Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Survey Track

Robust Counterfactual Explanations in Machine Learning: A Survey

Junqi Jiang , **Francesco Leofante** , **Antonio Rago** and **Francesca Toni**
Department of Computing, Imperial College London, UK
{junqi.jiang, f.leofante, a.rago, f.toni}@imperial.ac.uk

Abstract

Counterfactual explanations (CEs) are advocated as being ideally suited to providing algorithmic recourse for subjects affected by the predictions of machine learning models. While CEs can be beneficial to affected individuals, recent work has exposed severe issues related to the *robustness* of state-of-the-art methods for obtaining CEs. Since a lack of robustness may compromise the validity of CEs, techniques to mitigate this risk are in order. In this survey, we review works in the rapidly growing area of *robust CEs* and perform an in-depth analysis of the forms of robustness they consider. We also discuss existing solutions and their limitations, providing a solid foundation for future developments.

1 Introduction

As the field of explainable AI (XAI) has matured, *counterfactual explanations* (CEs) have emerged as one of the dominant post-hoc methods for explaining AI models (see, e.g. [Karimi *et al.*, 2022] for an overview). CEs are often advocated as a means to provide *recourse* for individuals that have been impacted by the decisions of a machine learning model. In particular, given an input x to a model M, a CE essentially presents a user with a new, slightly modified input x', which shows how a different outcome could be achieved if the proposed changes were to be applied to x. For illustration, consider a fictional loan application with features *income* $\pounds 50K$, *loan term* 35 months and *loan amount* $\pounds 10K$ being rejected by a model. In this example, a CE could demonstrate that increasing the *income* to $\pounds 55K$ would result in the application being accepted.

Given the critical nature of many scenarios in which CEs are deployed, e.g. in financial or medical settings, it is of utmost importance that the recourse they provide is valid, i.e. it gives the intended change in outcome, and thus trustworthy. However, recent work has demonstrated that state-of-the-art methods for obtaining CEs host major drawbacks when it comes to the *robustness*, i.e. the validity under changing conditions, of the CEs they generate. In particular, [Pawelczyk *et al.*, 2022] showed that popular approaches for generating CEs may return explanations that are indistinguishable from adversarial examples. Broadly speaking, this means that CEs are extremely susceptible to small changes occurring in the setting they were generated for. To understand the implications of this, let us return to our loan example: if increasing income to $\pounds 55.1K$ (as opposed to the exact amount recommended) does not result in the application being accepted, the applicant may begin to question whether the CE was actually explaining the decision making of the AI model and was not just an artefact of the explainer instead. Worryingly, this is just one of many examples in which a lack of robustness in CEs may compromise their reliability.

In this survey, we conduct the first[1] comprehensive analysis of techniques developed to ensure that CEs are robust. After introducing the necessary background on CEs in Section 2, we detail the methodology for our systematic survey of the existing literature on Robust CEs in Section 3. We then classify approaches based on the form of robustness they consider and discuss each resulting category in Sections 4-7. In Section 8, we discuss key findings of the survey and look ahead to future prospects in this emerging research field.

2 Counterfactual Explanations

Assume a classification model $M : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is a discrete set of output labels[2]. We focus on models M obtained by machine learning. Given an input point $x \in \mathcal{X}$, most machine learning methods (e.g. logistic regression, neural networks, tree ensembles) first produce a *class score* s ranged in $[0, 1]$ for each class in $\mathcal{Y}$, and then use it to determine the model classification (e.g. by choosing the class with the highest score). Given an input $x \in \mathcal{X}$ and a model M, existing approaches compute a CE x' as follows:

$$\arg\min_{x' \in \mathcal{X}} cost(x, x') \text{ s.t. } M(x') \neq M(x) \tag{1}$$

where $cost : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a suitable metric in the input space (e.g. ℓ_1 or ℓ_∞ norm) and $M(x') \neq M(x)$ amounts to *validity* of the CE. In practice, solving the formulation exactly, e.g. through Mixed Integer Programming (MIP) as in [Mohammadi *et al.*, 2021], may be viable only for certain types of classifiers. For differentiable classifiers, a relaxed

[1] To the best of our knowledge, only [Mishra *et al.*, 2021] have surveyed robustness in XAI methods previously. However, their primary focus was on feature attribution explanations instead of CEs.

[2] Most approaches focus on binary classification ($\mathcal{Y} = \{0, 1\}$).

formulation is typically considered instead:

$$\arg\min_{x' \in \mathcal{X}} loss(M(x'), M(x)) + \lambda \cdot cost(x, x') \quad (2)$$

where $loss$ is a differentiable loss function that pushes the search towards a valid CE x', i.e. one for which $M(x') \neq M(x')$, and λ is a parameter dictating the trade-off term between validity and cost. Other properties have also been considered in the CE literature (see [Karimi *et al.*, 2022] for a recent survey of CEs and their properties). For instance, *actionability* requires the CEs to act only on mutable features in realistic directions, *causality* means that the changes in the input should conform with a full or a partial structural causal model, *diversity* advocates the generation of a *set* of different CEs for each input instead of a single CE, and *plausibility* means a CE should be close to the data manifold.

3 Systematic Survey

We conducted a systematic search of the existing literature on robustness of CEs. We first performed keyword searches on Google Scholar using the following patterns for exact matches: *robust counterfactual explanation*, *robustness of counterfactual explanation*, *counterfactual explanation robustness*. For better coverage in each pattern, the words *consistent* (*consistency*) and *stable* (*stability*) were used in addition to *robust* (*robustness*). The phrase *counterfactual explanation* was also interchangeable with *counterfactual explanations, counterfactuals, recourse, algorithmic recourse*. To narrow down the topic, we filtered for technical papers from and after year 2017, when two seminal works in the CE literature were published [Tolomei *et al.*, 2017; Wachter *et al.*, 2017]. For completeness, we also checked in Google Scholar for additional papers that cited the influential early works on robust CEs, namely [Pawelczyk *et al.*, 2020; Upadhyay *et al.*, 2021; Slack *et al.*, 2021].

Table 1 summarises the results of our survey. We found that the type of robustness falls within four separately studied categories, namely robustness against *Model Changes* (MC), *Model Multiplicity* (MM), *Noisy Execution* (NE), and *Input Changes* (IC). For the identified approaches to these problems, we detail the suitable types of models to be explained, the level of access to the model which is required, the computational method used, whether formal guarantees of robustness are given, and other considered properties.

In the next sections, we review the problem definitions, evaluation metrics for robustness, solutions and theoretical results for each category of robustness.

4 Robustness against Model Changes

As shown in Table 1, the majority of approaches on robustness of CEs in the literature focus on robustness against MC. Model changes are typically defined as modifications in the parameters of a machine learning model that do not alter its architecture. These changes are typically assumed to result from retraining on inputs drawn from a slightly shifted data distribution [Rawal *et al.*, 2020]. However, model changes resulting from data deletion queries have also been studied [Pawelczyk *et al.*, 2023b; Krishna *et al.*, 2023].

Different characterisations of the notion of change have been proposed. [Upadhyay *et al.*, 2021] first considered *plausible model changes*, which are defined as updates to model parameters whose magnitude is bounded by a small constant. The same notion is also considered in [Black *et al.*, 2022c; Jiang *et al.*, 2023a; Jiang *et al.*, 2023b]. The assumption on bounded updates is lifted in [Hamman *et al.*, 2023], where *naturally-occurring model changes* are studied in the context of neural networks. In a nutshell, this notion allows for unbounded changes as long as the updated model preserves a similar behaviour around the input being explained.

The robustness against MC can be summarised as follows.

Robustness against MC: Assume an input x and a model M. Let x' be a CE for x. Robustness against MC requires that whenever the model M **changes** to M', and this change is **sufficiently small**, then $M(x') = M'(x')$.

Despite different definitions of what constitutes a (small) change in this setting, e.g., some approaches require explicitly that $M(x) = M'(x)$), all existing works agree that a lack of such *robustness against MC* could be a problem for both the user and the explanation providers. Consider the loan example: after being rejected, the applicant would expect that when applying again, achieving a £55K annual salary would result in the application being successful, as captured by the CE. However, a lack of robustness to MC may result in this CE being invalidated by small updates to the machine learning model (e.g. by retraining with new data). Ultimately this may lead to the application being rejected despite the applicant having implemented the CE. On the contrary, a robust CE should preserve its validity in such a scenario.

4.1 Robustness Metrics

A commonly used metric to assess the robustness of CEs against MC is *Validity after Retraining* (VaR). Roughly speaking, VaR measures the percentage of CEs that remain valid after the model for which they were generated is updated. Common update strategies involve retraining from scratch using shifted datasets [Upadhyay *et al.*, 2021] or different portions of the original dataset [Hamman *et al.*, 2023], incremental retraining [Jiang *et al.*, 2023b], and retraining with varying initialisation conditions [Black *et al.*, 2022c]. This metric is also referred to as *recourse outcome instability* [Pawelczyk *et al.*, 2023b] and *recourse reliability* [Fonseca *et al.*, 2023], the latter being defined for CEs in a multi-agent multi-step setting. A probabilistic formulation of this property is also given in [Bui *et al.*, 2022], where the authors compute lower and upper bounds on the probability of a CE remaining valid after retraining of linear models.

To quantify robustness against plausible model changes, the Δ-*robustness* metric was proposed in [Jiang *et al.*, 2023b]. Intuitively, a CE is said to be Δ-robust if its validity is preserved across the whole family of models that can be obtained after applying bounded model changes. Given a CE x', this can be can be captured by requiring that $M(x') = M'(x')$ for all M' such that $\|Param(M') - Param(M)\|_p \leq \delta\}$ for fixed p, δ and $Param$ being the vectorisation of a model's parameters.

		Models	Access	Method	Guarantee	Properties
MC	[Mochaourab *et al.*, 2021]	SVM	Pred.	Search	-	-
	[Upadhyay *et al.*, 2021]	LM	Grad.	GD+RO	-	Ac
	[Black *et al.*, 2022c]	NN	Grad.	GD	-	-
	[Bui *et al.*, 2022]	LM	Grad.	GD	Prob.	Di
	[da Silva and Bertini, 2022]	$*$	Pred.	Search	-	Pl
	[Dutta *et al.*, 2022]	TM	Pred.	Search	Prob.	Pl
	[Ferrario and Loi, 2022]	$*$	White	DA	-	-
	[Forel *et al.*, 2022]	TM	White	MIP	Prob.	Ac/Pl
	[Nguyen *et al.*, 2022]	DM	Grad.	GD	-	Pl
	[Bui *et al.*, 2023]	$*$	Pred.	GD, MIP	-	Ac/Pl
	[Guo *et al.*, 2023]	DM	White	GD+RO	-	Ac
	[Hamman *et al.*, 2023]	NN	Pred.	Search	Prob.	Pl
	[Jiang *et al.*, 2023a]	NN	White	MIP+RO	Det.	Ac/Pl
	[Jiang *et al.*, 2023b]	NN	White	MIP	Det.	-
	[Krishna *et al.*, 2023]	LM	Grad.	GD	Det.	-
	[Nguyen *et al.*, 2023]	$*$	Pred.	MIP	Det. / LM	Ac
	[Pawelczyk *et al.*, 2023b]	DM	Grad.	GD	-	-
	[Wang *et al.*, 2023a]	$*$	Pred.	Search	-	Ac/Di/Pl
MM	[Pawelczyk *et al.*, 2020]	$*$	Pred.	Search	-	Pl
	[Leofante *et al.*, 2023]	NN	White	MIP	Det.	-
	[Jiang *et al.*, 2024]	$*$	Pred.	CA	Det.	-
NE	[Hada and Carreira-Perpiñán, 2021]	TM	White	MIP	-	-
	[Dominguez-Olmedo *et al.*, 2022]	DM	Grad.	GD+RO	LM	Ac/Ca
	[Sharma *et al.*, 2022]	$*$	Pred.	Search	-	Di/Pl
	[Guyomard *et al.*, 2023]	DM	Grad.	GD	Prob.	-
	[Leofante and Lomuscio, 2023]	NN	White	FV	Det.	Ac
	[Maragno *et al.*, 2023]	NN, TM	White	MIP+RO	Det.	Ac/Pl
	[Pawelczyk *et al.*, 2023a]	DM	Grad.	GD	Prob.	Ac
	[Raman *et al.*, 2023]	DM	Grad.	Sampling	Prob.	Ac/Ca/Di/Pl
	[Virgolin and Fracaros, 2023]	$*$	Pred.	GA	-	Pl
IC	[Artelt *et al.*, 2021]	$*$	White	MIP	Prob. / LM	Pl
	[Slack *et al.*, 2021]	DM	Grad.	GD	-	-
	[Wang *et al.*, 2023b]	$*$	Pred.	SAT	-	Di
	[Zhang *et al.*, 2023]	$*$	Pred.	Search	-	Pl
	[Leofante and Potyka, 2024]	$*$	Pred.	Search	Det.	Di

Table 1: Robust CE methods, partitioned based on the targeted robustness form. We include: *(i)* the types of models they target (support vector machines (SVM), linear models (LM), neural networks (NN), differentiable models (DM), tree-based models (TM) and model agnostic methods ($*$)); *(ii)* the model access required (white box, gradients and predictions); *(iii)* the method used (gradient descent (GD), robust optimisation (RO), mixed-integer programming (MIP), computational argumentation (CA), formal verification (FV), genetic algorithms (GA), satisfiability solving (SAT) and data augmentation (DA)); *(iv)* whether a robustness guarantee is provided (Deterministic, Probabilistic, for linear models only (LM) and none ($-$)); *(v)* other properties satisfied by the method, including Actionability, Causality, Diversity, and Plausibility. Note that all methods consider validity and proximity by default, thus they are omitted.

Counterfactual Stability (CS) is a robustness metric proposed to capture naturally-occurring model changes. A formulation of CS was introduced in [Dutta *et al.*, 2022] for tree ensemble models to capture the intuition that *i.* the class score for a robust CE should be high and that *ii.* the class score of the CE's neighbouring input should also be high with a small variance. In practice, for a given CE, the CS score evaluates the mean and standard deviation of the class scores of samples from a Gaussian distribution centred at the CE. The metric was later specialised in [Hamman *et al.*, 2023] to target naturally-occurring models changes in neural networks.

4.2 Algorithms

Robust optimisation approaches. Targeting plausible model changes, [Upadhyay *et al.*, 2021] solve a min-max problem that generates the best CE under the largest admissible model change. At each step, an inner maximisation routine finds the weight perturbation vector that increases the prediction loss to the greatest extent. Then, an outer minimisation loop updates the CE which optimises the overall loss (Eq (2) under the previously computed worst-case perturbation). Although this formulation natively supports linear models only, the authors show that it can also be applied to non-linear models by leveraging linear surrogate models obtained, e.g. using LIME [Ribeiro *et al.*, 2016][3].

Similarly to this approach, [Jiang *et al.*, 2023a] present a constrained optimisation formulation of the same problem in the context of feed-forward neural networks with piece-wise activations. Using a MIP encoding instead of gradient-based optimisation, the authors are able to provide stronger robust-

[3]Note that in practice, all CE methods applicable to linear models can be also used for non-linear models by approximating them with linear surrogate models, but any theoretical properties could be lost.

ness guarantees than [Upadhyay *et al.*, 2021]. However, this comes with higher computational costs.

Increasing class scores. Other approaches generate robust explanations by guiding the search towards CEs that result in high class scores. For example, [Krishna *et al.*, 2023] leverage leave-k-out analysis to approximate the solutions to the robust CEs problem by finding CEs with higher class scores on the original model. [Forel *et al.*, 2022] increase the required class scores in the CE search and give probabilistic guarantees for ensembles of convex base learners. Similarly, [da Silva and Bertini, 2022] target CEs with high class scores by exploring the input's neighbouring points using k-associated optimal graph analysis. Meanwhile, [Mochaourab *et al.*, 2021; Wang *et al.*, 2023a] construct application-dependent CE prototypes and select those with higher class scores as robust CEs.

Complementing these results, [Black *et al.*, 2022c] show that for complex non-linear models, higher class scores alone may not be sufficient to ensure robustness. Therefore, additional requirements need to be imposed on the search to promote robustness. For instance, [Black *et al.*, 2022c] require that the CEs be located in a region with a low Lipschitz constant and propose a method that leverages this result during CE search. Instead, [Dutta *et al.*, 2022; Hamman *et al.*, 2023] generate robust CEs by requiring high class scores not only for the CE, but also for its neighbouring points. A first search algorithm exploiting this idea is presented in [Dutta *et al.*, 2022] in the context of decision trees; a gradient-based version is later developed in [Hamman *et al.*, 2023]. Both methods provide theoretical probabilistic guarantees for their CEs' VaR Finally, [Jiang *et al.*, 2023b] propose to strengthen the class score requirement with a robustness test based on the notion of Δ-robustness. The authors propose an algorithm that iteratively generates CEs with higher class scores until the candidate solution is certified to be robust. For the certification step, a MIP encoding is used to explore the space of model changes exhaustively.

Probabilistic modelling. Model changes can also be approximated using probabilistic modelling techniques as done in [Nguyen *et al.*, 2022], where kernel density estimation techniques are used to model the data distributions, along with a Gaussian mixture ambiguity set to capture the model shifts. A min-max optimisation problem is then formulated to maximise the worst-case validity probability of CEs and solved using gradient-based optimisation. A similar approach is taken in [Nguyen *et al.*, 2023] where the space of possible model shifts is modelled using Gaussian mixtures over model parameters. Finally, [Bui *et al.*, 2022] obtain a lower bound on the probability of VaR by modelling the mean and covariance matrix of some nominal distribution assumed for the model parameters. Their lower bound can be relaxed to a differentiable form, and thus can be added as an additional term to the loss function of the gradient-based CE search.

(Re-)Training for robustness. Training methods have been proposed to robustify the CE generation process. Augmentation-based training is proposed in [Ferrario and Loi, 2022], where CEs are used alongside training instances to train neural network classifiers. The authors show empirical results pointing to the fact that augmentation increases the chances of CEs remaining valid under successive shifts in the data-generating distribution. However, this technique may cause imbalances in the data, since the CEs are not naturally-collected data points from the same distribution as other training data. [Guo *et al.*, 2023] formulate a tri-level optimisation problem which aims to simultaneously train an accurate neural network for prediction purposes, for which robust CEs can be generated by passing any input into another jointly trained neural network. Finally, [Bui *et al.*, 2023] sample data points (with their prediction results) centred at the original model's decision boundary near the input, and perturb the data covariance matrix of these samples. Then, linear surrogate models which are aware of the potential decision boundary shifts in the original model can be trained using the data shifts. They show that CEs generated on such linear surrogates using non-robust methods are robust.

5 Robustness against Model Multiplicity

MM (also called *predictive multiplicity*) refers to the phenomenon that for a single machine learning task, multiple near-optimal models can be trained with similar test accuracies [Breiman, 2001; Marx *et al.*, 2020; Black *et al.*, 2022b]. MM can be dealt with by aggregating the models towards a single *aggregated prediction* for each input [Black *et al.*, 2022a; Black *et al.*, 2022b]. However, recent results have shown that these models could have distinct behaviours for the same individuals in terms of predictions and their explanations [Rudin, 2019; Coston *et al.*, 2021]. This has important implications for CEs, with one being that a CE for an aggregated prediction may not even be valid for all of the aggregated models, which would likely confound users' expectations, leading to the following high-level definition:

Robustness against MM: Assume an input x and a set of models $\mathcal{M}$. Let $agg(\mathcal{M}, x)$ be an **aggregated prediction** for x, and x' be a CE for $agg(\mathcal{M}, x)$. Robustness against MM requires that x' is **valid across some subset** $\mathcal{M}' \subseteq \mathcal{M}$, i.e. for any $M_i \in \mathcal{M}'$, $M_i(x') \neq agg(\mathcal{M}, x)$.

In the CE literature, [Pawelczyk *et al.*, 2020] observe that traditional CEs generated on one original model have a high probability of being invalidated by other models resulting from MM. In this setting, it is assumed that *the prediction result is fixed to one produced by one of the original models*, and the robustness target is to guard the CE validity against potential future changes in the classifier: a robust CE should ideally remain valid under the alternative models. This form of robustness is very similar to robustness against MC. The differences are that the MM problem does not assume a data distribution shift causing the model shifts, and there are no assumptions on the model types and architectures for the original model on which the CEs are generated and for the alternative models.

Recent advances have also raised concerns about *when the prediction result for one input is pending* and to be decided by

all the models under MM, instead of assuming a result from one original model [Black *et al.*, 2022a]. [Jiang *et al.*, 2024] investigate the problem of deciding the prediction result together with the CEs, with the robustness of the latter driving the former. That is, the solution should be a subset of models under MM and their (valid) CEs, from which the prediction results are decided.

5.1 Robustness Metrics

The robustness evaluation for the fixed prediction scenario is very similar to the VaR for robustness against MC problem, using validity under alternative models. These models could be arbitrary models trained on (different portions of) the same dataset with high accuracy. Thus, the same validity metric gives an intuitive evaluation of how robust the CEs are.

For the pending prediction scenario, [Jiang *et al.*, 2024] propose desirable properties to characterise the optimal resulting set of models and CEs. *Non-emptiness* requires that the resulting set be not empty, and ideally it should contain more than one model and CE such that the prediction result is collectively determined by multiple models with their explanations (*non-triviality*). *Model agreement* states that the models included in the solution set should have the same prediction result for the input. If these results are the same as the prediction results by the majority of models in the MM model set, then the solution is said to satisfy *majority vote*. Meanwhile, *counterfactual validity* requires that every selected CE remains valid on every selected model. Finally, *counterfactual coherence* enforces that if a model is selected as part of the solution set, then its corresponding CE should also be included, and vice versa.

5.2 Algorithms

Fixed prediction scenario. [Pawelczyk *et al.*, 2020] provide important theoretical results demonstrating that the plausible CEs within the data manifold are more likely to stay valid under alternative models, compared with the closest CEs found by optimising Eqs (1) or (2), and that there is a plausibility-cost tradeoff, indicating that the more robust CEs will admit higher costs. CEs empirically show increased (but much lower than 100%) validity under alternative models for an existing plausible CE method compared with non-plausible methods, along with increased costs. However, the method is not deployed to produce robust CEs. [Leofante *et al.*, 2023] propose a solution to generating CEs that are guaranteed to be valid for a set of ReLU neural networks. Their method builds a product construction combining all specified models into a single neural network, on which CEs can be generated by MIP. They also prove that finding a CE that is valid across a set of piece-wise linear models is NP-complete.

Pending prediction scenario. Leveraging methods in computational argumentation, [Jiang *et al.*, 2024] present a novel argumentative ensembling method for finding a subset of the models and their CEs which satisfies the properties mentioned in Section 5.1 (with the exception of majority vote). In essence, this consists in grouping similarly-behaving models together to eliminate conflicts therebetween, then finding the maximal group of models and CEs as the final output. Alternatively, one could first decide the prediction result for the input using ensembling methods like majority voting or the method of [Black *et al.*, 2022a], then select the group of models and CEs with this prediction as the final solution set.

6 Robustness against Noisy Executions

Traditional methods compute CEs under the assumption that the user receiving them will follow them to the letter. In practice, achieving the prescribed feature values exactly might not always be feasible, especially for the continuous features. Continuing from the example given in Section 1, a CE may suggest a salary increase of £5342.04 but achieving this exact change could be beyond the loan applicant's control. In some other cases, even if the prescribed change is achievable, the user may inadvertently introduce some noise in the implementation of the recourse. Broadly speaking, robustness to NE requires that a CE be invariant to these small changes:

Robustness against NE: Assume an input x and a model M. Let x' be a CE for x. Robustness against NE requires that whenever a **small perturbation** σ is applied to x', validity is not affected, i.e. $M(x') = M(x' + \sigma)$.

Different characterisations of σ have been proposed in the literature. For instance, [Pawelczyk *et al.*, 2023a; Raman *et al.*, 2023] assume noise could be sampled from some probability distribution [Pawelczyk *et al.*, 2023a; Raman *et al.*, 2023]. A more conservative formulation is used in [Dominguez-Olmedo *et al.*, 2022; Leofante and Lomuscio, 2023], where the authors take an adversarial ML and/or verification angle, and define σ as a p-norm ball around the CE. A specialised version of the latter is given in [Virgolin and Fracaros, 2023], in which a set of feature-specific perturbation vectors is obtained from domain knowledge.

6.1 Robustness Metrics

[Pawelczyk *et al.*, 2023a] formalise the notion of *invalidation rate* (IR) of a CE as the expectation of the difference between the predicted label of the CE and those predicted for its noisy variants, i.e. $\mathbb{E}_\sigma[M(x') - M(x' + \sigma)]$. When used to empirically evaluate CEs, the expectation component can be approximated by sampling a certain number of noise vectors.

Alternatively, inspired by the adversarial robustness literature, [Dominguez-Olmedo *et al.*, 2022] evaluate robustness by adding perturbation vectors (found by adversarial attack methods) to CEs to invalidate them, and measuring the minimum magnitude of the successful attack. Then, the CEs are said to be robust to perturbation vectors of up to this magnitude. Similarly, [Leofante and Lomuscio, 2023] employ formal verification tools to check the *local robustness* of a CE. In particular, a verification query is formulated to check if a CE remains valid within an ∞-norm ball around it. These notions are more conservative than the IR and complement it by capturing worst-case scenarios.

6.2 Algorithms

Robust optimisation and class scores. [Dominguez-Olmedo *et al.*, 2022] propose different solutions for both

linear and non-linear models. For the linear case, the authors prove that increasing class scores is a sufficient condition for finding robust CEs. Similar remarks are also made in [Hada and Carreira-Perpiñán, 2021; Sharma *et al.*, 2022]. In the non-linear case, a robust optimisation approach is proposed. This is similar to [Upadhyay *et al.*, 2021] for the MC problem (Section 4.2), except that the worst-case perturbation now applies to the CE itself instead of the model parameters. [Maragno *et al.*, 2023] provide a similar formulation using a MIP encoding, finding CEs with robustness guarantees for both neural networks and tree ensemble models.

Verification-based approach. [Leofante and Lomuscio, 2023] show that local robustness queries as commonly phrased in adversarial machine learning can be used to evaluate a given CE's robustness. Then, the authors provide an iterative algorithm to quantify robustness by embedding a local robustness check into a binary search procedure.

Novel loss functions. [Pawelczyk *et al.*, 2023a] propose a method that allows the end user to specify the IR of their CEs. The intuition behind this is that robust CEs typically have higher cost, which the user may not be willing to sustain. Specifying a threshold on the IR allows the end user to control the level of risk they intend to take, thus implicitly reducing the cost from the most robust CEs. The authors give a differentiable first-order approximation of the IR for linear models and Gaussian noise applied to the CE. This term can then be added to the loss function of Eq (2) and optimised with gradient descent. Under specific assumptions on the choices of models and CE generation algorithms (e.g. logistic regression models and using the method of [Wachter *et al.*, 2017] as the base CE method), IR can be expressed analytically, and a probabilistic robustness guarantee in terms of IR can be obtained through setting the class score. To tackle these strong assumptions and their inherent limitations on practicality, [Guyomard *et al.*, 2023] relax the definition of IR to capture class scores with which they provide a tighter upper bound on the exact IR. Then, the upper bound approximation by Monte Carlo estimation is plugged into the loss function and CEs are optimised in the same manner as in [Pawelczyk *et al.*, 2023a]. In [Virgolin and Fracaros, 2023], a robustness term which is similar to IR is added to Eq (2), and then minimised via a genetic algorithm.

Probabilistic approaches. The methods of [Dominguez-Olmedo *et al.*, 2022; Maragno *et al.*, 2023; Leofante and Lomuscio, 2023] effectively obtain a region containing valid CE points. Such regions could also be characterised by probability distributions, from which CEs could be sampled. [Raman *et al.*, 2023] take a Bayesian hierarchical modelling approach where noise is explicitly modelled by random variables with Gaussian or scaled Dirichlet distributions, for continuous and categorical variables respectively. Their method outputs a posterior distribution which estimates the counterfactual distribution, and a Hamiltonian Monte Carlo sampling method is applied to generate diverse CEs.

7 Robustness against Input Changes

Unlike the previous notions of robustness which mostly focus on the validity of CEs under certain perturbations, robustness against IC requires that CEs generated for similar inputs are consistent. A first argument in favour of this property is given by [Hancox-Li, 2020], who advocates that explanations generated for similar inputs should not differ radically to improve the justifiability of explanations. However, [Slack *et al.*, 2021] found that traditional CE generation algorithms may fail to satisfy this property, raising fairness concerns. As an example, the paper shows that neural networks can be trained in such a way that individuals belonging to a non-protected group can always obtain a lower cost recourse when compared to protected group. Following this example, we can formulate robustness against IC as follows.

Robustness against IC: Assume two inputs x_1, x_2 and a model M such that $M(x_1) = M(x_2)$. Let x_1', x_2' be CEs for x_1, x_2, respectively. Robustness against IC requires that whenever x_1, x_2 are **similar**, then x_1', x_2' are also **similar**.

Similarity between inputs is typically defined in terms p-norm balls, e.g. given an input x_1, an input x_2 is similar to x_1 if $\|x_1 - x_2\|_p \leq \delta$. Then, one could use this formulation to derive a similarity notion for CEs, bounding the maximum distance therebetween. For example, [Leofante and Potyka, 2024] discuss the case where $\|x_1' - x_2'\|_p \leq k \, \|x_1 - x_2\|_p, k \in \mathbb{R}^+$. A special case of robustness against IC is encountered when dealing with inputs with missing feature values, where x_2 differs from x_1 only in the missing attributes. The returned CEs should ideally capture validity for any missing feature values, as discussed in [Kanamori *et al.*, 2023].

7.1 Robustness Metrics

Robustness measures are usually characterised by the expected distance between the CEs of similar inputs, $\mathbb{E}_{x_2 \sim S}[d(x_1', x_2')]$, where S denotes a set of inputs that are similar to x_1 and d is a distance metric defined over the input space. This quantity is identified as the *local instability* of CEs [Artelt *et al.*, 2021], and is targeted by most of the surveyed studies. If the method generates a diverse set of CEs for a single input, then local instability needs to be generalised to account for distances between two sets of points, as in [Wang *et al.*, 2023b; Leofante and Potyka, 2024].

7.2 Algorithms

Adversarial manipulations and defenses. [Slack *et al.*, 2021] propose an adversarial training framework whereby the cost of CE changes can be made to change drastically for across protected and unprotected subgroup of data points. They demonstrate that such instabilities against small perturbations in the input are evident for traditional gradient-based CE methods, and discuss possible mitigation strategies. These include randomising the CE search initialisation, reducing the number of features used for CE search, or reducing the model size to limit overfitting.

Plausibility and robustness. [Artelt *et al.*, 2021] give some theoretical results about the upper bound of the local instability of CEs for linear binary classifiers under Gaussian and Uniform noise in the input. They show that plausible CEs, i.e. CEs that lie within the data manifold, exhibit a higher

degree of invariance to input perturbations. This insight is also the intuition behind the method by [Zhang *et al.*, 2023] which aims to maximise robustness to input changes by finding more plausible CEs. Similarly, [Wang *et al.*, 2023b] proposes a boolean satisfiability approach to generate plausible CEs, which also demonstrate a lower local instability than other traditional CE methods.

Robustness via diversity. The above methods do not explicitly target the local instability of CEs. [Leofante and Potyka, 2024] show that satisfying robustness notions such as $\|x_1' - x_2'\|_p \leq k \|x_1 - x_2\|_p, k \in \mathbb{R}^+$ may be impossible for traditional CE methods in the general case. Therefore, the authors propose to move away from single-instance CEs and instead consider (diverse) sets of CEs. Under some assumptions, the authors prove that a relaxed form of robustness can be satisfied whereby sets of explanations generated for similar inputs are guaranteed to contain similar CEs.

8 Summary and Outlook

We conducted a comprehensive and fine-grained analysis of robust CE generation approaches and categorised them into the type of robustness they consider. In doing so, we identified some open research questions that are shared across the robustness spectrum. We discuss them in this section, providing an outlook for the future of this emerging research field.

Robustness vs cost trade-offs. Several works discuss the existence of a trade-off between the cost of a CE and its robustness. Specifically, increasing cost is often discussed as a necessary condition for improving a CE's robustness [Upadhyay *et al.*, 2021; Pawelczyk *et al.*, 2023a]. Increasing the cost of a CE typically implies steering the CE search away from the decision boundaries of a model. This often results in increased class scores, which is a sufficient condition for increasing (some forms of) robustness in linear models [Dominguez-Olmedo *et al.*, 2022; Upadhyay *et al.*, 2021; Bui *et al.*, 2022]. However, the understanding of this trade-off within the context of non-linear models, such as neural networks, is still limited. For example, [Jiang *et al.*, 2023a] show that several robust methods often find less costly CEs than non-robust alternatives. Theoretical results have been obtained on bounding the maximum cost increase needed to achieve robustness [Pawelczyk *et al.*, 2023a; Guyomard *et al.*, 2023]. However, we argue that more research is needed to advance our understanding of robustness-cost trade-offs.

One form of robustness to fit them all? Existing robustness notions may share some similarities, depending on the types of models and methods considered. For instance, when considering tree-based models, the problem of guaranteeing robustness against MC and MM tends to converge and algorithms developed for one setting might be applicable to the other. However, the same considerations may not apply to other robustness notions. For example, [Maragno *et al.*, 2023] target robustness against MC and NE and demonstrate that, at least in the case of linear models, these two notions are orthogonal in general. Furthermore, [Krishna *et al.*, 2022] highlights an interesting interplay between robustness of CEs and general adversarial robustness of a model, showing that popular training techniques to robustify neural networks may also lead to increased robustness in CEs. This begs the question as to whether existing robustness notions, including adversarial robustness, are strictly disjoint. However, their potential connections are largely unexplored. We advocate for more in-depth studies to shed light on this promising research area.

Links to fairness. Some notions of CE robustness appear to have strong connections with existing literature on CE fairness. This is especially true for robustness against IC. In simple words, robustness against IC requires that *similar CEs be given to similar individuals*, whereas CE fairness requires that *recourse of similar cost be generated for similar individuals* (modulo differences on protected features). Therefore, robustness against IC only requires similarity between CEs without constraining their cost wrt specific features. Despite this difference, recent work by [Ehyaei *et al.*, 2023] showed that robustness often implies fairness, complementing existing results [Slack *et al.*, 2021; Gupta *et al.*, 2019; Sharma *et al.*, 2020; von Kügelgen *et al.*, 2022] and paving the way for new research directions at the intersection between CEs, robustness and fairness.

Lack of standardised benchmarks. Most works only include a limited number ($\leq$ 2) of robust CE baseline methods in their empirical studies. This is understandable since the topic emerged only recently, but it hinders understanding of how practical each method is. For example, optimisation methods based on MIP are able to generate exact solutions with strong guarantees, but this typically comes with higher computational cost. On the other hand, gradient-based methods may provide suboptimal results but can benefit from the highly-parallelised deep learning libraries. Additionally, several studies target more than one property concurrently, e.g. robustness and plausiblity [Pawelczyk *et al.*, 2020; Jiang *et al.*, 2023a], which further complicates empirical comparisons. Going forward, we argue that intensive research effort should be put into developing standardised libraries to evaluate robustness. Similar initiatives have been pursued within the CE arena, e.g. [Pawelczyk *et al.*, 2021], and could provide a solid starting point for this.

Lack of user studies. Robustness is typically framed as a functional requirement to be evaluated using mechanistic metrics. However, a lack of robustness in CEs has the potential to weaken their justifiability and thus jeopardise their explanatory function [Hancox-Li, 2020]. None of the studies reported in this survey have conducted user experiments to explore this phenomenon, which we argue should be given more prominence and potentially guide the development of novel algorithmic solutions in the future.

Acknowledgements

Jiang, Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Leofante is supported by an Imperial College Research Fellowship grant. Rago and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Any views or opinions expressed herein are solely those of the authors listed.

References

[Artelt *et al.*, 2021] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *IEEE SSCI*, 2021.

[Black *et al.*, 2022a] Emily Black, Klas Leino, and Matt Fredrikson. Selective ensembles for consistent predictions. In *ICLR*, 2022.

[Black *et al.*, 2022b] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT*, 2022.

[Black *et al.*, 2022c] Emily Black, Zifan Wang, and Matt Fredrikson. Consistent counterfactuals for deep models. In *ICLR*, 2022.

[Breiman, 2001] Leo Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.

[Bui *et al.*, 2022] Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. Counterfactual plans under distributional ambiguity. In *ICLR*, 2022.

[Bui *et al.*, 2023] Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Coverage-validity-aware algorithmic recourse. *arXiv:2311.11349*, 2023.

[Coston *et al.*, 2021] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *ICML*, 2021.

[da Silva and Bertini, 2022] Ariel Tadeu da Silva and João Roberto Bertini. Using the k-associated optimal graph to provide counterfactual explanations. In *FUZZ-IEEE*, 2022.

[Dominguez-Olmedo *et al.*, 2022] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *ICML*, 2022.

[Dutta *et al.*, 2022] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *ICML*, 2022.

[Ehyaei *et al.*, 2023] Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schölkopf, and Setareh Maghsudi. Robustness implies fairness in causal algorithmic recourse. In *FAccT*, 2023.

[Ferrario and Loi, 2022] Andrea Ferrario and Michele Loi. The robustness of counterfactual explanations over time. *IEEE Access*, 10:82736–82750, 2022.

[Fonseca *et al.*, 2023] João Fonseca, Andrew Bell, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. Setting the right expectations: Algorithmic recourse over time. In *EAAMO*, 2023.

[Forel *et al.*, 2022] Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Robust counterfactual explanations for random forests. *arXiv:2205.14116*, 2022.

[Guo *et al.*, 2023] Hangzhi Guo, Feiran Jia, Jinghui Chen, Anna Cinzia Squicciarini, and Amulya Yadav. Rocoursenet: Robust training of a prediction aware recourse model. In *CIKM*, 2023.

[Gupta *et al.*, 2019] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv:1909.03166*, 2019.

[Guyomard *et al.*, 2023] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Generating robust counterfactual explanations. In *ECML PKDD*, 2023.

[Hada and Carreira-Perpiñán, 2021] Suryabhan Singh Hada and Miguel Á. Carreira-Perpiñán. Exploring counterfactual explanations for classification and regression trees. In *ECML PKDD Workshops*, 2021.

[Hamman *et al.*, 2023] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *ICML*, 2023.

[Hancox-Li, 2020] Leif Hancox-Li. Robustness in machine learning explanations: does it matter? In *FAT**, 2020.

[Jiang *et al.*, 2023a] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In *ACML*, 2023.

[Jiang *et al.*, 2023b] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In *AAAI*, 2023.

[Jiang *et al.*, 2024] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Recourse under model multiplicity via argumentative ensembling. In *AAMAS*, 2024.

[Kanamori *et al.*, 2023] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Yuichi Ike. Counterfactual explanation with missing values. *arXiv:2304.14606*, 2023.

[Karimi *et al.*, 2022] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM CSUR*, 55(5):1–29, 2022.

[Krishna *et al.*, 2022] Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. On the impact of adversarially robust models on algorithmic recourse. In *NeurIPS Workshops*, 2022.

[Krishna *et al.*, 2023] Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. In *ICML*, 2023.

[Leofante and Lomuscio, 2023] Francesco Leofante and Alessio Lomuscio. Robust explanations for human-neural multi-agent systems with formal verification. In *EUMAS*, 2023.

[Leofante and Potyka, 2024] Francesco Leofante and Nico Potyka. Promoting counterfactual robustness through diversity. In *AAAI*, 2024.

[Leofante *et al.*, 2023] Francesco Leofante, Elena Botoeva, and Vineet Rajani. Counterfactual explanations and model multiplicity: a relational verification view. In *KR*, 2023.

[Maragno *et al.*, 2023] Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, S. Ilker Birbil, and Dick den Hertog. Finding regions of counterfactual explanations via robust optimization. *arXiv:2301.11113*, 2023.

[Marx *et al.*, 2020] Charles T. Marx, Flávio P. Calmon, and Berk Ustun. Predictive multiplicity in classification. In *ICML*, 2020.

[Mishra *et al.*, 2021] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv:2111.00358*, 2021.

[Mochaourab *et al.*, 2021] Rami Mochaourab, Sugandh Sinha, Stanley Greenstein, and Panagiotis Papapetrou. Robust explanations for private support vector machines. *arXiv:2102.03785*, 2021.

[Mohammadi *et al.*, 2021] Kiarash Mohammadi, Amir-Hossein Karimi, Gilles Barthe, and Isabel Valera. Scaling guarantees for nearest counterfactual explanations. In *AIES*, 2021.

[Nguyen *et al.*, 2022] Tuan-Duy H. Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. Robust bayesian recourse. In *UAI*, 2022.

[Nguyen *et al.*, 2023] Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Distributionally robust recourse action. In *ICLR*, 2023.

[Pawelczyk *et al.*, 2020] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *UAI*, 2020.

[Pawelczyk *et al.*, 2021] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. CARLA: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. In *NeurIPS Datasets and Benchmarks*, 2021.

[Pawelczyk *et al.*, 2022] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *AISTATS*, 2022.

[Pawelczyk *et al.*, 2023a] Martin Pawelczyk, Teresa Datta, Johannes van den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. In *ICLR*, 2023.

[Pawelczyk *et al.*, 2023b] Martin Pawelczyk, Tobias Leemann, Asia Biega, and Gjergji Kasneci. On the trade-off between actionable explanations and the right to be forgotten. In *ICLR*, 2023.

[Raman *et al.*, 2023] Natraj Raman, Daniele Magazzeni, and Sameena Shah. Bayesian hierarchical models for counterfactual estimation. In *AISTATS*, 2023.

[Rawal *et al.*, 2020] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv:2012.11788*, 2020.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *KDD*, 2016.

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

[Sharma *et al.*, 2020] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *AIES*, 2020.

[Sharma *et al.*, 2022] Shubham Sharma, Alan H. Gee, Jette Henderson, and Joydeep Ghosh. FASTER-CE: fast, sparse, transparent, and robust counterfactual explanations. *arXiv:2210.06578*, 2022.

[Slack *et al.*, 2021] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In *NeurIPS*, 2021.

[Tolomei *et al.*, 2017] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *KDD*, 2017.

[Upadhyay *et al.*, 2021] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In *NeurIPS*, 2021.

[Virgolin and Fracaros, 2023] Marco Virgolin and Saverio Fracaros. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artif. Intell.*, 316:103840, 2023.

[von Kügelgen *et al.*, 2022] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *AAAI*, 2022.

[Wachter *et al.*, 2017] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[Wang *et al.*, 2023a] Ming Wang, Daling Wang, Wenfang Wu, Shi Feng, and Yifei Zhang. T-COL: generating counterfactual explanations for general user preferences on variable machine learning systems. *arXiv:2309.16146*, 2023.

[Wang *et al.*, 2023b] Yongjie Wang, Hangwei Qian, Yongjie Liu, Wei Guo, and Chunyan Miao. Flexible and robust counterfactual explanations with minimal satisfiable perturbations. In *CIKM*, 2023.

[Zhang *et al.*, 2023] Songming Zhang, Xiaofeng Chen, Shiping Wen, and Zhongshan Li. Density-based reliable and robust explainer for counterfactual explanation. *Expert Syst. Appl.*, 226:120214, 2023.