that some disagreement amongst methods is existing. However, in an adversarial context, a malicious counterfactual generating user, wishing to avoid a certain feature e.g., sex or race, is able to do so by simply selecting a counterfactual method that does not include these features, i.e., wants to change these features with respect to the original instance, such as DiCE, NICE (*plaus*) or NICE (*spars*). Imagine we are predicting whether or not this person would be qualified to get a loan from a bank. A prediction model that uses features like sex or race would then be unethical and discriminatory. The decision maker would be able to *hide* this fact by secretly choosing a counterfactual method that does not include sex or race in the explanations. This way, the unfair prediction model can still be "rationally" explained. Vice versa, if the malicious user explicitly wants to *include* a certain feature in an explanation, e.g., hours per week, they can do so with CFproto, NICE (*none*) or NICE(*plaus*). And this once again by simply choosing among the diverse explanations, without any need to impose constraints on the counterfactual generating search. This shows the arbitrariness/disagreement of the methods and the power that it brings to the user of counterfactual generating methods. A user can include, as well as avoid, almost any desired feature in the given explanation.

This example clearly illustrates the possible existence of the disagreement problem and the ethical consequences resulting from this existence. In the following sections, we examine how easy it is to abuse the disagreement problem by malicious agents and the driving factors that cause the disagreement problem.

3.2 Experimental setup

Table 3 gives an overview of the 40 tabular data sets we use for our study. Note that this number is significantly higher compared to the 4 to 5 data sets that are used in previous studies of the disagreement problem (see Table 1). This allows us to confidently make more general conclusions.

A test set is created for each data set by comprising 20% of the data with a minimum of 200 instances. This means that e.g., for the threeOf9 data set, we do not use 102 instances in the test set, but we use 200. The remaining data is used as the

Table 2 Example instance (Adult data set)

	Workclass	Marital status	Relationship	Race	Sex	Age	Fnlwgt	Edu	Edu num	Occupation	Capital gain	Capital loss	Hours \week	Native country
Instance CBR	Local gov	Wid	Other relative	Native A/A	F	26	195,693	1st-4th	9	Craft-repair	0	0	40	France
CFproto	**Self empl.**	Wid	**Own-child**	Native A/A	**M**	**44**	**90,688**	**Doctorate**	**15**	**Other service**	0	0	**51**	**Portugal**
WIT	Local gov	**Never marr.**	**Unmarried**	**Black**	F	**29**	**197,932**	1st-4th	9	**Handlers-cleaners**	0	0	40	**Ecuador**
GeCo	Local gov	**Marr.**	Other relative	**Black**	F	**17**	**195,695**	1st-4th	9	Craft-repair	0	0	40	**Hong K.**
NICE(*none*)	Local gov	Wid	Other relative	**Black**	F	**43**	**112,763**	1st-4th	9	Craft-repair	**8614**	0	**43**	**Hong K.**
NICE(*plaus*)	Local gov	Wid	Other relative	**Black**	F	**43**	**112,763**	1st-4th	9	Craft-repair	**8614**	0	**43**	**Hong K.**
DiCE	Local gov	Wid	Other relative	Native A/A	F	**50**	195,693	1st-4th	9	Craft-repair	**28,533**	0	40	France
NICE(*prox*)	Local gov	Wid	Other relative	Native A/A	F	26	195,693	1st-4th	9	Craft-repair	**8614**	0	40	France
NICE(*spars*)	Local gov	Wid	Other relative	Native A/A	F	26	195,693	1st-4th	9	Craft-repair	**8614**	0	40	France
SEDC	Local gov	**Never marr.**	**Wife**	Native A/A	**M**	**39**	195,693	**Masters**	**10**	Craft-repair	0	0	40	France

(Edu. = Education, Local gov. = Local government, Wid. = Widowed, Native A/A = Native Alaskan/American, Self empl. = Self employed, Never marr. = Never Married, Marr. = Married, Hong K. = Hong Kong)

Table 3 Descriptive statistics and performance metrics of all 40 binary data sets

Name	#Inst.	#Feat.	#Cat. feat.	#Num. feat.	Class imbalance	AUC (ANN)	AUC (RF)
adult	48,842	14	5	9	0.761	0.903	0.913
agaricus_lepiota	8154	22	21	1	0.481	1.000	1.000
australian	690	14	7	7	0.445	0.905	0.940
breast_w	699	9	8	1	0.345	0.991	0.997
buggyCrx	690	15	8	7	0.555	0.921	0.949
chess	3196	36	36	0	0.522	1.000	0.999
churn	5000	20	4	16	0.142	0.872	0.919
clean2	6598	168	0	168	0.154	1.000	1.000
coil2000	9822	85	84	1	0.060	0.691	0.745
credit_a	690	15	8	7	0.555	0.902	0.910
credit_g	1000	20	17	3	0.700	0.663	0.731
crx	690	15	8	7	0.445	0.859	0.941
diabetes	768	8	0	8	0.349	0.823	0.851
dis	3772	29	23	6	0.985	0.895	0.989
GAMETES1	1600	20	20	0	0.500	0.636	0.648
GAMETES2	1600	20	20	0	0.500	0.746	0.780
GAMETES3	1600	20	20	0	0.500	0.664	0.722
GAMETES4	1600	20	20	0	0.500	0.690	0.705
german	1000	20	17	3	0.700	0.718	0.758
Hill_Valley	1212	100	0	100	0.505	0.993	0.557
hypothyroid	3163	25	18	7	0.952	0.975	0.988
kr_vs_kp	3196	36	36	0	0.522	1.000	0.999
magic	19,020	10	0	10	0.352	0.922	0.937
mofn_3_7_10	1324	10	10	0	0.779	1.000	1.000
monk1	556	6	6	0	0.500	1.000	1.000
monk2	601	6	6	0	0.342	1.000	0.896
monk3	554	6	6	0	0.520	0.992	0.986
mushroom	8124	22	21	1	0.482	1.000	1.000
parity5+5	1124	10	10	0	0.504	1.000	0.674
phoneme	5404	5	0	5	0.294	0.906	0.970
pima	768	8	0	8	0.349	0.867	0.819
profb	672	9	3	6	0.333	0.633	0.676
ring	7400	20	0	20	0.505	0.990	0.992
spambase	4601	57	0	57	0.394	0.974	0.988
threeOf9	512	9	9	0	0.465	0.972	0.999
tic_tac_toe	958	9	9	0	0.653	0.997	1.000
tokyo1	959	44	2	42	0.639	0.962	0.983
twonorm	7400	20	0	20	0.500	0.996	0.997
wdbc	569	30	0	30	0.371	0.973	0.981
xd6	973	9	9	0	0.331	1.000	1.000

Table 4 Overview of the counterfactual algorithms used for comparison

Name	Author	Spars	Prox	Plaus	Type
CBR	Keane and Smyth (2020)	x		x	NB
CFproto	Van Looveren and Klaise (2021)	x	x	x	GD
WIT	Wexler et al. (2019)		x	x	NB
GeCo	(Schleich et al. 2021)	x	x	x	GA
NICE (*none*)	Brughmans et al. (2023)		x	x	NB
NICE (*plaus*)	Brughmans et al. (2023)	x		x	NB
DiCE	Mothilal et al. (2020)		x		GD
NICE (*prox*)	Brughmans et al. (2023)		x		NB
NICE (*spars*)	Brughmans et al. (2023)	x			NB
SEDC	Martens and Provost (2014)	x			BF

GD = Gradient Descent, NB = Neighbor-Based, GA = Genetic Algorithm, BF = Best-first

training set for training a Random Forest classifier (RF) and an Artificial Neural Network (ANN). The final two columns of Table 3 display the AUC values obtained for both classifiers. The hyper-parameters of both models are trained using a five-fold cross-validation approach. Subsequently, we generate counterfactual explanations using all algorithms for a random sample of 200 instances from the test set. In total, we generate 200 counterfactual explanations for 10 counterfactual algorithms, Anchors, and SHAP for 2 classifiers on 40 data sets, resulting in a sample size of 192,000 explanations.

We selected a total of 12 post-hoc explanation methods to study the disagreement problem. Our focus lies on ten counterfactual algorithms that are suited for tabular data, are model-agnostic, and have their code publicly available. They are depicted in Table 4 and we refer to Appendix 1, Table 13, for the parameter values used in every algorithm. The final column of Table 4 indicates the type of heuristic used for the respective counterfactual algorithms. The final selection includes the following counterfactual algorithms: DiCe (Mothilal et al. 2020), CFproto (Van Looveren and Klaise 2021), WIT (Wexler et al. 2019), CBR (Keane and Smyth 2020), SEDC (Fernández-Loría et al. 2020), GeCo (Schleich et al. 2021) and four types of the NICE algorithm (Brughmans et al. 2023) to investigate the uniformity of their explanations. We refer to their respective manuscripts for detailed descriptions of the different counterfactual algorithms. Moreover, we also look at their disagreement with both SHAP and Anchors. As we mentioned in Sect. 2.2, there is no consensus on what defines the quality of a counterfactual explanation, which resulted in many algorithms optimizing explanations for different evaluation metrics. When we compare algorithms optimized or evaluated for other metrics, some form of disagreement is expected. However, when comparing explanations from algorithms that optimize for the same metric, one might expect less disagreement.

We notice two distinct groups in Table 4 (divided by a horizontal line). The first six algorithms optimize for plausibility and the last four do not. Hence we call the first group `Plaus`, and the second group `Prox`. We make this distinction because every algorithm that optimizes for plausibility, actually also optimizes for

proximity (or sparsity) in some way: CBR starts from the closest case, CFproto has proximity in its loss function, WIT selects the closest counterfactual instance from the training set, GeCo selects the fittest candidate in each iteration based on proximity, NICE (*none*) selects the closest counterfactual instance from the training set and lastly, NICE (*plaus*) has sparsity in its reward function.

The second group (`Prox`) is *only* interested in providing counterfactual instances that are close to the original instance. We take algorithms that optimize for sparsity and proximity together because sparsity is a special case of proximity, as is explained in Sect. 2.2. Moreover, even though DiCE and NICE (*prox*) mainly optimize for proximity, they both have an indirect optimization for sparsity. DiCE includes a sparsity enhancing step by adapting as many features as possible to their original value as long as the predicted class does not change. NICE (*prox*) has a sparsity loss function embedded in its proximity function. We therefore define the `Prox`-group as algorithms that optimize for any distance metric (e.g., L0, L1 or L2 distance). We start our counterfactual disagreement analysis in Sect. 3.3 by looking at counterfactual algorithms and the two groups globally, after which we also include pairwise comparisons between the algorithms individually.

For the counterfactual explanations, we consider a feature to be present in the explanation if the counterfactual instance indicates to change the feature compared to the original instance. For Anchors we consider a feature present simply when it is mentioned in the Anchor explanation. For SHAP we take into account the seven most important features as features present in an explanation (based on Miller (1956) and Krishna et al. (2022)).

3.3 Results

3.3.1 To what extent can counterfactual disagreement be abused by malicious agents?

The main issue with disagreement amongst counterfactual explanations is that malicious users can select certain explanations to rationalize decisions made by unfair or discriminating models. This can be done by either avoiding certain features to convince stakeholders that they are irrelevant or the other way around, by including certain features to insinuate that they are the main driver of the decision-making process.

We first check, how easy it is to *exclude* a certain feature from a counterfactual explanation. This can be done by looking at the percentage of features that are not present in at least one explanation. Equation (1) formalizes this metric which we call relative feature exclusion. In this metric, the numerator counts the unique features that are *not* present in the explanations of certain methods a to n. This number is divided by the total number of features F_D in a data set D. Tables 5 and 6 show the average relative feature exclusions for different data sets, XAI-methods and classifiers:

Table 5 Relative feature exclusion for the RF classifier

Data set	Prox	Plaus	All CF	All
adult	97.8	99.5	100	100
agaricus_lepiota	98.1	100	100	100
australian	99.8	98.9	100	100
breast_w	97.6	97.7	99.6	99.6
buggyCrx	99.2	99.6	100	100
chess	99.7	100	100	100
churn	98	99.1	99.9	100
clean2	99.6	99.3	99.7	99.7
coil2000	99.9	99.9	100	100
credit_a	99.6	97.5	99.9	100
credit_g	99.5	99.8	100	100
crx	99.4	99.4	100	100
diabetes	94.4	92.8	98	98.4
dis	97.6	99.6	99.9	100
GAMETES_1	99.7	100	100	100
GAMETES_2	99.3	99.9	100	100
GAMETES_3	99.7	100	100	100
GAMETES_4	99.6	100	100	100
german	98.8	99.9	100	100
Hill_Valley_without_noise	99.6	99.7	100	100
hypothyroid	98.6	98.9	99.9	100
kr_vs_kp	99.6	100	100	100
magic	93.6	94.6	98.6	99.9
mofn_3_7_10	97.2	99.9	100	100
monk1	97.5	98.3	99.6	99.8
monk2	97.2	98.2	99.8	99.9
monk3	95.1	93.5	96.6	99.1
mushroom	97.8	100	100	100
parity5+5	99.5	99.9	100	100
phoneme	86.8	89.8	97.5	98.9
pima	96.8	90.7	98.3	98.3
profb	98.6	95.5	99.7	100
ring	98.2	99.6	99.9	100
spambase	97.5	99.7	100	100
threeOf9	98.6	98.9	99.8	99.9
tic_tac_toe	99.9	99.2	100	100
tokyo1	99.4	99.5	100	100
twonorm	86.6	99	99.7	99.9
wdbc	98.6	99.3	99.8	100
xd6	98.4	97	99.3	100
Average	97	98.6	99.6	99.9
Standard Deviation	3.4	1.9	0.9	0.2

Table 6 Relative feature exclusion for the ANN classifier

Data set	Prox	Plaus	All CF	All
adult	99.4	99.4	100	100
agaricus_lepiota	97.7	97.3	99.8	100
australian	98.4	99.9	100	100
breast_w	99.5	99.6	99.9	100
buggyCrx	98.2	99.9	100	100
chess	99.5	100	100	100
churn	99.4	98.9	99.9	100
clean2	99.9	99.8	100	100
coil2000	99.6	99.7	100	100
credit_a	98.6	99.3	100	100
credit_g	99.2	99.9	100	100
crx	98.2	100	100	100
diabetes	89.8	95.9	97.4	99.9
dis	99.8	100	100	100
GAMETES_1	99.3	100	100	100
GAMETES_2	99.1	99.8	99.9	100
GAMETES_3	99.7	99.9	100	100
GAMETES_4	99.4	100	100	100
german	99	100	100	100
Hill_Valley_without_noise	99.1	100	100	100
hypothyroid	96	99.4	99.8	100
kr_vs_kp	99.4	100	100	100
magic	89.8	96	98.5	100
mofn_3_7_10	98.3	99.9	100	100
monk1	95	98.4	99.4	99.9
monk2	97.4	98.2	99.8	100
monk3	93.3	93.8	95.6	98.8
mushroom	97.4	96.4	99.6	100
parity5+5	99.7	100	100	100
phoneme	92.5	92.9	98.5	99.3
pima	93	98.6	99.4	99.5
profb	95.6	96.4	98.8	100
ring	98.3	98.5	99.9	99.9
spambase	99	100	100	100
threeOf9	98.3	96.7	99.5	99.8
tic_tac_toe	99.1	96.6	99.7	99.9
tokyo1	89.9	100	100	100
twonorm	90.3	99.6	100	100
wdbc	88.2	99.9	100	100
xd6	97.1	94.2	98.1	100
Average	97.8	98.4	99.6	99.8
Standard Deviation	3	2.6	0.8	0.4

$$\text{Relative feature exclusion}_{[a,n]} = \frac{|(F_D \setminus E_a) \cup (F_D \setminus E_b) ... \cup (F_D \setminus E_n)|}{|F_D|} \quad (1)$$

Next, we investigate the possibility to *include* a random feature into a counterfactual explanation. For this, we introduce a metric called relative feature span, see Eq. (2). It measures the percentage of all features that is present in at least one explanation. The numerator equals the absolute feature span and measures the size of the union of all explanations of all explanation methods a to n in the comparison. The absolute feature span divided by F_D is the relative feature span. A higher feature span most likely results from a higher disagreement amongst methods. Consequently, the user will be able to choose many features as part of the explanation. The maximum relative feature span of 1 is achieved when every single feature is used in at least one explanation. These relative feature spans are shown in Tables 7 and 8:

$$\text{Relative feature span}_{[a,n]} = \frac{|E_a \cup E_b ... \cup E_n|}{|F_D|} \quad (2)$$

If we revisit the example of Sect. 3.1 and assume that the user only has the first two counterfactual explanation algorithms, CFproto and WIT, available. The relative feature exclusion between these two methods amounts $\frac{|(F_D \setminus E_{CFproto}) \cup (F_D \setminus E_{WIT})|}{|F_{Adult}|}$ or 57.1%, meaning that 57.1% of the features can be avoided by the user when using only these two methods. The relative feature span of both methods amounts $\frac{|E_{CFproto} \cup E_{WIT}|}{|E_{Adult}|}$ or 85.7%. This means that 85.7% of the features are present in the explanations of CFproto and WIT, and can consequently be chosen by the user. If all counterfactual explanation methods of Table 2 are available to the user the overall relative feature exclusion equals 100%. This means that any of the features can be chosen to be left out of the explanation if all methods are available to the user. The overall relative feature span equals 92.9%. Only the feature 'capital loss' is never used in the explanations.

Our results in Tables 5 and 6 show that excluding certain features is particularly easy when multiple explanations are available. To obtain the relative feature exclusions for every data set, we first calculate the relative feature exclusion for each of the 200 counterfactual explanations individually. Then, we average these numbers. The average relative feature exclusion is over 99.6% for both classifiers over all counterfactual methods, and 99.8% if we include Anchors and SHAP. For many data sets, the average relative feature exclusion is even 100.0%. Meaning that for every instance that has to be explained, every feature of choice can be excluded from the explanation. These results show that it is fairly easy to avoid sensitive features in order to falsely justify model decisions.

Selecting random features that are desired to be in an explanation seems slightly more difficult. The average relative feature span for all counterfactual methods (depicted in Table 7 and 8) is 62.1% (63.3%) for an RF (ANN) classifier, and 73.9% (72.8%) if we include Anchors and SHAP. However, there are still data sets that have a relative features span of 100.0%. Similarly to the relative feature exclusion calculation, to obtain the relative feature span for every data set, we first calculate

Table 7 Relative feature span for the RF classifier

Data set	Prox	Plaus	All CF	All
adult	40.5	48.4	59.7	74.1
agaricus_lepiota	32.7	53	62.4	65.4
australian	45.3	57.7	63.8	74.1
breast_w	56.8	85.3	87.8	88.9
buggyCrx	29.3	60.8	64.4	75
chess	7.3	17.6	20.3	25.8
churn	47.2	87.9	88.9	93.2
clean2	14.7	92.4	92.8	100
coil2000	17.3	31.3	39.8	41.8
credit_a	41.1	57.6	62.7	75.1
credit_g	27.7	55.3	59.4	65.6
crx	27.1	59.4	64.5	74
diabetes	57.2	89.1	91.8	93.2
dis	17.1	24.4	24.8	49.6
GAMETES_1	17	45	49.8	53
GAMETES_2	15.8	39.4	44.2	67.8
GAMETES_3	16.8	39	43.6	62.3
GAMETES_4	17.4	45.5	50.6	70.6
german	24.4	57.2	60.2	65.7
Hill_Valley_without_noise	17.4	100	100	100
hypothyroid	24.5	27.9	31.6	38.2
kr_vs_kp	7.9	17.6	20.5	26.5
magic	99.9	100	100	100
mofn_3_7_10	31.9	39.9	45.8	57.4
monk1	36.5	42.5	49.5	63.8
monk2	41.1	49.4	59.8	85.8
monk3	24.8	36.9	41.2	60.3
mushroom	30.6	51.8	60.6	64.7
parity5+5	23.8	26.6	38	87.7
phoneme	98.4	100	100	100
pima	64.8	88.2	90.9	99.9
profb	25.2	72.1	73.6	90.1
ring	27.9	100	100	100
spambase	30.3	29.9	35.7	94
threeOf9	25.7	25.4	34.4	56.7
tic_tac_toe	26.2	45.1	52.1	73.5
tokyo1	82.4	81.2	87.9	88
twonorm	100	100	100	100
wdbc	98.4	99.1	100	100
xd6	26.1	21.6	31.1	53.6
Average	37.4	57.5	62.1	73.9
Standard Deviation	25.8	27.1	25.2	21

Table 8 Relative feature span for the ANN classifier

Data set	Prox	Plaus	All CF	All
adult	28.2	55.3	61.4	68
agaricus_lepiota	23.6	52.1	57.6	60
australian	36.2	63.8	67	76.5
breast_w	41.4	82.8	87.8	91.2
buggyCrx	24.3	61.6	64.2	65.7
chess	8.6	17.6	21.2	27.8
churn	54.8	88.1	90.9	93.8
clean2	8.4	92.6	93.1	93.1
coil2000	17.6	30	38.3	40.2
credit_a	25	58.5	61.8	73.1
credit_g	20.6	56.9	58.9	65.5
crx	31.3	57.3	63.4	72.1
diabetes	54.8	89.6	92	92.4
dis	17.5	24.8	27.6	29.3
GAMETES_1	17.5	43.8	48.9	71.3
GAMETES_2	15.7	40.2	44.7	66.4
GAMETES_3	16.4	41.6	45.8	65.4
GAMETES_4	16.8	46.1	51	69.5
german	21.6	57.8	61.2	67.4
Hill_Valley_without_noise	21.3	100	100	100
hypothyroid	19.8	28.7	30.9	37.9
kr_vs_kp	9.1	17.5	21.4	28
magic	99.7	100	100	100
mofn_3_7_10	34.2	40.4	46.9	57
monk1	38.7	42.4	52.1	87.5
monk2	44.9	48.8	60	89.3
monk3	32.1	38.5	46.2	61.3
mushroom	26.2	50.3	56.8	62
parity5+5	25.1	33.2	42.4	88.9
phoneme	97.7	99.8	99.9	100
pima	49	90.1	91.4	91.8
profb	35.2	77.1	79.2	92.4
ring	32.7	100	100	100
spambase	29.4	34.9	38.4	40.1
threeOf9	26.1	33.9	42.4	61.9
tic_tac_toe	24.6	58.2	64.8	77.5
tokyo1	73.6	85	85.5	88.4
twonorm	99.9	100	100	100
wdbc	97.8	100	100	100
xd6	26.8	30.2	38.8	58.5
Average	35.6	59.2	63.3	72.8
Standard Deviation	24.8	26	24	21.4

Table 9 Correlation between data set characteristics and relative feature span or relative feature exclusion

	Relative feature span				Relative feature exclusion			
	ANN		RF		ANN		RF	
	AUC	# Feat	AUC	# Feat	AUC	# Feat	AUC	# Feat
prox	−0.05	0.19	−0.26	0.28	0.22	−0.19	0.23	−0.23
plaus	0.23	0.05	0.14	−0.03	0.23	0	0.21	−0.25
All CF	0.23	0.07	0.12	−0.04	0.15	−0.09	0.17	−0.12
All	0.21	0.04	0.01	−0.09	0.11	−0.01	0.10	−0.14

the relative feature span for each of the 200 counterfactual explanations individually, after which we average these results. Relative feature spans seem to vary tremendously over different data sets, groups of XAI methods, and classifiers. We refer to Sect. 3.3.2 for a more detailed examination of the drivers of this counterfactual disagreement.

To conclude, a malicious agent can easily both exclude and include desired features when multiple counterfactual algorithms are available. Especially excluding certain features to hide their influence in the prediction model, while still using them to generate predictions, can easily be done by leveraging the disagreement problem.

3.3.2 What are the drivers of counterfactual disagreement?

The disagreement problem makes it easy for malicious agents to exclude or include certain features. While the variation in feature exclusion is minimal, Tables 7 and 8 show that feature span does have some variation. In this section, we investigate what are the drivers of this variation. We investigate whether, the data set, the counterfactual algorithms, or the classifier cause the variation in counterfactual disagreement. This should help to identify when the possibility of feature disagreement is high.

Data set

Tables 7 and 8 show that there is a lot of variance over the different data sets. Over all counterfactual explanations, the relative features span varies from around 20% to 100% for both classifiers. However, this variance seems to be random. There is no observed relationship between the characteristics of the data set and the disagreement metrics. Table 9 shows that the number of features in the data set or the AUC of the trained models only have a very weak correlation with these disagreement metrics.

Counterfactual algorithms

As shown in Table 4, the counterfactual algorithms can be divided into two groups `plaus` and `prox`. Those that optimize for plausibility, and those that only optimize for proximity. It might be that the existing disagreement only originates from

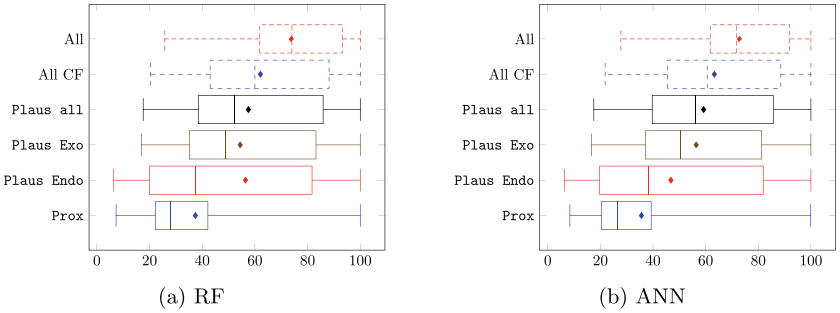


Fig. 2 Box-plots relative feature span for different algorithm groups for the RF and ANN classifiers

the disagreement between these groups and not from the disagreement within these groups. To verify this, we also calculated the relative feature span within these groups in Table 7 and 8 and a noticeable difference can be seen. The span within `plaus` is 20.1% to 23.6% higher compared to the span within the `prox` group.

Figure 2 stresses the difference between the two groups visually by the use of box plots. The center of gravity for the `prox` group is not only lower but also less broad, compared to the `plaus` group, meaning that even though the variance stretches over the entire x-axis, the gross of the relative feature spans for this group lies between 22% and 42% (20% and 39%) for the RF (ANN) classifier. In contrast, the relative feature span for the `plaus` group lies mainly between 39% and 86% (40% and 86%) for the RF (ANN) classifier.

To add more detail, we split up the `plaus` group into exogenous and endogenous counterfactual algorithms. Endogenous counterfactual algorithms provide explanations based on observed instances, while exogenous counterfactual algorithms give explanations based on unobserved instances (Crupi et al. 2022). In our experiments, NICE(*none*), NICE(*plaus*), CBR, WIT and Geco are endogenous, CFProto is exogenous. On average, the exogenous algorithm results in explanations with a higher feature span, compared to the endogenous ones. However, the `Plaus Endo` group has a wider spread, when looking at the gross of the relative feature spans.

This difference can simply be explained by the difference in sparsity between both groups as seen in Table 11. Sparsity refers to the number of features in a counterfactual explanation, normalised for the number of features present in the explanation. Optimizing for proximity (or sparsity directly) has a direct effect on this number of features in the explanations. Therefore, the average sparsity of the `prox` group is much lower compared to that of the `plaus` group. Consequently, having fewer features on average in each explanation also results in a lower relative feature span for this group. In fact, if we would look at only one counterfactual algorithm, the relative feature span would equal the normalized sparsity and the relative feature exclusion would equal 100%—the normalized sparsity.

Furthermore, the `plaus` group seems to account for most of the feature span of all counterfactual explanations. The difference between the `plaus` group and the group of all counterfactual explanations is less than 5% for both classifiers.

To first have a grasp of how similar the explanations are between different counterfactual algorithms, we examine the pairwise scaled L0 distances between the counterfactual instances of counterfactual methods. This metric counts the number of features that two explanations have in common and divide this number by the total number of features in the data set.

$$\text{L0 distance}_{ab} = \frac{|E_a \cap E_b|}{F_D} \tag{3}$$

To quantify the *pairwise* disagreement amongst counterfactual explanations, Anchors, and SHAP, we first introduce a new measure called feature disagreement in Eq. (4). This measure is similar to the feature agreement metric introduced by Krishna et al. (2022) but adapted to the variable explanation sizes of which counterfactuals consist.

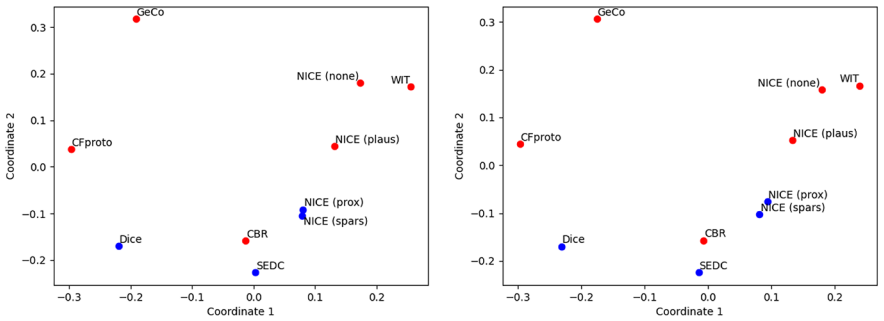
$$\text{Feature disagreement}_{ab} = \frac{|E_a \setminus E_b|}{|E_a|} \tag{4}$$

When comparing the explanations E_a and E_b of two methods A and B, the feature disagreement of A with B is equal to the size of the relative complement of E_a in E_b divided by the size of E_a. It measures the relative number of features that are in E_a but not in E_b. When the feature disagreement equals 1, none of the features of E_a are also in E_b.

It could be argued that the Jaccard similarity can be used to quantify the pairwise disagreement amongst counterfactual methods, this being an existing metric. However, feature disagreement has two advantages compared to Jaccard similarity. First, it contains information about the direction of disagreement. The feature disagreement of counterfactual A with counterfactual B is not equal to the feature disagreement of counterfactual A with counterfactual B, which is the case for Jaccard similarity. Second, the feature disagreement of counterfactual A with counterfactual B actually tells us the percentage of features in counterfactual A that contribute to a higher feature span on top of counterfactual B. For these reasons, we continue with this metric and refer to Appendix 1 for the Jaccard similarity analysis.

When revisiting the example in Sect. 3.1, and once again assuming the user only has the first two counterfactual explanation algorithms, CFproto and WIT, available. The pairwise scaled L0 distance between CFproto and WIT equals $\frac{E_{CFproto} \cap E_{WIT}}{F_{Adult}}$ or 35.7%. The feature disagreement between CFproto and WIT equals $\frac{|E_{CFproto} \setminus E_{WIT}|}{|E_{CFproto}|}$ or 50%. CFproto has 50% unique features with respect to WIT. Vice versa, the feature disagreement between WIT and CFproto equals $\frac{|E_{WIT} \setminus E_{CFproto}|}{|E_{WIT}|}$ or 28.6%. WIT has 28.6% unique features with respect to CFproto.

The pairwise scaled L0 distances are shown in Table 14 in Appendix B. We visualized these distances in a 2D-plot by using multidimensional scaling (MDS) in Fig. 3b and Fig. 3a. Note that the numbers on the x- and y-axis of these figures have no translatable meaning, only the relative Euclidean distances



(a) ANN (b) RF

Fig. 3 Multidimensional scaling (MDS) for the L0 distance between the counterfactual explanation methods (Blue: `Prox`, Red: `Plaus`)

Table 10 Relative feature disagreement for the RF and ANN classifier

RF	CBR	CFproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC	Anchors	SHAP	Average
CBR	0	37.0	14.3	28.9	15.1	20.3	35.4	25.2	25.0	39.1	12.1	43.3	24.6
CFproto	72.0	0	20.7	30.3	20.8	33.6	44.5	45.1	47.8	61.5	34.7	57.4	39.0
WIT	81.6	56.0	0	40.3	4.6	25.8	64.1	48.8	53.5	78.4	57.6	79.7	49.2
GeCo	87.3	63.4	35.7	0	36.7	51.0	70.1	69.0	72.7	82.0	63.9	77.9	59.1
NICE(none)	81.9	55.7	4.0	40.6	0	22.6	64.0	47.2	52.0	78.7	58.0	80.2	48.7
NICE(plaus)	78.0	54.4	3.5	41.1	0	0	60.2	37.4	40.2	70.4	50.3	80.2	43.0
DiCE	80.3	52.0	32.5	43.5	32.4	43.4	0	55.3	59.4	71.4	48.3	69.7	49.0
NICE(prox)	75.2	51.3	3.0	42.1	0	14.3	54.2	0	17.2	66.7	41.5	81.9	37.3
NICE(spars)	72.1	50.9	3.3	43.3	0	11.8	53.5	7.7	0	63.0	37.2	80.9	35.3
SEDC	63.7	45.2	20.2	35.1	21.3	28.1	40.4	33.2	34.3	0	21.2	58.2	33.4
Anchors	80.4	58.0	33.9	50.1	35.7	45.3	55.4	53.4	55.9	67.7	0	79.0	51.2
SHAP	89.7	72.0	57.7	59.6	59.9	67.7	73.3	78.4	79.1	82.3	72.1	0	66.0

ANN	CBR	CFproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC	Anchors	SHAP	Average
CBR	0	39.7	17.1	30.9	17.9	25.7	43	32.7	32.7	47.5	21.9	59.2	30.7
CFproto	74.3	0	24.8	36.9	24.7	39.6	51.6	54.6	56.7	66.6	42.8	75.3	45.7
WIT	80.8	53.5	0	38.9	4.3	29.9	67.3	55.7	57.4	80.2	59.8	87.3	51.3
GeCo	85.7	56.4	33.9	0	34.6	53.1	65.8	73.7	75.4	81.7	66.4	84.6	59.3
NICE(none)	80.8	53	3.3	38.8	0	26.9	67	54.3	56.1	80.2	59.9	87.8	50.7
NICE(plaus)	75.9	52.2	3.1	39.3	0	0	65.3	40.1	40	70.9	50.9	86.8	43.7
DiCE	82	51	36.8	44.2	36.5	51.8	0	65.7	67.5	75.7	54.8	82.7	54.1
NICE(prox)	70.2	46.7	3.1	41.1	0	12.8	58.6	0	10.7	63.9	38.8	87	36.1
NICE(spars)	68.4	48	3.5	41.8	0	10.6	59	7	0	62.2	37.1	86.3	35.3
SEDC	66.1	44.8	24	36.4	24.5	32.9	50.3	40.5	41.1	0	29.4	71	38.4
Anchors	77.8	54.5	34.2	46.2	35.3	48.3	60.7	57.2	58.8	69.7	0	86.9	52.5
SHAP	76	62.2	52.4	50.2	54.8	62	61.6	68.8	69.1	70.6	58.6	0	57.2

between two points are meaningful. The closer two points lie together, the more similar the resulting counterfactual instances of each method are. NICE(*prox*) and NICE(*spars*) optimize for very similar metrics with the same optimization method and therefore result in very similar counterfactual instances. The same can be said for NICE(*none*) and WIT. Both these methods use real instances from the training set in their explanations, and those instances seem to be quite close to each other. Surprisingly, SEDC and CBR are similar as well. GeCo provides counterfactual instances that are the farthest away from all other methods. This might be because GeCo's explanations contain many features in general. Table 11 shows, that GeCo on average uses around 74% of all features in a single counterfactual explanation.

Table 11 The average L0-distance (normalised sparsity) between the instance to explain and the counterfactual instance

Classifier	CBR	CFproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC
ANN	44.1	50.2	45.4	73.6	44.7	30.1	32.4	14.9	13.9	31.1
RF	55.1	57.9	44.8	74.0	44.4	32.6	40.4	18.7	15.7	38.3

Table 12 Intra and inter group relative feature disagreement

	RF		ANN	
	Intra group	Inter group	Intra group	Inter group
CBR	23.1	31.2	26.2	39
CFproto	35.5	49.7	40.1	57.4
WIT	41.7	61.2	41.5	65.2
GeCo	54.8	73.5	52.7	74.1
NICE(*none*)	41.0	60.5	40.5	64.4
NICE(*plaus*)	35.4	52.1	34.1	54.1
DiCE	62.0	47.4	69.6	50.4
NICE(*prox*)	46.0	31.0	44.4	29.0
NICE(*spars*)	41.4	30.2	42.7	28.7
SEDC	36.0	35.6	44.0	38.1

Which algorithms disagree the most with others, while taking into account the number of features present in their explanations, can be identified by looking at the pairwise feature disagreement. Table 10 shows that most post-hoc explanation methods have a high number of features that are not present in CBR, SEDC or SHAP (the darkest columns in Table 10) even though the sparsity of these methods is not necessarily low (see Table 11). Overall GeCo is the counterfactual algorithm that generates the most features that are not available in the explanations of other algorithms. Once again, this can be largely attributed to the fact that GeCo has the worst sparsity, meaning that it has the most features in its generated explanations.

The high feature disagreement of SHAP and Anchors with counterfactual explanations confirms that the disagreement between different post-hoc explanation methods is larger than the disagreement within counterfactual explanations.

Lastly Table 12, presents the average pairwise disagreement between a counterfactual method presented in the first column and the other counterfactual algorithms within the same (intra) and between (inter) group(s). For example, the intra group average for CBR (a member of the `plaus` group) equals the average of the relative feature disagreements between CBR and the other members of the `plaus` group: CFproto, WIT, GeCo, NICE(*none*) and NICE(*plaus*). On the other hand, the inter group average, amounts the average of the relative feature disagreements between CBR and the members of the `prox` group: DiCe, NICE(*prox*), NICE(*spars*), and SEDC. It is clear that for the `plaus` group, the intra group averages are significantly lower compared to the inter group averages. In contrast, for the `prox` group the inter group averages are lower. Since counterfactual algorithms of the `prox` group generate explanations with less features compared to the other group, chances

of disagreement are higher. Vice versa, it is easier to find agreement when a lot of features are present in the explanantions, which is the case in the `plaus` group.

Classifier

Surprisingly, the used classifier does not have a critical influence on the size of the disagreement problem. For each data set the difference between the average relative feature span between both classifiers is minimal. Moreover, we calculated the correlation between both classifiers, which is more than 99%. The same conclusion can be drawn from the L0 distances in Fig. 3 between the counterfactual instances or the relative feature disagreements in Table 10. Both metrics show little variation between the RF and ANN.

In conclusion, both the data set and the group of counterfactual algorithms determine the variation in disagreement metrics. Perhaps more surprisingly, we find that the classifier has a small to no influence on the results and variation obtained.

4 Discussion

Our experiments reveal a severe disagreement amongst various counterfactual explanation methods, suggesting that this disagreement may surpass the feature disagreement observed in feature-importance techniques by previous research Krishna et al. (2022); Roy et al. (2022); Neely et al. (2021). Note that direct comparisons with these prior studies are challenging due to their focus on rank and sign agreement of the top k features, metrics not directly applicable to counterfactual explanations. Our analysis centers on a variant of the feature agreement metric, used by Krishna et al. (2022), but adapted to the variable explanation sizes of counterfactual methods. Krishna et al. (2022) reported one hundred percent feature agreement among the tested feature-importance techniques for $k = 7$. This finding starkly contrasts with the significantly higher feature disagreement rates identified in our study, highlighting the unique challenges and potentially greater biases within counterfactual explanation methods.

This discrepancy is not entirely surprising, given the intrinsic challenges associated with their evaluation (cf. Section 2.2). The diversity of desired properties these methods aim to optimize contributes to their diversity in outcomes. One resolution might be to identify an optimal counterfactual property, providing a unified optimization objective for all counterfactual algorithms, thereby minimizing the disagreement. However, currently, the field is far from reaching a consensus on what this property should be. Perhaps we should ask ourselves if the field of XAI will ever be able to fully address this problem, or if it is simply a consequence of the field's core objective: XAI seeks to explain decisions made by complex predictive algorithms in terms understandable to humans. The gap between the complexity of these predictive algorithms and human comprehension abilities is vast. Therefore, the resulting explanations are often simplifications, leading to an inevitable loss of information. Most counterfactual explanations capture only a portion of this information, representing a unique point of view on the involved prediction logic. It is these varied

points of view that lead to the disagreement problem. Different stakeholders inevitably have preferences for distinctive points of view, amplifying this challenge. This is compounded by the existence of information asymmetry, where one stakeholder (the decision-maker) possesses more information and can determine what is shared with others (the decision subject). Such a scenario potentially opens the door to manipulative practices. Consequently, there is a pressing need for awareness and the implementation of mechanisms that strive to circumvent the disagreement problem as much as feasible.

Currently, we identify three potential solutions, though none emerge as flawless. The first proposed solution is to present all counterfactual explanations to the decision subject. This approach empowers decision subjects with the autonomy to select the explanation that best suits them, effectively limiting the opportunity for decision-maker manipulation. However, this presents challenges. The information might be too extensive for the decision subject, thus undermining the primary objective of XAI, which is to make algorithmic decision-making comprehensible. Moreover, there is a risk of providing more information than necessary. Companies invest heavily in predictive modeling, and the confidentiality of these models often preserves their competitive edge. Sharing too much information could inadvertently reveal insights to competitors. The ideal scenario would strike a balance, sharing sufficient information to validate the decision while preventing excessive conflict of interest about the model. Yet, there is no guarantee that such a balance can be attained.

A second potential solution is to summarize multiple counterfactual explanations into a combination, as suggested by Fernández et al. (2022) and Carrizosa et al. (2024). However, these approaches introduce an additional step in the framework introduced by Goethals et al. (2023). Further research would be needed to determine whether these combinatorial algorithms (dis)agree in their explanations. Also the usefulness and properties of these combinations should be further investigated.

We conclude that the current state of research has yet to sufficiently address the disagreement problem from a technical point of view. And currently, the most viable solution might be to give the decision-subject insight in the process of coming to explanations. Providing transparency should ensure that decision-makers can be held accountable for their actions. Ideally, such a principle would be enforced by regulation. This is particularly crucial in high-risk decision-making scenarios where individuals bear the consequences of algorithmic processes. One potential method to enforce this transparency is by introducing an audit mechanism when deploying these XAI algorithms. Decision-makers could only provide one or a limited number of explanations to the decision subject in this case. However, they would need to document and submit the reasoning, that led to the deployment of these XAI algorithms, for review to a third party. This entity would then evaluate whether the interests of the decision subject have been appropriately considered. However, this oversight should not be confined to the selection of the XAI algorithm alone. As mentioned in Sect. 2.3, Goethals et al. (2023) identify several more avenues for manipulation in the decision-making process. To close these vulnerabilities and maximize accountability, the entire decision-making process (in aspects of data selection, modeling, parameter choice, algorithm selection, etc.) should be thoroughly justified and presented to a third party for evaluation.

5 Conclusion and future research

In our large-scale empirical analyses, on 40 data sets, yielding over 192,000 explanations generated, we provide evidence of the existence of the disagreement problem amongst different counterfactual explanation algorithms. If a malicious agent has the option to choose between the 10 counterfactual algorithms examined in our experiments, it will be very easy to exclude features of their choice in an explanation. Including a feature of choice, is slightly more difficult, but still in many cases the relative feature span is 100%, giving the decision maker the full choice to include a certain feature.

Moreover, we conclude that the size of the disagreement problem is highly dependent on the data set and counterfactual methods used and not so much on the classifier used. However, we want to stress again that in contrast to other post-hoc explanation methods, disagreement between counterfactual explanations does not mean any explanation is wrong. On one hand, a counterfactual explanation cannot be wrong, as the suggested feature changes will by definition lead to a class change. It can, however, be that these suggested changes are not useful to certain stakeholders. Therefore, situations with high disagreement between counterfactual explanations signal instead that one single explanation fails to capture the full complexity of a decision made by a prediction model.

By proving the existence of the disagreement problem amongst counterfactual explanation methods, we demonstrate the potential rise of ethical issues. Especially in an adversarial context, where the goals of the stakeholders are not aligned, these ethical issues occur when users are able to choose which explanations are used, giving them a lot of power. To avoid the occurrence of moral issues, ideally, this power should be in the hands of the decision subject, as they carry the, possibly life-changing, consequences of the decision. We discussed how giving the explanatory decision power to the decision subject, can in turn create new issues. Some of these issues could potentially be solved by developing new algorithms that can combine many counterfactual explanations into one. However, this may cause the problem to simply move to the level of algorithms that combine these counterfactual explanations. Therefore, we argue that currently, the most viable solution might be to provide transparency. Decision makers should document all decisions that led to the explanations. This should allow a third party to determine whether the interest of the decision-subject are taken into account.

Appendix A: Parameter values

See Table 13.

Table 13 Parameters values used in XAI algorithms

Algorithm	Code	Parameter values
CBR	https://github.com/ADMAntwerp/NICE_experiments	distance_metric= 'HEOM' explanation_length = 2 tolerance = 0.2
CFproto	https://github.com/SeldonIO/alibi/	beta = 0.01 c_init = 1 c_steps = 5 max_iterations = 500 theta = 10 use_kdtree = True
WIT	https://github.com/ADMAntwerp/NICE	optimization = 'none' distance_metric= 'HEOM' num_normalization = 'std' justified_cf= False
GeCo	https://github.com/interpretml/DiCE	backend = 'sklearn' method = 'genethic'
NICE(none)	https://github.com/ADMAntwerp/NICE	optimization='none' distance_metric= 'HEOM' num_normalization= 'minmax' justified_cf = True
NICE(plaus)	https://github.com/ADMAntwerp/NICE	optimization='plausibility' distance_metric = 'HEOM' num_normalization = 'minmax' justified_cf = True
DiCe	https://github.com/interpretml/DiCE	backend = 'sklearn' method = 'random'
3NICE(prox)	https://github.com/ADMAntwerp/NICE	optimization ='proximity' distance_metric = 'HEOM' num_normalization = 'minmax' justified_cf = True
NICE(spars)	https://github.com/ADMAntwerp/NICE	optimization ='sparsity' distance_metric = 'HEOM' num_normalization = 'minmax' justified_cf = True
SEDC	https://github.com/ADMAntwerp/NICE_experiments	prune = True omit_default = True max_ite=20 stop_at_first = False cost_func = None
Anchor	https://github.com/marcotcr/anchor	threshold = 0.95
SHAP	https://github.com/shap/shap	k_obs = 50 n_samples = 50 l1_reg = False

Appendix B: Scaled L0 distance

See Table 14.

Table 14 Scaled L0 distance for the RF and ANN classifier

RF	CBR	Cfproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC
CBR	0	38.5	43.1	43.7	42.8	25.6	30.1	15.7	13.4	13.6
Cfproto	38.5	0	46.8	49.1	46.4	42.8	37.1	39.1	38.9	39.3
WIT	43.1	46.8	0	46.3	12.9	21.4	47.7	30.7	32.8	44.9
GeCo	43.7	49.1	46.3	0	46.7	45.3	45.7	44.6	44.3	43
NICE(none)	42.8	46.4	12.9	46.7	0	11.8	47.3	25.7	28.6	44.8
NICE(plaus)	25.6	42.8	21.4	45.3	11.8	0	41.7	18.1	19.3	30.5
DiCE	30.1	37.1	47.7	45.7	47.3	41.7	0	33.7	32.8	31.2
NICE(prox)	15.7	39.1	30.7	44.6	25.7	18.1	33.7	0	5.1	20.9
NICE(spars)	13.4	38.9	32.8	44.3	28.6	19.3	32.8	5.1	0	18.3
SEDC	13.6	39.3	44.9	43	44.8	30.5	31.2	20.9	18.3	0

ANN	CBR	Cfproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC
CBR	0	39	43.1	45.7	42.7	28.5	29.6	15.1	14	14.7
Cfproto	39	0	47.5	49.7	47.2	43.3	36.7	38.9	38.9	38.6
WIT	43.1	47.5	0	48	16.2	25.7	48.2	34.8	35.4	45.8
GeCo	45.7	49.7	48	0	48.2	47	46.8	46	46	46.4
NICE(none)	42.7	47.2	16.2	48.2	0	14.7	47.5	29.8	30.8	45.6
NICE(plaus)	28.5	43.3	25.7	47	14.7	0	40.5	18	18.1	29.1
DiCE	29.6	36.7	48.2	46.8	47.5	40.5	0	31.5	31.1	30.8
NICE(prox)	15.1	38.9	34.8	46	29.8	18	31.5	0	2.7	17.3
NICE(spars)	14	38.9	35.4	46	30.8	18.1	31.1	2.7	0	16.5
SEDC	14.7	38.6	45.8	46.4	45.6	29.1	30.8	17.3	16.5	0

Appendix C: Jaccard distance

Equation (5) counts the number of features two explanations have in common as the numerator and the union of both explanations in the denominator. The Jaccard distance always lies between 0 and 1. The complement, 1 minus the Jaccard distance, gives an indication of the dissimilarity of two explanations (Eq. 6).

$$\text{Jaccard distance or similarity}_{ab} = \frac{|E_a \cap E_b|}{|E_a \cup E_b|} \tag{5}$$

$$\text{Jaccard dissimilarity}_{ab} = 1 - \frac{|E_a \cap E_b|}{|E_a \cup E_b|} \tag{6}$$

Table 15 shows the pairwise inverse Jaccard distance or dissimilarity.

Table 15 Inverse Jaccard distance (dissimilarity) for the RF and ANN classifier

RF	CBR	CFproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC	Anchors	SHAP	Average
CBR	0	89.7	82.4	92.3	82.7	80.5	90.3	79.1	77.2	82.7	83.7	95.2	78
CFproto	89.7	0	58.7	72.5	58.7	66.6	74.5	74.2	76.7	84	80.9	90.6	68.9
WIT	82.4	58.7	0	50.6	6.7	27.1	68.8	49.2	53.8	79.7	74.6	85.4	53.1
GeCo	92.3	72.5	50.6	0	51.2	62.3	77.9	76.7	79.8	87.8	80	84.4	68
NICE(none)	82.7	58.7	6.7	51.2	0	22.6	68.6	47.2	52	80	75	85.9	52.6
NICE(plaus)	80.5	66.6	27.1	62.3	22.6	0	71.5	40.4	41.1	72.6	74.8	87.6	53.9
DiCE	90.3	74.5	68.8	77.9	68.6	71.5	0	74.7	76.9	83.2	79.4	90.9	71.4
NICE(prox)	79.1	74.2	49.2	76.7	47.2	40.4	74.7	0	18.8	70.1	71.5	91.5	57.8
NICE(spars)	77.2	76.7	53.8	79.8	52	41.1	76.9	18.8	0	67.4	69.8	91.2	58.7
SEDC	82.7	84	79.7	87.8	80	72.6	83.2	70.1	67.4	0	76.4	92.3	73
Anchors	83.7	80.9	74.6	80	75	74.8	79.4	71.5	69.8	76.4	0	88.9	71.3
SHAP	95.2	90.6	85.4	84.4	85.9	87.6	90.9	91.5	91.2	92.3	88.9	0	82

ANN	CBR	CFproto	WIT	GeCo	NICE(none)	NICE(plaus)	DiCE	NICE(prox)	NICE(spars)	SEDC	Anchors	SHAP	Average
CBR	0	88.4	81.6	90.8	81.6	78.5	90.2	75.2	74	83.1	82	94.5	76.7
CFproto	88.4	0	56.5	66.5	56.2	65.8	73.8	75	76.8	83.7	81.1	93.5	68.1
WIT	81.6	56.5	0	47.7	6.1	30.9	70.7	55.9	57.6	81.4	76.4	91.1	54.7
GeCo	90.8	66.5	47.7	0	48.1	63.4	72.9	80.1	81.6	87.3	81.8	90.4	67.6
NICE(none)	81.6	56.2	6.1	48.1	0	26.9	70.5	54.3	56.1	81.5	76.6	91.5	54.1
NICE(plaus)	78.5	65.8	30.9	63.4	26.9	0	76.2	42.2	40.9	73.4	75.9	91.5	55.5
DiCE	90.2	73.8	70.7	72.9	70.5	76.2	0	78.6	79.5	86.4	80.5	93	72.7
NICE(prox)	75.2	75	55.9	80.1	54.3	42.2	78.6	0	12.2	69.3	70.2	92	58.8
NICE(spars)	74	76.8	57.6	81.6	56.1	40.9	79.5	12.2	0	68.1	70.3	91.6	59.1
SEDC	83.1	83.7	81.4	87.3	81.5	73.4	86.4	69.3	68.1	0	78.4	92.1	73.7
Anchors	82	81.1	76.4	81.8	76.6	75.9	80.5	70.2	70.3	78.4	0	90.2	72
SHAP	94.5	93.5	91.1	90.4	91.5	91.5	93	92	91.6	92.1	90.2	0	84.3

Acknowledgements Dr. Lissa Melis was supported by a Fellowship of the Belgian American Educational Foundation (BAEF) and the President's Postdoctoral Fellowship Program (PPFP).

Data availability All 40 data sets used in this work are retrieved from the Penn Machine Learning Benchmark (PMLB). https://doi.org/10.1186/s13040-017-0154-4.

References

Arrieta AB, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion 58:82–115

Aïvodji U, Arai H, Fortineau O, et al (2019) Fairwashing: the risk of rationalization. International Conference on Machine Learning pp 161–170

Bordt S, Finck M, Raidl E, et al (2022) Post-hoc explanations fail to achieve their purpose in adversarial contexts. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp 891–905

Brughmans D, Leyman P, Martens D (2023) Nice: an algorithm for nearest instance counterfactual explanations. Data Mining and Knowledge Discovery pp 1–39

Carrizosa E, Ramírez-Ayerbe J, Morales DR (2024) Generating collective counterfactual explanations in score-based classification via mathematical optimization. Expert Syst Appl 238:121954

Crupi R, Castelnovo A, Regoli D, et al (2022) Counterfactual explanations as interventions in latent space. Data Mining and Knowledge Discovery pp 1–37

Dandl S, Molnar C, Binder M, et al (2020) Multi-objective counterfactual explanations. In: International Conference on Parallel Problem Solving from Nature, Springer, pp 448–469
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608
Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml
Dwivedi R, Dave D, Naik H et al (2023) Explainable ai (xai): Core ideas, techniques, and solutions. ACM Comput Surv 55(9):1–33
Fernández RR, de Diego IM, Moguerza JM et al (2022) Explanation sets: A general framework for machine learning explainability. Inf Sci 617:464–481
Fernández-Loría C, Provost F, Han X (2020) Explaining data-driven decisions made by ai systems: the counterfactual approach. arXiv preprint arXiv:2001.07417
Goethals S, Martens D, Evgeniou T (2023) Manipulation risks in explainable ai: The implications of the disagreement problem. arXiv preprint arXiv:2306.13885
Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation." AI magazine 38(3):50–57
Guidotti R (2022) Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery pp 1–55
Han T, Srinivas S, Lakkaraju H (2022) Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. arXiv preprint arXiv:2206.01254
Hasan MGMM, Talbert D (2022) Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach. In: The International FLAIRS Conference Proceedings
Hinns J, Fan X, Liu S, et al (2021) An initial study of machine learning underspecification using feature attribution explainable ai algorithms: A covid-19 virus transmission case study. In: PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part I 18, Springer, pp 323–335
Huysmans J, Baesens B, Vanthienen J (2006) Using rule extraction to improve the comprehensibility of predictive models
Karimi AH, Barthe G, Balle B, et al (2020) Model-agnostic counterfactual explanations for consequential decisions. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 895–905
Keane MT, Smyth B (2020) Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In: Case-Based Reasoning Research and Development: 28th International Conference, ICCBR 2020. Springer-Verlag, p 163-178
Krishna S, Han T, Gu A, et al (2022) The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602
Lakkaraju H, Bastani O (2020) " how do i fool you?" manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 79–85
Laugel T, Lesot MJ, Marsala C, et al (2018) Comparison-based inverse classification for interpretability in machine learning. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, pp 100–111
Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable ai: A review of machine learning interpretability methods. Entropy 23(1):18
Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in neural information processing systems 30
Martens D (2022) Data Science Ethics: Concepts, Techniques, and Cautionary Tales. Oxford University Press
Martens D, Provost F (2014) Explaining data-driven document classifications. MIS Quarterly 38(1):73–100. https://www.jstor.org/stable/26554869
Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol Rev 63(2):81
Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38
Molnar C (2018) A guide for making black box models explainable. URL: https://christophm github io/interpretable-ml-book
Mothilal RK, Sharma A, Tan C (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 607–617

Neely M, Schouten SF, Bleeker MJ, et al (2021) Order in the court: Explainable ai methods prone to disagreement. arXiv preprint arXiv:2105.03287
Păvăloaia VD, Necula SC (2023) Artificial intelligence as a disruptive technology-a systematic literature review. Electronics 12(5):1102
Pawelczyk M, Broelemann K, Kasneci G (2020) On counterfactual explanations under predictive multiplicity. In: Conference on Uncertainty in Artificial Intelligence, PMLR, pp 809–818
Ribeiro MT, Singh S, Guestrin C (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, pp 1135–1144
Ribeiro MT, Singh S, Guestrin C (2018) Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI conference on artificial intelligence
Rosenfeld A (2021) Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th international conference on autonomous agents and multiagent systems, pp 45–50
Roy S, Laberge G, Roy B, et al (2022) Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In: 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME), IEEE, pp 444–448
Schleich M, Geng Z, Zhang Y et al (2021) GeCo: Quality counterfactual explanations in real time. Proceedings of the VLDB Endowment 14(9):1681–1693
Schwarzschild A, Cembalest M, Rao K, et al (2023) Reckoning with the disagreement problem: Explanation consensus as a training objective. arXiv preprint arXiv:2303.13299
Slack D, Hilgard S, Jia E, et al (2020) Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 180–186
Van Looveren A, Klaise J (2021) Interpretable counterfactual explanations guided by prototypes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, pp 650–665
Verma S, Boonsanong V, Hoang M, et al (2020) Counterfactual explanations and algorithmic recourses for machine learning: A review. arXiv preprint arXiv:2010.10596
Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76:89–106
Wexler J, Pushkarna M, Bolukbasi T et al (2019) The what-if tool: Interactive probing of machine learning models. IEEE Trans Visual Comput Graphics 26(1):56–65

Authors and Affiliations

Dieter Brughmans[1] · Lissa Melis[2,3] · David Martens[1]

✉ Lissa Melis
lissa.melis@maastrichtuniversity.nl

Dieter Brughmans
dieter.brughmans@uantwerpen.be

David Martens
david.martens@uantwerpen.be

[1] Engineering Management Department, University of Antwerp, Prinsstraat 13, Antwerp 2000, Belgium

[2] Civil and Environmental Engineering Department, Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, USA

[3] School of Business and Economics, Maastricht University, Tongersestraat 53, Maastricht 6211 LM, The Netherlands