Robust Algorithmic Recourse Under Model Multiplicity With Probabilistic Guarantees

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta

***Abstract*—There is an emerging interest in generating robust algorithmic recourse that would remain valid if the model is updated or changed even slightly. Towards finding robust algorithmic recourse (or counterfactual explanations), existing literature often assumes that the original model m and the new model M are bounded in the parameter space, i.e., $\|\text{Params}(M) - \text{Params}(m)\| < \Delta$. However, models can often change significantly in the parameter space with little to no change in their predictions or accuracy on the given dataset. In this work, we introduce a mathematical abstraction termed *naturally-occurring* model change, which allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. Next, we propose a measure – that we call *Stability* – to quantify the robustness of counterfactuals to potential model changes for differentiable models, e.g., neural networks. Our main contribution is to show that counterfactuals with sufficiently high value of *Stability* as defined by our measure will remain valid after potential "naturally-occurring" model changes with high probability (leveraging concentration bounds for Lipschitz function of independent Gaussians). Since our quantification depends on the local Lipschitz constant around a data point which is not always available, we also examine estimators of our proposed measure and derive a fundamental lower bound on the sample size required to have a precise estimate. We explore methods of using stability measures to generate robust counterfactuals that are close, realistic, and remain valid after potential model changes. This work also has interesting connections with model multiplicity, also known as the Rashomon effect.**

***Index Terms*—Counterfactual explanation, model multiplicity, algorithmic recourse, explainable AI, responsible machine learning.**

I. Introduction

ALGORITHMIC recourse and counterfactual explanations [2], [3], [4] have garnered significant interest in various high-stakes applications, such as lending, hiring, etc. Algorithmic recourse aim to guide an applicant on how they can change a model outcome by providing suggestions for improvement. Given an original data-point (e.g., an applicant who is denied a loan), the goal is to try to find a point on the other (desired) side of the decision boundary (a hypothetical applicant who is approved for the loan) which also satisfies several other preferred constraints, such as, (i) proximity to the original point; (ii) changes in as few features as possible; and (iii) conforming to the data manifold. Such a data-point that alters the model decision is widely referred to as a "counterfactual explanation," as illustrated in Fig. 1.

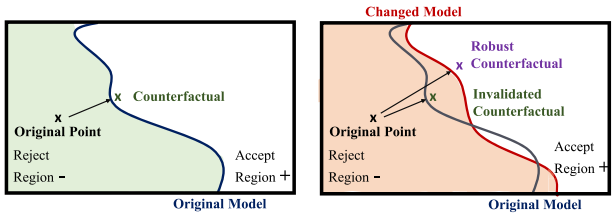


Fig. 1. *(left)* We depict the original model's decision boundary. Given a point in the rejected region, a counterfactual explanation (counterfactual) typically refers to the closest point on the accepted side. Counterfactual explanations provide guidance on actions for recourse. *(right)* We depict a changed model where the initial counterfactual becomes invalid. A robust counterfactual should remain valid even when the model changes while maintaining proximity to the original data point.

However, in several real-world scenarios, such as credit lending, the models have to be updated due to various reasons [5], [6], [7], e.g., to retrain on a few additional data points, change the hyper-parameters or seed, or transition to a different model class [8]. Such model changes can often cause the counterfactuals to become invalid because typically they are quite close to the original data point, and hence, also quite close to the decision boundary. For instance, suppose the counterfactual explanation suggests an applicant to increase their income by $10K$ to get approved for a loan, and they act upon that, but now, due to updates to the original model, they are still denied by the updated model (see Fig. 1).

If recourse becomes invalid due to model updates, this can lead to confusion and distrust in the use of algorithms in high-stakes applications altogether. Users would typically act on the suggested counterfactuals over a period of time, e.g., increase their income for credit lending, but only to find that it is no longer enough since the model has slightly changed (perhaps due to retraining with a new seed or hyperparameter). This cycle of invalidation and regenerating new counterfactuals can not only be frustrating and time-consuming for users but also

Manuscript received 26 October 2023; revised 13 February 2024; accepted 29 April 2024. Date of publication 15 May 2024; date of current version 24 June 2024. This work was supported in part by the JPMorgan Faculty Award and Startup Funding from the University of Maryland. This work was presented in part at the International Conference on Machine Learning (ICML) 2023. *(Corresponding author: Faisal Hamman.)*

Faisal Hamman, Erfaun Noorani, and Sanghamitra Dutta are with the Department of Electrical and Computer Engineering, University of Maryland at College Park, College Park, MD 20742 USA (e-mail: fhamman@umd.edu; enoorani@umd.edu; sanghamd@umd.edu).

Saumitra Mishra and Daniele Magazzeni are with JP Morgan AI Research, E14 5JP London, U.K. (e-mail: saumitra.mishra@jpmorgan.com; daniele.magazzeni@jpmorgan.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/JSAIT.2024.3401407, provided by the authors.

Digital Object Identifier 10.1109/JSAIT.2024.3401407

potentially hurt an institution's reputation. This motivates our primary question:

How to provide theoretical guarantees on the robustness of counterfactuals to potential model changes?

Towards addressing this question, in this work, we introduce the abstraction of "naturally-occurring" model change for differentiable models. Our abstraction allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. This abstraction is centered on the inherent need for model explanations to remain robust against variations such as weight initialization or minor adjustments in hyperparameters (changing seed) [5], [6], [9], [10], [11], [12]. Another insightful angle is the concept of machine unlearning, particularly in light of regulatory frameworks like the GDPR [13]. The *right to be forgotten* necessitates the removal of an individual's data upon request, potentially leading to model updates. These updates could, in turn, impact the validity of previously issued explanations, thus challenging the balance between the *right to explanation* and the *right to be forgotten* [7].

This abstraction motivates a measure of robustness for counterfactuals that arrives with provable probabilistic guarantees on their validity under naturally-occurring model change. We also introduce the notion of adversarial or targeted model change and provide an impossibility result for such model change. We examine estimators of our proposed measure and derive a fundamental lower bound on the sample size required to have a precise estimate. Next, by leveraging this computable estimator, we explore methods of using stability measures to generate robust counterfactuals that are close, realistic, and remain valid after potential model changes. Our experimental results validate our theoretical understanding and illustrate the efficacy of our proposed algorithms. We summarize our contributions here:

Abstraction of "naturally-occurring" model change for differentiable models: Existing literature [5], [6] on robust counterfactuals often assumes that the original model m and the new model M are bounded in the parameter space, i.e., $\|\text{Params}(M) - \text{Params}(m)\| < \Delta$. Building on [9] for tree-based models, we note that models can often change significantly in the parameter space with little to no change on their predictions or accuracy on the given dataset. To capture this, we introduce an abstraction (see Definition 5), that we call *naturally-occurring* model change, which instead allows for arbitrary changes in the parameter space such that the change in predictions on points that lie on the data manifold is limited. Our proposed abstraction of naturally-occurring model change also has interesting connections with predictive/model multiplicity, also known as, the Rashomon Effect [14], [15].

We also make a clear distinction between our proposed naturally-occurring and *adversarial* model change. Under the adversarial model change, we provide an impossibility result (Theorem 2) that given any counterfactual for a model, one can always design a new model that is quite similar to the original model and that renders that particular counterfactual invalid. However, in this work, our focus is on non-targeted model change such as retraining on a few additional data points, changing some hyperparameters or seed, etc. which is captured in "naturally-occurring" model change (see Definition 5).

A measure of robustness with probabilistic guarantees on validity: Next, we propose a novel mathematical measure – that we call Stability – to quantify the robustness of counterfactuals to potential model changes. Stability of a counterfactual $x \in \mathbb{R}^d$ with respect to a model $m(\cdot)$ is given by:

$$R_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} (m(x_i) - \gamma_x \|x - x_i\|),$$

where $N_{x,k}$ is a set of k points in $\mathbb{R}^d$ drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$ with I_d being the identity matrix, and γ_x is the local Lipschitz constant of the model $m(\cdot)$ around x (see Definition 6).

Our main contribution is to provide a theoretical guarantee (Theorem 3) that counterfactuals with a sufficiently high value of Stability (as defined by our measure) will remain valid with high probability after *naturally-occurring* model changes. In Theorem 3, we assume a strict upper bound $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$, where $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. We generalize this by introducing a probabilistic bound $\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon') \leq \delta$ (see Corollary 1). Further, we characterize this bound δ under the conditions of naturally-occurring model change and specific assumptions about the expected variability in a data point's neighborhood (see Assumption 1). Leveraging this characterization, we introduce Lemma 3l, which serves as the foundation for proving Theorem 4. This theorem offers a comprehensive probabilistic guarantee on the validity of counterfactuals with a high value of Stability (as per our measure) on the data manifold. Our results leverage concentration bounds for Lipschitz functions of independent Gaussian random variables (see Lemma 2).

Estimators of Stability and Their Properties: Since our proposed Stability measure depends on the local Lipschitz constant which is not always known, we also examine two practical estimators of our measure: (1) The *Stability-Lipschitz estimator* (see Definition 7) aims to approximate the local Lipschitz constant using $\hat{\gamma}_x = \max_{x_i \in N_{x,k}} \frac{|m(x) - m(x_i)|}{\|x - x_i\|}$. This captures the *worst-case variability* in the model's outputs in the neighborhood of x. We also derive a fundamental lower bound on sample size to ensure that the *Stability-Lipschitz estimator* approximates the true stability within an ϵ error (see Theorem 5). (2) We introduce the *Stability-Soft estimator* (see Definition 8) as a less computationally expensive, albeit less accurate, alternative for estimating stability:

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k} \sum_{x_i \in N_{x,k}} (m(x_i) - |m(x) - m(x_i)|).$$

The first term essentially captures the mean value of the model output in a region around it (higher mean is expected to be more robust and reliable). The second term captures the local *average variability* of the model output around it (lower variability is expected to be more reliable). This intuition is in alignment with the results in [9] for tree-based models (see Section III-D).

Generating Robust Counterfactuals Using Stability: We explore strategies for using stability measures to generate robust counterfactuals for neural networks. We introduce T-Rex:I (Algorithm 1), which finds robust counterfactuals

that are close to the original data point. T-Rex:I can be integrated into any base technique for generating counterfactuals to improve robustness. We also propose T-Rex:NN (Algorithm 2), which generates robust counterfactuals that are data-supported (along the lines of [9] for tree-based models). We propose a hybrid method T-Rex:Hybrid (Algorithm 3) that focuses on finding robust counterfactuals on the data manifold, making them more realistic. The hybrid method employs generative models to learn a latent representation of the data manifold, within which we conduct our search for counterfactuals.

Experimental Results: We conduct experiments on several benchmark datasets, namely, HELOC [16], German Credit, Cardiotocography (CTG), Adult [17], and Taiwanese Credit [18] to support our theoretical findings (see Section V). Our experiments show that T-Rex:I can improve robustness for neural networks without significantly increasing the cost, and T-Rex:NN consistently generates counterfactuals that are similar to the data manifold, as measured using the Local Outlier Factor (LOF). The Local Outlier Factor (LOF) is a popular evaluation metric that assesses the relative isolation of a data point within their local neighborhood to identify anomalies (see Definition 4).

Related Works: Algorithmic recourse and counterfactual explanations have seen growing interest in recent years [2], [3], [11]. Regarding their robustness to model changes, [19], [20], [21] argue that counterfactuals situated on the data manifold are more likely to be more robust than the closest counterfactuals. Later, [9] demonstrate that generating counterfactuals on the data manifold may not be sufficient for robustness. While the importance of robustness in local explanation methods has been emphasized [22], the problem of specifically generating robust counterfactuals has been less explored, with the notable exceptions of some recent works [5], [6], [9], [10], [23]. In [5], the authors propose an algorithm called ROAR that uses min-max optimization to find the *closest* counterfactuals that are also robust. In [23], the focus is on analytical trade-offs between validity and cost. Reference [10] introduces a method for identifying close and robust counterfactuals based on a framework that utilizes interval neural networks. Reference [6] propose that local Lipschitzness can be leveraged to generate consistent counterfactuals and propose an algorithm called Stable Neighbor Search to generate consistent counterfactuals for neural networks. Our research builds on this perspective and further performs Gaussian sampling around the counterfactual, leading to a novel estimator for which we are also able to provide probabilistic guarantees going beyond the bounded model change assumption. Furthermore, examining all three performance metrics, namely, cost, validity (robustness), and likeness to the data-manifold has received less attention with the notable exception of [9] but they focus only on tree-based models (non-differentiable). Following our conference publication, [24] proposed a robust optimization framework to generate provably robust and plausible counterfactuals for neural networks and proved its soundness, completeness, and convergence. We also refer to [25] for a survey.

We note that [26], [27] propose an alternate perspective of robustness in explanations (called L-stability in [27]) which is built on similar individuals receiving similar explanations. [28], [29], [30] focus on finding counterfactuals that are robust to small input perturbations (noisy counterfactuals). In contrast, our focus is on counterfactuals remaining valid after some changes to the model, and providing theoretical guarantees thereof.

Our work also shares interesting conceptual connections with a body of work on model multiplicity or predictive multiplicity, also known as the Rashomon effect [14], [15], [31], [32]. [14] suggested that models can be very different from each other but have almost similar performance on the data manifold. The term predictive multiplicity was suggested by [15] which defined it as the ability of a prediction problem to admit competing models with conflicting predictions. Reference [31] investigates ways to leverage model multiplicity beneficially in model selection processes while simultaneously addressing its concerning implications. Reference [33] offered a framework for measuring predictive multiplicity in classification, introducing measures that encapsulate the variation in risk estimates over the ensemble of competing models. Reference [32] unveiled a novel metric, Rashomon Capacity, for measuring predictive multiplicity in probabilistic classification. Our proposed abstraction of naturally-occurring model change in this work can be viewed as a fresh perspective on model multiplicity that further emphasizes the models that are more likely to occur.

II. Preliminaries

Let $m(\cdot) : \mathbb{R}^d \to [0, 1]$ denote the original machine learning model that takes a d-dimensional input value and produces an output probability lying between 0 and 1. The final decision is denoted by $\mathbb{1}(m(x) \geq 0.5)$ where $\mathbb{1}(\cdot)$ denotes the indicator function.

Definition 1 ($\gamma-$Lipschitz): A function $m(\cdot)$ is said to be $\gamma-$Lipschitz if $|m(x) - m(x')| \leq \gamma \|x - x'\| \ \forall \ x, x' \in \mathbb{R}^d$.

Here $\|\cdot\|$ denotes the Euclidean norm, i.e., for $u \in \mathbb{R}^d$, we have $\|u\| = \sqrt{u_1^2 + u_2^2 + \ldots + u_d^2}$. In Remark 2, we also discuss relaxations to local Lipschitz constants from global Lipschitz constants. We denote the updated or changed model as $M(\cdot) : \mathbb{R}^d \to [0, 1]$ where M is a random entity. We mostly use capital letters to denote random entities, e.g., M, X, etc., and small letters to denote non-random entities, e.g., m, x, γ, n, etc.

Definition 2 (Closest Counterfactual $\mathcal{C}_p(x, m)$): Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its closest counterfactual (in terms of l_p-norm) with respect to the model $m(\cdot)$ is defined as a point $x' \in \mathbb{R}^d$ that minimizes the l_p norm $\|x - x'\|_p$ such that $m(x') \geq 0.5$.

$$\mathcal{C}_p(x, m) = \arg\min_{x' \in \mathbb{R}^d} \|x - x'\|_p \text{ such that } m\big(x'\big) \geq 0.5.$$

When one is interested in finding counterfactuals by changing as few features as possible, the l_1 norm is used (enforcing a sparsity constraint). These are called *sparse* counterfactuals [19]. However, such closest counterfactuals often fall too far from the data manifold, resulting in unrealistic and

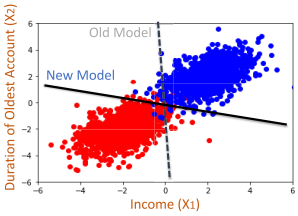


Fig. 2. Models can often change drastically in the parameter space causing little to no change in the actual decisions on the points on the data manifold.

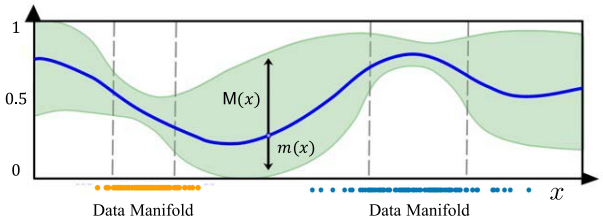


Fig. 3. Illustrates our proposed abstraction of naturally-occurring model change: The distribution of the changed model outputs $M(x)$ (stochastic) is centered around the original model output $m(x)$. The points specifically lying on the data-manifold acting as anchors without much change as they exhibit lower variance in model outputs compared to points outside the manifold. This visualization also connects with the Rashomon effect, encapsulating the diverse yet similarly accurate models that can be learned from a given dataset.

anomalous instances, as noted in [3], [11], [19], [20], [21], [34]. This highlights the need for generating counterfactuals that lie on the data manifold.

Definition 3 (Closest Data-Manifold Counterfactual $\mathcal{C}_{p,\mathcal{X}}(x,m)$): Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its closest data-manifold counterfactual $\mathcal{C}_{p,\mathcal{X}}(x,m)$ with respect to the model $m(\cdot)$ and data manifold $\mathcal{X} \subseteq \mathbb{R}^d$ is defined as a point $x' \in \mathcal{X}$ that minimizes the l_p norm $\|x - x'\|_p$ such that $m(x') \geq 0.5$.

$$\mathcal{C}_{p,\mathcal{X}}(x,m) = \arg\min_{x' \in \mathcal{X}} \|x - x'\|_p \text{ such that } m\big(x'\big) \geq 0.5.$$

In order to assess the similarity or anomalous nature of a point concerning the given dataset $\mathcal{S} \subseteq \mathcal{X}$, various metrics can be employed, e.g., K-nearest neighbors, Mahalanobis distance, Kernel density, LOF. These metrics play a crucial role in understanding the quality of counterfactual explanations generated by a model. One metric widely used in literature [9], [19], [20] is the Local Outlier Factor (LOF).

Definition 4 (Local Outlier Factor [35]): For $x \in \mathcal{S}$, let $L_k(x)$ be its k-Nearest Neighbors (k-NN) in $\mathcal{S}$. The k-reachability distance rd_k of x with respect to x' is defined by $rd_k(x,x') = \max\{\delta(x,x'), d_k(x')\}$, where $d_k(x')$ is the distance δ between x' and its k-th nearest instance on $\mathcal{S}$. The k-local reachability density of x is defined by $lrd_k(x) = |L_k(x)|(\sum_{x' \in L_k(x)} rd_k(x,x'))^{-1}$. Then, the k-LOF of x on $\mathcal{S}$ is defined as follows:

$$LOF_{k,\mathcal{S}}(x) = \frac{1}{|L_k(x)|} \sum_{x' \in L_k(x)} \frac{lrd_k\big(x'\big)}{lrd_k(x)}.$$

Here, $\delta(x,x')$ is the distance between two d-dimensional feature vectors. The LOF Predicts -1 for anomalous points and $+1$ for inlier points.

Goals: In this work, our main goal is to provide *probabilistic guarantees* on the robustness of counterfactuals to potential model changes for differential models such as neural networks. Towards achieving this goal, our objective involves: (i) introducing an abstraction that rigorously defines the class of model changes that we are interested in; and (ii) establishing a measure, denoted as $R_\Phi(x,m)$, for a counterfactual x and a given model $m(\cdot)$, that quantifies its robustness to potential model changes. Here, Φ represents the hyperparameters of the robustness measure. Ideally, we desire that the measure $R_\Phi(x,m)$ should be high if the counterfactual x is less likely to be invalidated by potential model changes. We seek to provide: (i) theoretical guarantees on the validity of counterfactuals with sufficiently high value of $R_\Phi(x,m)$ with a deeper understanding of the guarantee under various assumptions; (ii) various estimators $\hat{R}_\Phi(x,m)$ and study fundamental requirements needed to ensure precise estimates; and (iii) strategies to incorporate our measure into an algorithmic framework for generating robust counterfactuals while meeting other requirements, such as low cost or likeness to the data manifold.

III. Main Contributions

In this section, we first introduce our proposed abstraction of *naturally-occurring* model change and then propose a novel measure – that we call *Stability* – to quantify the robustness of counterfactuals to potential model changes. We derive a theoretical guarantee that counterfactuals that have a sufficiently high value of *Stability* will remain valid after potential *naturally-occurring* model change with high probability. But since our quantification would depend on the local Lipschitz constant around a data point, which is not always known, we also examine estimators of our proposed measure and demonstrate its applicability.

A. Naturally-Occurring Model Change

A popular assumption in existing literature [5], [6] to quantify potential model changes is to assume that the model changes are bounded in the parameter space, i.e.,

$$\|\text{Params}(M) - \text{Params}(m)\| < \Delta \text{ for a constant } \Delta.$$

Here, Params(M) denote the parameters of the model M, e.g., weights of a neural network. However, we note that models can often change drastically in the parameter space causing little to no change in the actual decisions on the points on the data manifold (see Fig. 2 for an example). In this work, we avoid the bounded-model-change assumption and instead introduce the notion of a naturally-occurring model change as defined in Definition 5. Our abstraction allows for arbitrary model changes such that the change in predictions on points that lie on the data manifold is limited (see Fig. 3).

This abstraction is motivated from the observation that points residing in the data-manifold generally demonstrate reduced variance in model outputs compared to those outside the manifold. This behavior can be attributed to the fact that during training, the model is predominantly exposed to data points from the data-manifold, leading to higher confidence

in its predictions in that regions. Consequently, the model's behavior for points outside the manifold can be unpredictable (also see Fig. 4).

Definition 5 (Naturally-Occurring Model Change): A naturally-occurring model change is defined as follows:

1) $\mathbb{E}[M(X)|X = x] = \mathbb{E}[M(x)] = m(x)$ where the expectation is over the randomness of M given a fixed value of $X = x \in \mathbb{R}^d$.
2) Whenever $m(x)$ is γ_m-Lipschitz, any updated model $M(x)$ is also $\gamma-$Lipschitz for some constant γ. Note that, this constant γ does not depend on M since we may define γ to be an upper bound on the Lipschitz constants for all possible M as well as m.
3) $\text{Var}[M(X)|X = x] = \text{Var}[M(x)] = \nu_x$ which depends on the fixed value of $X = x \in \mathbb{R}^d$. Furthermore, ν_x is small for x lying on the data manifold $\mathcal{X}$.

Closely connected to naturally-occurring model change is the idea of the Rashomon effect, alternatively known as predictive or model multiplicity. [14], [15], [19], [32] which suggests that models can be very different from each other but have almost similar performance on the data manifold. Model multiplicity arises when models trained on the same dataset (with different weight initialization) assign varying predictions to a given sample. This is mainly because the primary objective of training is to minimize empirical risk loss. Consequently, several models can be distinctly different (even yielding opposing predictions) but still maintain comparable accuracy levels. These models are generally more confident on the data manifold, e.g., $\frac{1}{n}\sum_{i=1}^{n}|M(x_i) - m(x_i)|$ is small when the points x_i lie on the data manifold. Under the naturally-occurring model change, this holds in expectation:

Theorem 1 (Connection to Rashomon Effect): For points $x_1, \ldots, x_n \in \mathcal{X}$ (lying on the data-manifold) under naturally-occurring model change, the following holds:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|M(x_i) - m(x_i)|\right] \leq \sqrt{\nu}, \tag{1}$$

where $\nu = \frac{1}{n}\sum_{i=1}^{n}\nu_{x_i}$.

Rashomon effect [14], [15], [19], [32] or model multiplicity typically refers to the phenomenon of a diverse models yielding similar accuracy levels on the same dataset. In this context, the Rashomon set [32] aims to characterize the entire set of models whose predictions differ by a small amount with additional constraints, e.g., models within a certain model class. We adopt a probabilistic stance on model multiplicity. Our proposed abstraction of naturally-occurring model change attempts to characterize the distribution of the models which are more likely to occur naturally rather than the entire set. Theorem 1 ties to the Rashomon effect by demonstrating how, under naturally-occurring model multiplicity, different models (despite their varied structures and predictions) can exhibit a surprisingly consistent performance when evaluating on points lying on the data manifold. This consistency is quantified by the expectation that the absolute difference in predictions across models is bounded, implying that a diverse set of models can indeed yield similar accuracy levels on the same dataset. Thus, Definition 5 might be better suited over boundedness in the parameter space. Proof of Theorem 1 is in Appendix B.

Remark 1 (Adversarial Model Change): In contrast to naturally-occurring model change, we also introduce adversarial model change (targeted) which essentially refers to a model change that is more deliberately targeted to make a particular counterfactual invalid.

Theorem 2 (Impossibility Under Adversarial Change): Given a model and a counterfactual, one can always design another similar model such that the particular targeted counterfactual can be invalidated.

The proof, provided in Appendix E, shows that there is a new model $M(x) = m(x)$ almost everywhere except at or around the targeted point x', i.e., $M(x') = 1 - m(x')$. Such a model could emerge from training with a poisoned data point or as a split model. These scenarios represent adversarial manipulations rather than naturally occurring model variations, and illustrate non-standard model behaviors that we distinguish from the naturally occurring model changes.

B. Measure of Robustness of a Counterfactual

Definition 6 (Stability): Given a model $m(\cdot)$, the stability of a counterfactual $x \in \mathbb{R}^d$ is defined as follows:

$$R_{k,\sigma^2}(x, m) = \frac{1}{k}\sum_{x_i \in N_{x,k}}(m(x_i) - \gamma\|x - x_i\|), \tag{2}$$

where $N_{x,k}$ is a set of k points drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2\text{I}_d)$ with I_d being the identity matrix, and γ is an upper bound on the Lipschitz constant for all models $M(\cdot)$ under naturally-occurring change.

Our stability measure generalizes to any predictive class, as it can be fundamentally tied to the confidence of predicting a class (in our case class 1). For cases where a prediction needs to shift from 1 to 0, the concept can seamlessly apply by considering the logits (or softmax outputs) for predicting class 0. This could also extend to multi-class classification providing logits for each class.

Since obtaining the precise Lipschitz constant for neural networks is a complex task; hence, we operate under the assumption of a finite upper bound on the Lipschitz continuity for both our original model and changed models. This assumption might be more likely to hold if all the models belong to the same model class with roughly similar architectures. Furthermore, in practice models can also be trained using regularization to prevent their Lipschitz constant from being very high [36].

Remark 2 (Relaxations to Local Lipschitz): While we prove our theoretical result (Theorem 3) with the global Lipschitz constant γ, we can relax this to local Lipschitz constants γ_x, around a given point x. This is because we sample from a Gaussian centered around the point x and hence mainly capture the variability around x. So most points will be very close to x but a few points can still lie far away. Potential extensions of our guarantees could apply to truncated Gaussian and uniform sampling methods, given their sub-Gaussian properties. This is because Lipschitz concentration inherently extends to sub-Gaussian random variables [37].

C. Probabilistic Guarantees on Validity

To justify stability as a measure of robustness for a counterfactual to natural-occurring model changes, we provide a probabilistic guarantee on the validity of the counterfactual in Theorem 3.

Theorem 3 (Probabilistic Guarantee): Let $X_1, X_2, \ldots, X_k$ be k iid random variables with distribution $\mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. Then, for any $\epsilon > 2\epsilon'$, a counterfactual $x \in \mathcal{X}$ under naturally-occurring model change satisfies:

$$\Pr\big(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon\big) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m + \gamma)^2\sigma^2}\right).$$

Probability is over the randomness of both M and X_i's.

This stability metric (see Definition 6) is a way to measure the robustness of counterfactuals that are subject to natural model changes (see Definition 5). The first term in the metric, represented by $\frac{1}{k}\sum_{i=1}^{k} m(X_i)$, captures the average model outputs for a group of points centered around the counterfactual x. The second term, represented by $\gamma\|x - X_i\|$, is an upper bound on the potential difference in outputs of any new model on the points x and X_i (Recall the Lipschitz property of M around the point x). Using our measure, the guarantee in Theorem 3 can be rewritten as:

$$\Pr\left(\frac{1}{k}\sum_{i=1}^{k} m(X_i) - M(x) \leq \frac{\gamma}{k}\sum_{i=1}^{k}\|x - X_i\| + \epsilon\right) \geq 1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma + \gamma_m)^2\sigma^2}\right).$$

This form of the inequality allows for the following interpretation of Theorem 3: The distance between the output of the new model on an input x, i.e., $M(x)$, and the average prediction of the neighborhood of the given input by the old model, i.e., $\frac{1}{k}\sum m(X_i)$ is upper bounded by ϵ-corrected, γ multiplied average distance of the datapoints within the neighborhood of the input x, i.e., $\frac{1}{k}\sum\|x - X_i\|$.

Proof Sketch: The complete proof of Theorem 3 is provided in Appendix C. Here, we include a proof sketch. Notice that, using the Lipschitz property of $M(\cdot)$ around x, we have $M(x) \geq M(X_i) - \gamma\|x - X_i\|$ for all X_i. Thus,

$$M(x) \geq \frac{1}{k}\sum_{i=1}^{k}(M(X_i) - \gamma\|x - X_i\|) \overset{(a)}{\geq} \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - \gamma\|x - X_i\|) - \epsilon, \tag{3}$$

where (a) holds from Lemma 1 with probability at least $1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$. [Deviation Bound]lembound Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. Then, under naturally-occurring model change, $\mathbb{E}[Z] = 0$. Moreover, for any $\epsilon > 2\epsilon'$,

$$\Pr(Z \geq \epsilon) \leq \exp\left(\frac{-k\epsilon^2}{8(\gamma + \gamma_m)^2\sigma^2}\right). \tag{4}$$

Proof Sketch: The proof of Lemma 1 leverages concentration bounds for Lipschitz functions of independent Gaussian random variables (see Lemma 2). The complete proof of Lemma 1 is provided in Appendix C. ■

Lemma 1 (Gaussian Concentration Inequality): Let $W = (W_1, W_2, \ldots, W_n)$ consist of n i.i.d. random variables belonging to $\mathcal{N}(0, \sigma^2)$, and $Z = f(W)$ be a γ-Lipschitz function, i.e., $|f(W) - f(W')| \leq \gamma\|W - W'\|$. Then:

$$\Pr(Z - \mathbb{E}[Z] \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\gamma^2\sigma^2}\right) \text{ for all } \epsilon > 0. \tag{5}$$

For the proof of Lemma 2 refer to [38, p.125]. Our robustness guarantee (Theorem 3) essentially states that $\Pr(M(x) \leq R_{k,\sigma^2}(x, m) - \epsilon) \leq \exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$ under naturally-occurring model change. For instance, if we find a counterfactual x such that $R_{k,\sigma^2}(x, m) - \epsilon$ is greater or equal to 0.5, then $M(x)$ would also be greater than 0.5 with high probability. The term $\exp\left(\frac{-k\epsilon^2}{8(\gamma+\gamma_m)^2\sigma^2}\right)$ decays with k.

In Theorem 3, we assume the bound $|\mathbb{E}[Z|M] - \mathbb{E}[Z]| < \epsilon'$. In Corollary 1, we relax this assumption to $\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon') \leq \delta$, i.e., the bound is relaxed to allow a probability of δ for the deviation to exceed ϵ' (see proof in Appendix C-B). For small δ, a high stability measure implies a high probability of being valid for changed models.

Corollary 1: Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Suppose $\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[Z]| > \epsilon') \leq \delta$. Then, for any $\epsilon > 2\epsilon'$, a counterfactual $x \in \mathcal{X}$ under naturally-occurring model change satisfies:

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq (1 - \delta)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m + \gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M and X_i's.

Under certain assumptions, we are able to characterize δ. For instance, in Assumption 1, we build on condition (3) of Definition 5 by further bounding the expected variance around the neighborhood of a point.

Assumption 1: Let x be a point that lies on the data manifold $\mathcal{X}$. Assume that the random variable X is drawn from a Gaussian distribution $\mathcal{N}(x, \sigma^2 I_d)$. Under these conditions, we make the following assumption:

$$\mathbb{E}_X[\mathrm{Var}(M(X)|X)] = \mathbb{E}_X[\nu_X] \leq \alpha \tag{6}$$

where α is a small constant. The expectations are over X, and the variance over M.

This assumption posits that the expected variance of the changed models' prediction around the neighborhood is bounded by a small constant α. Points residing on the data-manifold generally demonstrate reduced variance in model outputs compared to those outside the manifold since the model is predominantly exposed to training data points from the data-manifold, leading to higher confidence in its predictions in those regions (see Fig. 4 for illustration).

Leveraging this Assumption 1, we introduce Lemma 3, which serves as the foundation for proving Theorem 4 which offers a comprehensive probabilistic guarantee on the validity of counterfactuals with a high Stability value on the data manifold (see Appendix C-B for proof).

Lemma 2: Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Then, for all $\epsilon > 0$, under naturally-occurring model change and Assumption 1,

$$\Pr(|\mathbb{E}[Z|M] - \mathbb{E}[\mathbb{E}[Z|M]]| \geq \epsilon) \leq \frac{\alpha}{\epsilon^2}. \tag{7}$$

Theorem 4: Let $X_1, X_2, \ldots, X_k \sim \mathcal{N}(x, \sigma^2 I_d)$ on the data-manifold and $Z = \frac{1}{k}\sum_{i=1}^{k}(m(X_i) - M(X_i))$. Then, a counterfactual $x \in \mathcal{X}$ under Assumption 1 and naturally-occurring model change satisfies:

$$\Pr(M(x) \geq R_{k,\sigma^2}(x, m) - \epsilon) \geq \left(1 - \frac{\alpha}{\epsilon^2}\right)\left(1 - \exp\left(\frac{-k\epsilon^2}{8(\gamma_m+\gamma)^2\sigma^2}\right)\right).$$

Probability is over the randomness of both M and X_i's.

D. Estimators of Stability and Their Properties

Here, we provide practical estimators of stability measure since true stability (see Definition 6) relies on the Lipschitz constant γ (or the local Lipschitz constant γ_x around the point x), which is often unknown. We propose two practical estimators and study their properties.

Definition 7 (Stability-Lipschitz Estimator): Let $N_{x,k}$ be a set of k points drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$, the stability (Lipschitz Estimate) of a counterfactual $x \in \mathbb{R}^d$ is defined as follows:

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k}\sum_{x_i \in N_{x,k}} \big(m(x_i) - \hat{\gamma}_x \|x - x_i\|\big),$$
$$\text{where} \quad \hat{\gamma}_x = \max_{x_i \in N_{x,k}} \frac{|m(x) - m(x_i)|}{\|x - x_i\|}. \tag{8}$$

The Stability-Lipschitz Estimate aims to approximate the local Lipschitz constant γ_x through the term $\hat{\gamma}_x$. By design, this estimate focuses on capturing the worst-case variability in the local neighborhood of the point x.

Insight into the Stability-Lipschitz Relaxation: Through this localized assessment, our Stability-Lipschitz Estimate offers a fine-grained, yet computationally feasible, metric for stability. However, the accuracy of this estimate is closely tied to the number of samples k drawn from the local neighborhood of x. In essence, a sufficient k is crucial for the robust approximation of the local Lipschitz constant γ_x. We formally address this requirement in Theorem 5, where we derive a fundamental lower bound on k to ensure that the Stability-Lipschitz estimator approximates the true stability within an ϵ error.

Theorem 5 (Fundamental Lower Bound on Sample Size): Let $\mathcal{M}$ be a class of all models with Lipschitz constant γ in domain $[-T, T]^d \subset \mathbb{R}^d$ and bound on the second-order partial derivatives, i.e., $\forall m \in \mathcal{M}$, $|\frac{\partial^2 m}{\partial x_i \partial x_j}| \leq \psi$ for all $x \in \mathbb{R}^d$ and $i, j \in \{1, 2, \ldots, d\}$. If,

$$\sup_{m \in \mathcal{M}} \mathbb{E}\left[\left|\hat{R}_{k,\sigma^2}(x, m) - R_{k,\sigma^2}(x, m)\right|\right] < \epsilon,$$

then $k \geq \left(\frac{\sqrt{2\sigma^2}\, T \psi\, \Gamma(\frac{d+1}{2})}{9.69\epsilon\, \Gamma(\frac{d}{2})}\right)^d$, where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ (Gamma function).

Theorem 5 highlights that the estimation of our measure is adversely affected by the curse of dimensionality, meaning that as the dimensionality of the data increases, so does the number of samples required for accurate estimation. This poses a computational challenge, particularly when employing gradient-based methods to identify robust counterfactuals based on stability metrics. To mitigate this computational burden, we introduce the Stability-Soft Estimator as a more efficient, albeit less accurate, alternative for estimating stability. To arrive at this estimator, we utilize the Lipschitz property to approximate the aspect that involves the Lipschitz constant, specifically, by approximating $\gamma_x ||x - x_i||$ with $|m(x) - m(x_i)|$.

Remark 3: A reverse statement of Theorem 5 would depend on the particular estimation technique. Estimating the Lipschitz constant is challenging in general, and most estimators tend to underestimate the true Lipschitz constant. This happens because even if there is a small region of the input manifold where the model has erratic behavior, the global Lipschitz constant is high and this can be missed in estimation if there are no samples collected from that small region. Proving a reverse might require additional assumptions, e.g., the Lipschitz constant is further known to be bounded or has limited variation which will be explored in future work.

Definition 8 (Stability-Soft Estimator): Let $N_{x,k}$ be a set of k points drawn from the Gaussian distribution $\mathcal{N}(x, \sigma^2 \mathrm{I}_d)$, the stability variance estimator of a counterfactual $x \in \mathbb{R}^d$ is defined as follows:

$$\hat{R}_{k,\sigma^2}(x, m) = \frac{1}{k}\sum_{x_i \in N_{x,k}} (m(x_i) - |m(x) - m(x_i)|).$$

Properties of Stability Estimators: To gain a deeper understanding of stability, we now consider some desirable properties of counterfactuals from [9], which proposed these properties for tree-based ensembles. The first property is based on the fact that the output of a model $m(x) \in [0, 1]$ is expected to be higher if the model has more confidence in that prediction.

Property 1: For $x \in \mathbb{R}^d$, a higher value of $m(x)$ makes it less likely to be invalidated due to model changes.

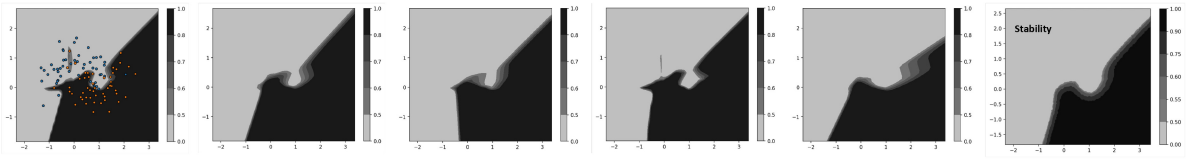
A high $m(x)$ alone does not guarantee robustness, as local variability around x can make predictions less reliable, e.g., points with high $m(x)$ near the decision boundary are more vulnerable to invalidation.

Property 2: An x is less likely to be invalidated if several points close to x (denoted by x') have a high $m(x')$.

Counterfactuals may also be more likely to be invalidated if it lies in a highly variable region of the model output function. This is because the confidence of the model predictions in that region may be less reliable.

Property 3: An x is less likely to be invalidated if model outputs around x have low variability.

We recognize the insights provided by the three axiomatic properties which highlight individual aspects contributing to robustness. We note that robustness cannot be ascribed to any single property in isolation. Rather, it is that the collective integration of these properties—high confidence in predictions (*Property 1*), the reinforcement of confidence through neighborhood consensus (*Property 2*), and low variability in model



(a) m (b) $M = m_1$ (c) $M = m_2$ (d) $M = m_3$ (e) $M = m_4$ (f) $\hat{R}_{k,\sigma^2}(x, m)$

Fig. 4. Effect of stability measure on naturally-occurring model changes: (a) corresponds to the original data distribution and the trained model. (b)-(e) demonstrate some examples of changed models obtained on retraining with different weight initializations. One may notice that the model decision boundary is changing a lot in the sparse regions of the data-manifold (few data-points), possibly violating the bounded-parameter change assumption but the predictions on the dense regions of the data-manifold do not change much (in alignment with Rashomon effect). This motivates our proposed abstraction of naturally-occurring model change which allows for arbitrary changes in the parameter space with little change in the actual predictions on the dense regions of the data manifold. (f) demonstrates our proposed measure of stability $\hat{R}_{k,\sigma^2}(x, m)$ (high mean model output, low variability, *almost* like a Gaussian filter) for which we derive probabilistic guarantees on validity. In essence, we show that under the abstraction of naturally-occurring model change, the stability measure captures the reliable intersecting region of changed models with high probability. In the original model, we observe that certain non-robust regions (i.e., those caused by overfitting to certain data points in the original model) have higher local Lipschitz values and variability. Counterfactuals assigned to these regions (even if $m(x)$ is high) would be invalidated in the changed models. The stability measure, which samples around a region, penalizes these higher local Lipschitz values.

outputs around a point (*Property 3*)—is essential for the robustness of a counterfactual. Our stability measures excel at collectively respecting these properties.

Given a point $x \in \mathbb{R}^d$, it generates a set of k points centered around x. The first term $\frac{1}{k}\sum_{x' \in N_{x,k}} m(x')$ is expected to be high if the model output value $m(x)$ is high for x as well as several points close to x. But the mean value of $m(x')$ around a point x may not always capture the variability in that region, hence, the second term of the stability measure. In the Stability-Lipschitz estimator, the second term $\frac{1}{k}\sum_{x' \in N_{x,k}} \hat{\gamma}_x \|x - x'\|$ captures the *worst-case variability* of the model outputs in the neighborhood of x. The second term of the Stability-Soft estimator, $\frac{1}{k}\sum_{x' \in N_{x,k}} |m(x) - m(x')|$, captures the *average variability* of the model outputs around x. The variability term is only useful in conjunction with the mean term which captures the average confidence in the neighborhood of a given point. This mean term along with the variability term make up our stability measure.

Fig. 4 provides an example on a synthetic dataset to show the effect of our stability measure on naturally changed models realized from actual experiments by retraining with different weight initializations.

IV. Generating Robust Counterfactuals Using Our Proposed Measure: Stability

In this section, we examine several techniques of incorporating our proposed stability measure for generating robust counterfactuals for neural networks. We first define a counterfactual robustness test along the lines of [9].

Definition 9 (Counterfactual Robustness Test): A counterfactual $x \in \mathbb{R}^d$ satisfies the test if: $\hat{R}_{k,\sigma^2}(x, m) \geq \tau$.

A. Closest Robust Counterfactual

Here, we focus on finding a point that satisfies the robustness test, $\hat{R}_{k,\sigma^2}(x', m) \geq \tau$. The threshold value of τ can be adjusted based on the desired effective validity. Hence, a larger threshold would likely ensure that the new model, M, remains valid with high probability.

We propose Algorithm 1, T-Rex:I, which incorporates our measure to find robust counterfactuals on top of any preferred base method for generating counterfactuals. It evaluates the stability of the generated counterfactual and, if necessary, iteratively updates the generated counterfactual through a gradient descent process until a robust counterfactual that meets the desired criteria is obtained. We anticipate that since the robustness measure maximizes the mean value of the model prediction probabilities, it would steer toward the more favorable region. We also check for $m(x_c) \geq 0.5$ as a stopping criterion. An alternative T-Rex variant could start directly from the original instance x, aiming to find a counterfactual that is both close and robust by integrating multiple (differentiable) loss functions including a distance metric and our stability measure.

Remark 4 (Gradient of Stability): In Algorithm 1 and 3, we compute the gradient of $R_{k,\sigma^2}(x, m)$ with respect to x (not model m parameters). Such gradients w.r.t. x instead of m are also computed commonly in adversarial machine learning and also in feature-attributions for explainability. We use TensorFlow `tf.GradientTape` for automatic differentiation, which allows for the computation of gradients with respect to certain inputs.

B. Robust Counterfactuals on Data Support

In certain cases, it may be desirable to generate counterfactuals from a predefined set of data points (i.e., training dataset $\mathcal{S}$. This is to remove the risk of producing unrealistic or anomalous results. In this context, we define the Robust Data Support Counterfactual.

Definition 10 (Robust Data Support Counterfactual): Given $x \in \mathbb{R}^d$ such that $m(x) < 0.5$, its robust nearest neighbor counterfactual $\mathcal{C}^{(\tau)}_{p,\mathcal{S}}(x, m)$ with respect to the model $m(\cdot)$ and dataset $\mathcal{S}$ is defined as another point $x' \in \mathcal{S}$ that minimizes the l_p norm $\|x - x'\|_p$ such that $m(x') \geq 0.5$ and $\hat{R}_{k,\sigma^2}(x', m) \geq \tau$.

The closest data-supported counterfactual serves as a reliable reference, as it inherently has a high Local Outlier Factor (LOF). We propose Algorithm 2, T-Rex:NN, for finding data-supported counterfactuals. The algorithm begins by locating the K nearest neighbor counterfactuals to a given point x within the dataset $\mathcal{S}$. It then iterates through each of these candidates, evaluating them against a robustness test, $\hat{R}_{k,\sigma^2}(x', m) \geq \tau$. If

Algorithm 1 T-Rex:I: Theoretically Robust EXplanations: Iterative Version

Input: Model $m(\cdot)$, Datapoint x with $m(x) < 0.5$,
Algorithm parameters $(k, \sigma^2, \eta, \tau,$ max_steps).
Generate initial counterfactual x' using any technique.
Initialize robust counterfactual $x_c = x'$ and steps $= 0$.
while $\hat{R}_{k,\sigma^2}(x_c, m) < \tau$ and steps $<$ max_steps **do**
  Compute $\hat{R}_{k,\sigma^2}(x_c, m)$
  Compute gradient $\nabla_{x_c}\hat{R}_{k,\sigma^2}(x_c, m)$
  Update x_c via gradient ascent:
    $x_c = x_c + \eta\nabla_{x_c}\hat{R}_{k,\sigma^2}(x_c, m)$
  Increment steps
end while
Output x_c and exit

Algorithm 2 T-Rex:NN: Theoretically Robust EXplanations: Nearest Neighbor Version

Input: Model $m(\cdot)$, Datapoint x with $m(x)<0.5$, Dataset $\mathcal{S}$,
Algorithm parameters (K, σ^2, k, τ).
Let $\text{NN}_x = (x'_1, x'_2, \ldots, x'_K)$ be the K nearest neighbors to x with $m(x') \geq 0.5$,
for $x'_i \in \text{NN}_x$ **do**
  Perform counterfactual robustness test on x'_i:
    Check if $\hat{R}_{k,\sigma^2}(x'_i, m) \geq \tau$
  if counterfactual robustness test is satisfied: **then**
    Output x'_i and exit
  end if
end for
Output: *No robust counterfactual found* and exit

a counterfactual meets this criterion, it is considered robust and the algorithm terminates.

C. Robust Counterfactuals on Data Manifold

Here, we focus on finding robust counterfactuals on the data manifold $\mathcal{X} \subseteq \mathbb{R}^d$ (realistic samples; see Definition 3). We leverage generative models to learn a lower dimensional latent representation of the data manifold in $\mathbb{R}^l$ where $l < d$. We focus on the Variational Auto-Encoders (VAEs) [39]. We designate the encoder component of the VAE, parameterized by θ, as $F_\theta : \mathbb{R}^d \to \mathbb{R}^l$ which transforms any data point $x \in \mathcal{X}$ into its corresponding latent variable $z \in \mathbb{R}^l$. The decoder denoted as $G_\phi : \mathbb{R}^l \to \mathbb{R}^d$, parameterized by ϕ, maps the latent variable back to the original data space.

Our method uses this latent space learned by a VAE for robust counterfactual search. Given that the latent space captures the data manifold, searching for counterfactuals in this representation enables us to discover instances coherent with the intrinsic data distribution and hence more plausible (higher LOF). The objective is as follows:

$$z' = \arg\min_z \ell(m(G_\phi(z)), 1) + \lambda_1 \|x - G_\phi(z)\|_p - \lambda_2 \hat{R}_{k,\sigma^2}(G_\phi(z), m). \quad (9)$$

Here $\ell(\cdot, \cdot)$ denotes a differentiable loss function (e.g. mean square loss, $\ell(u, v) = (u - v)^2$ or binary cross-entropy, loss $\ell(u, v) = -[v \log(u) + (1 - v)\ \log(1 - u)]$) that minimizes the gap between the prediction and the favorable outcome of 1, and the counterfactual returned is $G_\phi(z')$. The counterfactual lies in the data manifold since our algorithm obtains the latent encoding of our sample x using the encoder $z = F_\phi(x)$. The gradient steps are in the latent space of the encoder to minimize our overall loss function until we reach a z with robustness threshold $R_{k,\sigma^2}(G_\phi(z), m) > \tau$ on the desired side of the decision boundary. The details are in Algorithm 3: T-Rex: Hybrid.

Algorithm 3 T-Rex: Hybrid: Theoretically Robust EXplanations: Hybrid Version

Input: Model $m(\cdot)$, Dataset $\mathcal{S}$, Datapoint x with $m(x) < 0.5$,
Algorithm parameters $(k, \sigma^2, \eta, \tau, \lambda_1, \lambda_2,$ max_steps)
Train VAE encoder $F_\theta(\cdot)$ and decoder $G_\phi(\cdot)$ with dataset $\mathcal{S}$
Initialize $z = F_\theta(x)$
while steps $<$ max_steps **do**
  $z \leftarrow z - \eta\nabla_z(\ell(m(G_\phi(z)), 1) + \lambda_1\|x - G_\phi(z)\| - \lambda_2\hat{R}_{k,\sigma^2}(G_\phi(z), m))$
  steps $\leftarrow$ steps $+1$
  if $m(G_\phi(z)) > 0.5$ and $\hat{R}_{k,\sigma^2}(G_\phi(z), m) > \tau$ **then**
    Return $x' = G_\phi(z)$ and exit
  end if
end while
Return *No robust counterfactual found* and exit

V. Experiments

Here, we present experimental results to demonstrate how our proposed Algorithm 1 & 2 utilizes our stability measure to generate robust counterfactuals effectively.

1) Datasets: We conduct experiments on several benchmark datasets, namely, HELOC [16], German Credit, Cardiotocography (CTG), Adult [17], and Taiwanese Credit [18]. These have two classes, with one class representing the most favorable outcome, and the other representing the least desirable outcome for which we aim to generate counterfactuals. For simplicity, we normalize the features to lie between $[0, 1]$.

2) Performance Metrics: Our metrics of interest are:

- Cost: Average l_1 or l_2 distance between counterfactuals x' and original points x.
- Validity (%): Percentage of counterfactuals that remain valid under the new model M.
- LOF: Predicts -1 for anomalous points, and $+1$ for inliers. A high average LOF essentially suggests the points lie on the data manifold and hence more realistic, i.e., *higher is better* (see Definition 4). We use an existing implementation to compute LOF from [40].

3) Methodology: We begin by training a baseline neural network model and find counterfactuals for data points with true negative predictions. To test the robustness of these counterfactual examples, we then train 50 new models (M) and evaluate the *validity* of the counterfactuals under different model change scenarios, which include: (i) Weight Initialization (WI): Retraining new models using the same

TABLE I
EXPERIMENTAL RESULTS

		l_1 based				l_2 based			
	Method	COST	LOF	WI VAL.	LO VAL.	COST	LOF	WI VAL.	LO VAL.
HELOC	min Cost	**0.40**	**0.49**	38.8%	35.2%	**0.11**	**0.75**	13.5%	13.5%
	min Cost+T-Rex:I (Ours)	1.02	0.38	**98.2**%	**98.1**%	0.29	0.68	**98.5**%	**98.2**%
	min Cost+SNS	1.20	0.30	98.0%	97.8%	0.31	0.64	97.9%	97.0%
	ROAR	1.69	0.41	92.6%	91.2%	1.91	0.43	86.3 %	84.8%
	NN	**1.91**	0.80	51.1%	50.3%	**0.56**	0.80	51.1%	50.3%
	T-Rex:NN (Ours)	2.50	**0.92**	**84.0**%	**84.0**%	0.77	**0.92**	**84.0**%	**84.0**%
GERMAN	min Cost	**1.42**	**0.77**	58.8%	56.7%	**0.48**	**0.81**	26.6%	26.6%
	min Cost+T-Rex:I (Ours)	4.81	0.72	**98.0**%	96.5%	1.20	0.75	99.2%	98.7%
	min Cost+SNS	5.71	0.67	97.5%	**98.1**%	1.44	0.68	**99.9**%	**98.9**%
	ROAR	7.63	0.54	96.3%	92.3%	6.81	0.58	87.8%	85.2%
	NN	**7.05**	**1.00**	95.3%	95.4%	**2.50**	**1.00**	95.3%	95.3%
	T-Rex:NN (Ours)	10.13	**1.00**	**100**%	**100**%	3.04	**1.00**	**100**%	**100**%
CTG	min Cost	**0.21**	**0.94**	74.6%	70.2%	**0.08**	**1.00**	19.7%	14.1%
	min Cost+T-Rex:I (Ours)	1.11	0.83	**100**%	**98.8**%	0.42	0.94	**100**%	**99.7**%
	min Cost+SNS	3.34	-1.00	**100**%	98.2%	1.07	-1.00	**100**%	99.3%
	ROAR	3.68	0.64	98.7%	96.4%	1.35	0.59	98.9%	97.2%
	NN	**0.39**	**1.00**	70.5%	67.5%	**0.15**	**1.00**	70.5%	67.5%
	T-Rex:NN (Ours)	2.22	-0.33	**100**%	**100**%	1.00	-0.33	**100**%	**100**%

hyperparameters but with different weight initialization by using different random seeds for each new model; and (ii) Leave Out (LO): Retraining new models by randomly removing a small portion (1%) of the training data each time (with replacement) as well as different weight initialization. This can be justified by the concept of machine unlearning, especially within the context of regulatory frameworks like the GDPR [13]. The "right to be forgotten" mandates the deletion of an individual's data upon request, which may necessitate updates to the model. These updates can affect the reliability of previously provided explanations, thereby posing a challenge to reconciling the "right to explanation" with the "right to be forgotten" [7].

4) Hyperparameter Selection: Our findings indicate that higher k improves robustness, but comes at the cost of increased computational cost. Our choice of k $= 1000$ also aligns with practices in the adversarial robustness literature, where similar trade-offs between performance and computational feasibility are considered. The value of σ^2 was determined by analyzing the variance of the features. In the dataset with the features between [0, 1], we found that a value of $\sigma^2 = 0.01$ produced good results. The threshold τ is a critical aspect of our method and can be adjusted based on the desired effective validity. A higher τ value improves validity at the expense of l_1 or l_2 cost. See Appendix F for more details.

5) Baseline: We compare our approaches with established baselines. First, we find the min Cost (l_1 and l_2) counterfactual [2] and use it as our base method for generating counterfactuals. We then compare T-Rex:I to the Stable Neighbor Search (SNS) [6] and Robust Algorithmic Recourse (ROAR) [5]. We evaluate the performance of our Robust Nearest Neighbor (Algorithm 2:T-Rex:NN) against the Nearest Neighbor (NN) counterfactuals (closest data-support robust counterfactual in Definition 10). We choose a value of τ to get high validity and compare cost and LOF with baselines.

6) Results: Results for HELOC, German Credit, and CTG datasets are in Table I. Observe that the min Cost counterfactual is not robust to variations in the training data or weight initialization as expected. ROAR generates counterfactuals with high validity, albeit at the expense of a higher cost. Our proposed method, T-Rex:I, significantly improves the validity of the counterfactuals compared to the minimum cost. The T-Rex:I algorithm achieves comparable validity results to the SNS method for both types of model changes, and often accomplishes this with lower costs and higher LOF. This can be observed across all three datasets for both l_1 and l_2 cost metrics. The T-Rex:NN algorithm also significantly improves the validity of the counterfactuals compared to the traditional Nearest Neighbor (NN) method and maintains a high LOF, except for the CTG dataset with a low LOF score. This appears to be an exception rather than the norm, possibly due to the specific characteristics of the CTG dataset itself. T-Rex:NN shows competitive performance on other datasets, such as the Taiwanese Credit and Adult datasets (see Appendix F). It comes at a price of increased cost, but the counterfactuals are guaranteed to be realistic since they are data-supported. We observe a lower LOF score on the CTG dataset. Refer to Appendix F for additional results for Adult and Taiwanese credit datasets.

7) Ablation: To evaluate the efficacy of our proposed stability measure, we conduct an ablation study on the German credit dataset. We first evaluate a robustness measure that solely relies on the model's prediction of the counterfactual, denoted as $r(x', m) = m(x')$. We then examine a measure that only incorporates the mean, the average predictions for k points sampled from the distribution $N(x', \sigma^2 I_d)$, denoted as

$r_{k,\sigma^2}(x', m) = \frac{1}{k}\sum_{x'_i \in N_{x',k}} m(x'_i)$. We compare these with our proposed robustness measure $\hat{R}_{k,\sigma^2}(x', m)$, which also takes into account the variability around the counterfactual. The results of the ablation study, for various τ thresholds, are summarized in Table VI in Appendix F.

VI. Discussion

We introduce an abstraction called naturally-occurring model change and propose a measure, Stability, to quantify the robustness of counterfactuals with probabilistic guarantees. We show that counterfactuals with high Stability will remain valid after potential model changes with high probability. We investigate various techniques for incorporating stability in generating robust counterfactuals and introduce the T-Rex:I, T-Rex:NN, and T-Rex:Hybrid algorithms. We also make a novel conceptual connection with the body of work on model multiplicity, further emphasizing on the models that are more likely to occur.

The naturally-occurring model changes rest on assumptions that may not apply to all models or datasets. Our stability estimators, although practically implementable, lack the same theoretical guarantees as the initial stability measure. Estimating the Lipschitz constant around a counterfactual can be computationally demanding, particularly when leveraging gradient descent to optimize stability. Though generating robust counterfactuals is a key step towards trustworthy AI, it can fall short of other important factors such as fairness [41], [42], [43], [44], [45]. Future research could explore links between robustness and fairness, improving the estimation of stability, or integrating Stability into training-time-based approaches for generating robust counterfactuals. Our current framework assumes a continuous space, but exploring extensions to discrete feature spaces would also be interesting.

Disclaimer: This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy, or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust counterfactual explanations for neural networks with probabilistic guarantees," in *Int. Conf. Mach. Learn.*, ser. PMLR, vol. 202, Jul. 2023, pp. 12351–12367.

[2] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL Tech.*, vol. 31, p. 841, 2017.

[3] A. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects," *CoRR*, vol. abs/2010.04050, 2020.

[4] S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in *Proc. 2020 Conf. Fairness, Accountability, Transparency*, 2020, pp. 80–89.

[5] S. Upadhyay, S. Joshi, and H. Lakkaraju, "Towards robust and reliable algorithmic recourse," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[6] E. Black, Z. Wang, M. Fredrikson, and A. Datta, "Consistent counterfactuals for deep models," *arXiv preprint arXiv:2110.03109*, 2021.

[7] S. Krishna, J. Ma, and H. Lakkaraju, "Towards bridging the gaps between the right to explanation and the right to be forgotten," *arXiv preprint arXiv:2302.04288*, 2023.

[8] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proc. of Web Conf. 2020*, 2020, pp. 3126–3132.

[9] S. Dutta, J. Long, S. Mishra, C. Tilli, and D. Magazzeni, "Robust counterfactual explanations for tree-based ensembles," in *Int. Conf. Mach. Learn.*. PMLR, 2022, pp. 5742–5756.

[10] J. Jiang, F. Leofante, A. Rago, and F. Toni, "Formalising the robustness of counterfactual explanations for neural networks," *arXiv preprint arXiv:2208.14878*, 2022.

[11] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[12] J. Jiang, F. Leofante, A. Rago, and F. Toni, "Robust counterfactual explanations in machine learning: A survey," *arXiv preprint arXiv:2402.01928*, 2024.

[13] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer Int. Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[14] L. Breiman, "Statistical modeling: The two cultures," *Qual. Eng.*, vol. 48, pp. 81–82, 2001.

[15] C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Int. Conf. Mach. Learn.*. PMLR, 2020, pp. 6765–6774.

[16] FICO, "FICO XML challenge," [Online]. Available: https://community.fico.com/s/explainable-machine-learning-challenge, 2018.

[17] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[18] I.-C. Yeh and C. hui Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Syst. Appl.*, vol. 36, no. 2, Part 1, pp. 2473–2480, 2009.

[19] M. Pawelczyk, K. Broelemann, and G. Kasneci, "On counterfactual explanations under predictive multiplicity," in *Conf. Uncertainty Artif. Intell.*. PMLR, 2020, pp. 809–818.

[20] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, "DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization." in *IJCAI*, 2020, pp. 2855–2862.

[21] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and actionable counterfactual explanations," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2020, pp. 344–350.

[22] L. Hancox-Li, "Robustness in machine learning explanations: Does it matter?" in *Proc. 3rd ACM Conf. Fairness, Accountability, Transparency (FAT*)*. Barcelona, Spain, Jan. Feb.Jul.–Mar.0 2020, pp. 640–647.

[23] K. Rawal, E. Kamar, and H. Lakkaraju, "Can I still trust you?: Understanding the impact of distribution shifts on algorithmic recourses," *arXiv preprint arXiv:2012.11788*, 2020.

[24] J. Jiang, J. Lan, F. Leofante, A. Rago, and F. Toni, "Provably robust and plausible counterfactual explanations for neural networks via robust Optimisation," *arXiv preprint arXiv:2309.12545*, 2023.

[25] S. Mishra, S. Dutta, J. Long, and D. Magazzeni, "A survey on the robustness of feature importance and counterfactual explanations," *arXiv e-prints*, vol. arXiv:2111.00358, 2021.

[26] T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "Issues with post-hoc counterfactual explanations: A discussion," *arXiv preprint arXiv:1906.04774*, 2019.

[27] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[28] M. Pawelczyk, T. Datta, J. van-den Heuvel, G. Kasneci, and H. Lakkaraju, "Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse," *arXiv preprint arXiv:2203.06768*, 2022.

[29] D. Maragno, J. Kurtz, T. E. Röber, R. Goedhart, Ş. I. Birbil, and D. d. Hertog, "Finding regions of counterfactual explanations via robust optimization," *arXiv preprint arXiv:2301.11113*, 2023.

[30] R. Dominguez-Olmedo, A. H. Karimi, and B. Schölkopf, "On the adversarial robustness of causal algorithmic recourse," in *Int. Conf. Mach. Learn.*. PMLR, 2022, pp. 5324–5342.

[31] E. Black, M. Raghavan, and S. Barocas, "Model multiplicity: Opportunities, concerns, and solutions," in *Proc. 2022 ACM Conf. Fairness, Accountability, Transparency*, ser. FAccT '22, 2022, pp. 850–863.

[32] H. Hsu and F. Calmon, "Rashomon capacity: A metric for predictive multiplicity in classification," in *Adv. Neural Inf. Process. Syst.*, vol. 35. Curran Associates, Inc., 2022, pp. 28988–29000.

[33] J. Watson-Daniels, D. C. Parkes, and B. Ustun, "Predictive multiplicity in probabilistic classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 9, 2023, pp. 10306–10314.

[34] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, "Counterfactual Shapley additive explanations," *ACM Conf. Fairness, Accountability, Transparency*, 2022.

[35] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. 2000 ACM SIGMOD Int. Conf. Manage. of data*, 2000, pp. 93–104.

[36] H.-T. D. Liu, F. Williams, A. Jacobson, S. Fidler, and O. Litany, "Learning smooth neural functions via lipschitz regularization," in *ACM SIGGRAPH 2022 Conf. Proc.*, 2022, pp. 1–13.

[37] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive approximation*, vol. 28, pp. 253–263, 2008.

[38] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.

[39] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[40] Scikit-Learn, "LOF implementation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

[41] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models," *arXiv preprint arXiv:1905.07857*, 2019.

[42] V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian, "Equalizing recourse across groups," *arXiv preprint arXiv:1909.03166*, 2019.

[43] D. Ley, S. Mishra, and D. Magazzeni, "Global counterfactual explanations: Investigations, implementations and improvements," 2022.

[44] N. Raman, D. Magazzeni, and S. Shah, "Bayesian hierarchical models for counterfactual estimation," in *Int. Conf. Artif. Intell. Statist.*. PMLR, 2023, pp. 1115–1128.

[45] A.-R. Ehyaei, A.-H. Karimi, B. Schölkopf, and S. Maghsudi, "Robustness implies fairness in casual algorithmic recourse," *arXiv preprint arXiv:2302.03465*, 2023.