Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Anonymous submission

Abstract

We propose a novel training regime termed counterfactual training that leverages counterfactual explanations to increase the explanatory capacity of models. Counterfactual explanations have emerged as a popular post-hoc explanation method for opaque machine learning models: they inform how factual inputs would need to change in order for a model to produce some desired output. To be useful in real-word decision-making systems, counterfactuals should be plausible with respect to the underlying data and actionable with respect to the feature mutability constraints. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. In this work, we instead hold models directly accountable for the desired end goal: counterfactual training employs counterfactuals during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable counterfactual explanations and exhibit greatly improved adversarial robustness.

1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by the success of representation learning with high degrees of freedom: instead of relying on features and rules hand-crafted by humans, modern machine learning (ML) models are tasked with learning highly complex representations directly from the data, guided by narrow objectives such as predictive accuracy (Goodfellow, Bengio, and Courville 2016). These models tend to be so complex that humans cannot easily interpret their decision logic.

Counterfactual explanations (CE) have become a key part of the broader explainable AI (XAI) toolkit (Molnar 2022) that can be applied to make sense of this complexity. Originally proposed by Wachter, Mittelstadt, and Russell (2017), CEs prescribe minimal changes for factual inputs that, if implemented, would prompt some fitted model to produce an alternative, more desirable output. This is useful and necessary to not only understand how opaque models make their predictions, but also to provide algorithmic recourse (AR) to individuals subjected to them: a retail bank, for example, could use CE to provide meaningful feedback to unsuccessful loan applicants that were rejected based on an opaque automated decision-making (ADM) system (Figure 1).

For such feedback to be meaningful, counterfactual explanations need to fulfill certain desiderata (Verma et al. 2022; Karimi et al. 2021)—they should be faithful to the model (Altmeyer et al. 2024), plausible (Joshi et al. 2019) and actionable (Ustun, Spangher, and Liu 2019). Plausibility is typically understood as counterfactuals being *in-domain*: unsuccessful loan applicants that implement the provided recourse should end up with credit profiles that are genuinely similar to that of individuals who have successfully repaid their loans in the past. Actionable explanations comply with practical constraints: a young, unsuccessful loan applicant cannot increase their age in an instance.

Existing state-of-the-art (SOTA) approaches in the field have largely focused on designing model-agnostic CE methods that identify subsets of counterfactuals, which comply with specific desiderata. This is problematic, because the narrow focus on any specific desideratum can adversely affect others: it is possible, for example, to generate plausible counterfactuals for models that are also highly vulnerable to implausible, possibly adversarial counterfactuals (Altmeyer et al. 2024). In this work, we therefore embrace the paradigm that models (as opposed to explanation methods) should be held accountable for explanations that are plausible and actionable. While previous work has shown that at least plausibility can be indirectly achieved through existing techniques aimed at models' generative capacity, generalization and robustness (Altmeyer et al. 2024; Augustin, Meinke, and Hein 2020; Schut et al. 2021), we directly incorporate both plausibility and actionability in the training objective of models to improve their overall explanatory capacity.

Specifically, we propose **counterfactual training (CT)**: a novel training regime that leverages counterfactual explanations on-the-fly to ensure that differentiable models learn plausible and actionable explanations for the underlying data, while at the same time also being more robust to adversarial examples (AE). Figure 1 illustrates the outcomes of CT compared to a conventionally trained model. First, in panel (a), faithful and valid counterfactuals end up near the decision boundary forming a clearly distinguishable cluster in the target class (orange). In panel (b), CT is applied to the same underlying linear classifier architecture resulting in much more plausible counterfactuals. In panel (c), the classifier is again trained conventionally and we have introduced a mutability constraint on the *age* feature at test

time—counterfactuals are valid but the classifier is roughly equally sensitive to both features. By contrast, the decision boundary in panel (d) has titled, making the model trained with CT relatively less sensitive to the immutable *age* feature. To achieve these outcomes, CT draws inspiration from the literature on contrastive and robust learning: we contrast faithful CEs with ground-truth data while protecting immutable features, and capitalize on methodological links between CE and AE by penalizing the model's adversarial loss on interim (*nascent*) counterfactuals. To the best of our knowledge, CT represents the first venture in this direction with promising empirical and theoretical results.

The remainder of this manuscript is structured as follows. Section 2 presents related work, focusing on the links to contrastive and robust learning. Then follow our two principal contributions. In Section 3, we introduce our methodological framework and show theoretically that it can be employed to respect global actionability constraints. In our experiments (Section 4), we find that thanks to counterfactual training, (1) the implausibility of CEs decreases by up to 90%; (2) the cost of reaching valid counterfactuals with protected features decreases by 19% on average; and (3) models' adversarial robustness improves across the board. Finally, we discuss open challenges in Section 5 and conclude in Section 6.

2 Related Literature

To make the desiderata for our framework more concrete, we follow previous work in tying the explanatory capacity of models to the quality of CEs that can be generated for them (Altmeyer et al. 2024; Augustin, Meinke, and Hein 2020). For simplicity, we refer to "explanatory capacity" as "explainability" in the rest of this manuscript (see Def. 3.1).

2.1 Explainability and Contrastive Learning

In a closely related work, Altmeyer et al. (2024) show that model averaging and, in particular, contrastive model objectives can produce more explainable and hence trustworthy models. The authors propose a way to generate counterfactuals that are maximally faithful in that they are consistent with what models have learned about the underlying data. Formally, they rely on tools from energy-based modelling (Teh et al. 2003) to minimize the contrastive divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by a model. Their algorithm, *ECCCo*, yields plausible counterfactual explanations if and only if the underlying model has learned representations that align with them. The authors find that both deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and joint energy-based models (JEMs) (Grathwohl et al. 2020), a form of constrastive learning, tend to do well in this regard.

It helps to look at these findings through the lens of representation learning with high degrees of freedom. Deep ensembles are approximate Bayesian model averages, which are particularly effective when models are underspecified by the available data (Wilson 2020). Averaging across solutions mitigates the risk of overrelying on a single locally optimal representation that corresponds to semantically meaningless explanations. Likewise, previous work of Schut et al. (2021) found that generating plausible ("interpretable") CEs is almost trivial for deep ensembles that have undergone adversarial training. The case for JEMs is even clearer: they optimize a hybrid objective that induces both high predictive performance and strong generative capacity (Grathwohl et al. 2020), which resembles the idea of aligning models with plausible explanations and has inspired CT.

2.2 Explainability and Robust Learning

Augustin, Meinke, and Hein (2020) show that CEs tend to be more meaningful ("explainable") if the underlying model is more robust to adversarial examples. Once again, we can make intuitive sense of this finding if we look at adversarial training (AT) through the lens of representation learning with high degrees of freedom: highly complex and flexible models may learn representations that make them sensitive to implausible or even adversarial examples (Szegedy et al. 2014). Thus, by inducing models to "unlearn" susceptibility to such examples, adversarial training can effectively remove implausible explanations from the solution space.

This interpretation of the link between explainability through counterfactuals on the one side, and robustness to adversarial examples on the other is backed by empirical evidence. Sauer and Geiger (2021) demonstrate that using counterfactual images during classifier training improves model robustness. Similarly, Abbasnejad et al. (2020) argue that counterfactuals represent potentially useful training data in machine learning, especially in supervised settings where inputs may be reasonably mapped to multiple outputs. They, too, show that augmenting the training data of (image) classifiers can improve generalization performance. Finally, Teney, Abbasnedjad, and van den Hengel (2020) argue that counterfactual pairs tend to exist in training data. Hence, their approach aims to identify similar input samples with different annotations and ensure that the gradient of the classifier aligns with the vector between such pairs of counterfactual inputs using a cosine distance loss function.

CEs have also been used to improve models in the natural language processing domain. For example, Wu et al. (2021) propose *Polyjuice*, a general-purpose CE generator for language models and demonstrate that the augmentation of training data with *Polyjuice* improves robustness in a number of tasks, while Luu and Inoue (2023) introduce the *Counterfactual Adversarial Training* (CAT) framework that aims to improve generalization and robustness of language models by generating counterfactuals for training samples that are subject to high predictive uncertainty.

There have also been several attempts at formalizing the relationship between counterfactual explanations and adversarial examples. Pointing to clear similarities in how CEs and AEs are generated, Freiesleben (2022) makes the case for jointly studying the opaqueness and robustness problems in representation learning. Formally, AEs can be seen as the subset of CEs for which misclassification is achieved (Freiesleben 2022). Similarly, Pawelczyk et al. (2022) show that CEs and AEs are equivalent under certain conditions.

Two other works are closely related to ours in that they use counterfactuals during training with the explicit goal of affecting certain properties of the post-hoc counterfactual

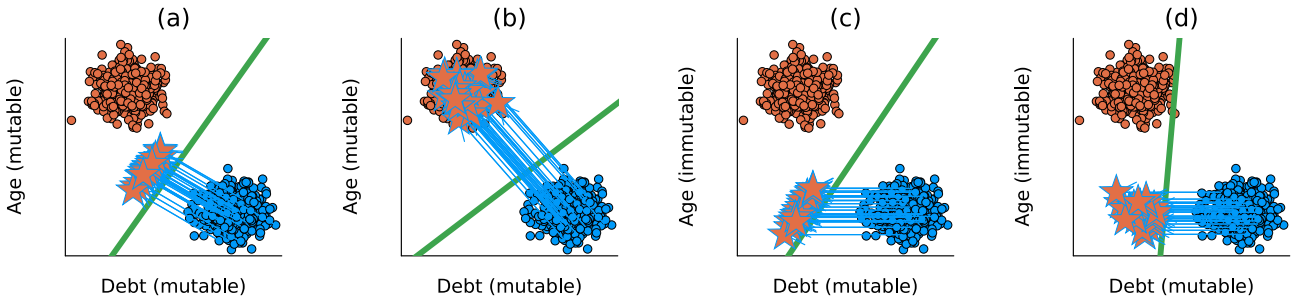

Figure 1: Counterfactual explanations (stars) for linear classifiers trained under different regimes on synthetic data: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable. The linear decision boundary is shown in green along with training data colored according to ground-truth labels: y^- = "loan withheld" (blue) and y^+ = "loan provided" (orange). Class and feature annotations (*debt* and *age*) are for illustrative purposes.

explanations. Firstly, Ross, Lakkaraju, and Bastani (2024) propose a way to train models that guarantee recourse to a positive target class with high probability. Their approach builds on adversarial training by explicitly inducing susceptibility to targeted AEs for the positive class. Additionally, the method allows for imposing a set of actionability constraints ex-ante. For example, users can specify that certain features are immutable. Secondly, Guo, Nguyen, and Yadav (2023) are the first to propose an end-to-end training pipeline that includes CEs as part of the training procedure. Their *CounterNet* network architecture includes a predictor and a CE generator, where the parameters of the CE generator are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor. In contrast, we impose no restrictions on the ANN architecture at all.

3 Counterfactual Training

This section introduces the counterfactual training framework, applying ideas from contrastive and robust learning to counterfactual explanations. CT produces models whose learned representations align with plausible explanations that comply with user-defined actionability constraints.

Counterfactual explanations are typically generated by solving variations of the following optimization problem,

$$\min_{\mathbf{x}' \in \mathcal{X}^D} \left\{ \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'), \mathbf{y}^+) + \lambda \text{reg}(\mathbf{x}') \right\} \tag{1}$$

where $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denotes a classifier, $\mathbf{x}'$ denotes the counterfactual with D features and $\mathbf{y}^+ \in \mathcal{Y}$ denotes some target class. The yloss$(\cdot)$ function quantifies the discrepancy between current model predictions for $\mathbf{x}'$ and the target class (a conventional choice is cross-entropy). Finally, we use reg$(\cdot)$ to denote any form of regularization used to induce certain properties on the counterfactual. In their seminal paper, Wachter, Mittelstadt, and Russell (2017) propose regularizing the distance between counterfactuals and their original factual values to ensure that individuals seeking recourse through CE face minimal costs in terms of feature changes. Different variations of Equation 1 have been proposed in the literature to address many desiderata including the ones discussed above (faithfulness, plausibility and actionability). Like Wachter, Mittelstadt, and Russell (2017), most of these approaches rely on gradient descent to optimize Equation 1. For more details on the approaches tested in this work, we refer the reader to the supplementary appendix. In the following, we describe in detail how counterfactuals are generated and used in counterfactual training.

3.1 Proposed Training Objective

The goal of CT is to improve model explainability by aligning models with faithful explanations that are plausible and actionable. Formally, we define explainability as follows:

Definition 3.1 (Model Explainability)**.** Let $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denote a supervised classification model that maps from the D-dimensional input space $\mathcal{X}$ to representations $\phi(\mathbf{x}; \theta)$ and finally to the K-dimensional output space $\mathcal{Y}$. Assume that for any given input-output pair $\{\mathbf{x}, \mathbf{y}\}_i$ there exists a counterfactual $\mathbf{x}' = \mathbf{x} + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$, where $\arg\max_y \mathbf{y}^+ = y^+$ is the index of the target class.

We say that $\mathbf{M}_\theta$ has an **explanatory capacity** to the extent that faithfully generated, valid counterfactuals are also plausible and actionable. We define these properties as:

- (Faithfulness) $\int^A p_\theta(\mathbf{x}'|\mathbf{y}^+) d\mathbf{x} \to 1$; A is an arbitrarily small region around $\mathbf{x}'$.
- (Plausibility) $\int^A p(\mathbf{x}'|\mathbf{y}^+) d\mathbf{x} \to 1$; A as specified above.
- (Actionability) Perturbations Δ may be subject to some actionability constraints.

Here, $p_\theta(\mathbf{x}|\mathbf{y}^+)$ denotes the conditional posterior distribution over inputs. For simplicity, we refer to a model with high explanatory capacity as **explainable** in this manuscript.

The characterization of faithfulness and plausibility in Def. 3.1 follows Altmeyer et al. (2024), with adapted notation. Intuitively, plausible counterfactuals are consistent with the data and faithful counterfactuals are consistent with what the model has learned about the input data. Ac tionability constraints in Def. 3.1 vary and depend on the context in which $\mathbf{M}_\theta$ is deployed. In this work, we choose to only consider domain and mutability constraints for individual features x_d for $d = 1, ..., D$. We also limit ourselves to classification tasks for reasons discussed in Section 5.

Let $\mathbf{x}'_t$ for $t = 0, ..., T$ denote a counterfactual generated through gradient descent over T iterations as originally

proposed by Wachter, Mittelstadt, and Russell (2017). CT adopts gradient-based CE search in training to generate on-the-fly model explanations $\mathbf{x}'$ for the training samples. We use the term *nascent* to denote interim counterfactuals $\mathbf{x}'_{t\leq T}$ that have not yet converged. As we explain below, these nascent counterfactuals can be stored and repurposed as adversarial examples. Conversely, we consider counterfactuals $\mathbf{x}'_T$ as *mature* explanations if they have either exhausted all T iterations or converged by reaching a pre-specified threshold, τ, for the predicted probability of the target class: $\mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau$, where $\mathcal{S}$ is the softmax function.

Formally, we propose the following counterfactual training objective to train explainable (as in Def. 3.1) models,

$$\begin{aligned} &\min_\theta \text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}}\text{div}(\mathbf{x}^+, \mathbf{x}'_T, y; \theta) \\ &+\lambda_{\text{adv}}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t\leq T}), \mathbf{y}) + \lambda_{\text{reg}}\text{ridge}(\mathbf{x}^+, \mathbf{x}'_T, y; \theta) \end{aligned} \tag{2}$$

where $\text{yloss}(\cdot)$ is any classification loss that induces discriminative performance (e.g., cross-entropy). The second and third terms are explained in detail below. For now, they can be summarized as inducing explainability directly and indirectly by penalizing (1) the contrastive divergence, $\text{div}(\cdot)$, between mature counterfactuals $\mathbf{x}'_T$ and observed samples $\mathbf{x}^+ \in \mathcal{X}^+ = \{\mathbf{x} : y = y^+\}$ in the target class y^+, and (2) the adversarial loss, $\text{advloss}(.)$, wrt. nascent counterfactuals $\mathbf{x}'_{t\leq T}$. Finally, $\text{ridge}(\cdot)$ denotes a Ridge penalty (ℓ_2-norm) that regularizes the magnitude of the energy terms involved in $\text{div}(\cdot)$ (Du and Mordatch 2020). The trade-offs between these components are adjusted through λ_{div}, λ_{adv} and λ_{reg}. The full training regime is sketched out in Algorithm 1.

Algorithm 1: Counterfactual Training

Require: Training dataset $\mathcal{D}$, initialize model $\mathbf{M}_\theta$
1: **while** not converged **do**
2: Sample $\mathbf{x}$ and $\mathbf{y}$ from dataset $\mathcal{D}$.
3: Sample $\mathbf{x}'_0$, $\mathbf{y}^+$ and $\mathbf{x}^+$.
4: **for** $t = 1$ to T **do**
5: Backpropagate $\nabla_{\mathbf{x}'}$ through Equation 1. Store $\mathbf{x}'_t$.
6: **end for**
7: Backpropagate ∇_θ through Equation 2.
8: **end while**
9: **return** $\mathbf{M}_\theta$

3.2 Directly Inducing Explainability with Contrastive Divergence

Grathwohl et al. (2020) observe that any classifier can be re-interpreted as a joint energy-based model that learns to discriminate output classes conditional on the observed (training) samples from $p(\mathbf{x})$ and the generated samples from $p_\theta(\mathbf{x})$. The authors show that JEMs can be trained to perform well at both tasks by directly maximizing the joint log-likelihood: $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$, where the first term can be optimized using cross-entropy as in Equation 2. To optimize $\log p_\theta(\mathbf{x})$, they minimize the contrastive divergence between the observed samples from $p(\mathbf{x})$ and samples generated from $p_\theta(\mathbf{x})$.

To generate samples, Grathwohl et al. (2020) use Stochastic Gradient Langevin Dynamics (SGLD) with an uninformative prior for initialization but we depart from their methodology: we propose to leverage counterfactual explainers to generate counterfactuals of observed training samples. Specifically, we have:

$$\text{div}(\mathbf{x}^+, \mathbf{x}'_T, y; \theta) = \mathcal{E}_\theta(\mathbf{x}^+, y) - \mathcal{E}_\theta(\mathbf{x}'_T, y) \tag{3}$$

where $\mathcal{E}_\theta(\cdot)$ denotes the energy function defined as $\mathcal{E}_\theta(\mathbf{x}, y) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$, with y^+ denoting the index of the randomly drawn target class, $y^+ \sim p(y)$. Conditional on the target class y^+, $\mathbf{x}'_T$ denotes a mature counterfactual for a randomly sampled factual from a non-target class generated with a gradient-based CE generator for up to T iterations. Intuitively, the gradient of Equation 3 decreases the energy of observed training samples (positive samples) while increasing the energy of counterfactuals (negative samples) (Du and Mordatch 2020). As the counterfactuals get more plausible (Def. 3.1) during training, these opposing effects gradually balance each other out (Lippe 2024).

Since maturity of counterfactuals in terms of a probability threshold is often reached before T, this form of sampling is not only more closely aligned with Def. 3.1., but can also speed up training times compared to SGLD. The departure from SGLD also allows us to tap into the vast repertoire of explainers that have been proposed in the literature to meet different desiderata. For example, many methods support domain and mutability constraints. In principle, any existing approach for generating CEs is viable, so long as it does not violate the faithfulness condition. Like JEMs (Murphy 2022), counterfactual training can be considered a form of contrastive representation learning.

3.3 Indirectly Inducing Explainability with Adversarial Robustness

Based on our analysis in Section 2, counterfactuals $\mathbf{x}'$ can be repurposed as additional training samples (Balashankar et al. 2023; Luu and Inoue 2023) or adversarial examples (Freiesleben 2022; Pawelczyk et al. 2022). This leaves some flexibility with regards to the choice for the $\text{advloss}(\cdot)$ term in Equation 2. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

$$\begin{aligned} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t\leq T}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t\{t : ||\Delta_t||_\infty < \varepsilon\} \end{aligned} \tag{4}$$

Under this choice, we consider nascent counterfactuals $\mathbf{x}'_{t\leq T}$ as AEs as long as the magnitude of the perturbation to any single feature is at most ε. This is closely aligned with Szegedy et al. (2014) who define an adversarial attack as an "imperceptible non-random perturbation". Thus, we work with a different distinction between CE and AE than Freiesleben (2022) who considers misclassification as the distinguishing feature of adversarial examples. One of the key observations of this work is that we can leverage CEs during training and get AEs essentially for free to reap the aforementioned benefits of adversarial training.

3.4 Encoding Actionability Constraints

Many existing counterfactual explainers support domain and mutability constraints. In fact, both types of constraints can be implemented for any explainer that relies on gradient descent in the feature space for optimization (Altmeyer, van Deursen, and Liem 2023). In this context, domain constraints can be imposed by simply projecting counterfactuals back to the specified domain, if the previous gradient step resulted in updated feature values that were out-of-domain. Similarly, mutability constraints can be enforced by setting partial derivatives to zero to ensure that features are only perturbed in the allowed direction, if at all.

Since actionability constraints are binding at test time, we also impose them when generating $\mathbf{x}'$ during each training iteration to inform model representations. Through their effect on $\mathbf{x}'$, both types of constraints influence model outcomes via Equation 3. Here it is crucial that we avoid penalizing implausibility that arises due to mutability constraints. For any mutability-constrained feature d this can be achieved by enforcing $\mathbf{x}^+[d] - \mathbf{x}'[d] := 0$ whenever perturbing $\mathbf{x}'[d]$ in the direction of $\mathbf{x}^+[d]$ would violate mutability constraints. Specifically, we set $\mathbf{x}^+[d] := \mathbf{x}'[d]$ if:

1. Feature d is strictly immutable in practice.
2. $\mathbf{x}^+[d] > \mathbf{x}'[d]$, but d can only be decreased in practice.
3. $\mathbf{x}^+[d] < \mathbf{x}'[d]$, but d can only be increased in practice.

From a Bayesian perspective, setting $\mathbf{x}^+[d] := \mathbf{x}'[d]$ can be understood as assuming a point mass prior for $p(\mathbf{x}^+)$ wrt. feature d. Intuitively, we think of this as ignoring implausibility costs of immutable features, which effectively forces the model to instead seek plausibility through the remaining features. This can be expected to result in relatively lower sensitivity to immutable features; and higher relative sensitivity to mutable features should make mutability-constrained recourse less costly (Section 4). Under certain conditions, this result holds theoretically; for the proof, see the supplementary appendix:

Proposition 3.1 (Protecting Immutable Features)**.** *Let $f_\theta(\mathbf{x}) = \mathcal{S}(\mathbf{M}_\theta(\mathbf{x})) = \mathcal{S}(\Theta\mathbf{x})$ denote a linear classifier with softmax activation $\mathcal{S}$ where $y \in \{1, ..., K\} = \mathcal{K}$ and $\mathbf{x} \in \mathbb{R}^D$. Assume multivariate Gaussian class densities with common diagonal covariance matrix $\Sigma_k = \Sigma$ for all $k \in \mathcal{K}$, then protecting an immutable feature from the contrastive divergence penalty will result in lower classifier sensitivity to that feature relative to the remaining features, provided that at least one of those is discriminative and mutable.*

4 Experiments

We seek to answer the following four research questions:

- (RQ1) To what extent does the CT objective in Equation 1 induce models to learn plausible explanations?
- (RQ2) To what extent does CT result in more favorable algorithmic recourse outcomes in the presence of actionability constraints?
- (RQ3) To what extent does CT influence the adversarial robustness of trained models?
- (RQ4) What are the effects of hyperparameter selection on counterfactual training?

4.1 Experimental Setup

Our focus is the improvement in explainability (Def. 3.1). Thus, we primarily look at the plausibility and cost of faithfully generated counterfactuals at test time. Other metrics, such as validity and redundancy, are reported in the supplementary appendix. To measure the cost, we follow the standard proxy of distances (ℓ_1-norm) between factuals and counterfactuals. For plausibility, we assess how similar CEs are to observed samples in the target domain, $\mathbf{X}^+ \subset \mathcal{X}^+$. We rely on the metric used by Altmeyer et al. (2024),

$$\text{IP}(\mathbf{x}', \mathbf{X}^+) = \frac{1}{|\mathbf{X}^+|} \sum_{\mathbf{x} \in \mathbf{X}^+} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{5}$$

and introduce a novel divergence-based adaptation,

$$\text{IP}^*(\mathbf{X}', \mathbf{X}^+) = \text{MMD}(\mathbf{X}', \mathbf{X}^+) \tag{6}$$

where $\mathbf{X}'$ denotes a collection of counterfactuals and $\text{MMD}(\cdot)$ is the unbiased estimate of the squared population maximum mean discrepancy, proposed by Gretton et al. (2012). The metric in Equation 6 is equal to zero if and only if the two distributions are exactly the same, $\mathbf{X}' = \mathbf{X}^+$.

To assess outcomes with respect to actionability for non-linear models, we look at the average costs of valid counterfactuals in terms of their distances from factual starting points. While this an imperfect proxy of sensitivity, we hypothesize that CT can reduce these costs by teaching models to seek plausibility with respect to mutable features, much like we observe in Figure 1 in panel (d) compared to (c). We supplement this analysis with qualitative findings for integrated gradients (Sundararajan, Taly, and Yan 2017). Finally, for predictive performance, we use standard metrics, such as robust accuracy estimated on adversarially perturbed data using FGSM (Goodfellow, Shlens, and Szegedy 2015).

We run experiments with three gradient-based generators: *Generic* of Wachter, Mittelstadt, and Russell (2017) as a simple baseline approach, *REVISE* (Joshi et al. 2019) that aims to generate plausible counterfactuals using a surrogate Variational Autoencoder (VAE), and *ECCCo* (Altmeyer et al. 2024), which targets faithfulness.

We make use of nine classification datasets common in the CE/AR literature. Four of them are synthetic with two classes and different characteristics: linearly separable clusters (*LS*), overlapping clusters (*OL*), concentric circles (*Circ*), and interlocking moons (*Moon*). Next, we have four real-world binary tabular datasets: *Adult* (Census data) of Becker and Kohavi (1996), California housing (*CH*) of Pace and Barry (1997), Default of Credit Card Clients (*Cred*) of Yeh (2016), and Give Me Some Credit (*GMSC*) from Kaggle (2011). Finally, for the convenience of illustration, we use the 10-class *MNIST* (LeCun 1998).

To assess CT, we investigate the improvements in performance metrics when using it on top of a weak baseline (BL): a multilayer perceptron (*MLP*). This is the best way to get a clear picture of the effectiveness of CT, and it is consistent with evaluation practices in the related literature (Goodfellow, Shlens, and Szegedy 2015; Ross, Lakkaraju, and Bastani 2024; Teney, Abbasnedjad, and van den Hengel 2020).

Table 1: Key evaluation metrics for valid counterfactual along with bootstrapped standard errors for all datasets. **Plausibility** (columns 1-2): percentage reduction in implausibility for IP and IP*, respectively; **Cost / Actionability** (column 3): percentage reduction in costs when selected features are protected. Outcomes are aggregated across bootstrap samples and varying degrees of the energy penalty λ_{egy} used for *ECCCo* at test time. Asterisks (*) indicate that the bootstrapped 99%-confidence interval of differences in mean outcomes does *not* include zero.

Data	**IP** $(-\%)$	**IP*** $(-\%)$	**Cost** $(-\%)$
LS	$29.05 \pm 0.67^*$	$55.33 \pm 2.03^*$	$14.07 \pm 0.60^*$
Circ	$56.29 \pm 0.44^*$	$89.38 \pm 9.30^*$	$45.55 \pm 0.76^*$
Moon	$20.62 \pm 0.69^*$	$19.26 \pm 8.12^*$	$2.86 \pm 1.03^*$
OL	-1.13 ± 0.88	-24.52 ± 14.52	$38.39 \pm 2.21^*$
Adult	0.77 ± 1.34	$32.29 \pm 6.87^*$	-2.82 ± 4.88
CH	$12.05 \pm 1.41^*$	$70.27 \pm 3.72^*$	$40.71 \pm 1.55^*$
Cred	$12.31 \pm 1.84^*$	$54.89 \pm 11.21^*$	$-17.43 \pm 5.17^*$
GMSC	$23.44 \pm 1.99^*$	$73.31 \pm 4.83^*$	$62.64 \pm 2.04^*$
MNIST	$7.05 \pm 1.80^*$	-25.09 ± 109.05	-12.34 ± 6.52
Avg.	17.83	38.35	19.07

4.2 Experimental Results

Our main results for plausibility and actionability for *MLP* models are summarised in Table 1 that presents counterfactual outcomes grouped by dataset along with standard errors averaged across bootstrap samples. Asterisks (*) are used when the bootstrapped 99%-confidence interval of differences in mean outcomes does *not* include zero, so the observed effects are statistically significant at the 0.01 level.

The first two columns (IP and IP*) show the percentage reduction in implausibility for our two metrics when using CT on top of the weak baseline. As an example, consider the first row for *LS* data: the observed positive values indicate that faithful counterfactuals are around 30-55% more plausible for models trained with CT, in line with our observations in panel (b) of Figure 1 compared to panel (a).

The third column shows the results for a scenario when mutability constraints are imposed on the selected features. Again, we are comparing CT to the baseline, so reductions in the positive direction imply that valid counterfactuals are "cheaper" (more actionable) when using CT with feature protection. Relating this back to Figure 1, the third column represents the reduction in distances travelled by counterfactuals in panel (d) compared to panel (c). In the following paragraphs, we summarize the results for all datasets.

Plausibility (RQ1). *CT generally produces substantial and statistically significant improvements in plausibility.*

Average reductions in IP range from around 7% for *MNIST* to almost 60% for *Circ*. For the real-world tabular datasets they are around 12% for *CH* and *Cred* and almost 25% for *GMSC*; for *Adult* and *OL* we find no significant impact of CT on IP. Reductions in IP* are even more substantial and generally statistically significant, although the average degree of uncertainty is higher than for IP: reductions range from around 20% (*Moons*) to almost 90% (*Circ*). The only negative findings are for OL and MNIST, but they are not statistically significant. A qualitative inspection of the counterfactuals in Figure 2 (columns 2-5) suggests recognizable digits 1-4 for the model trained with CT (bottom row), unlike the baseline (top row).

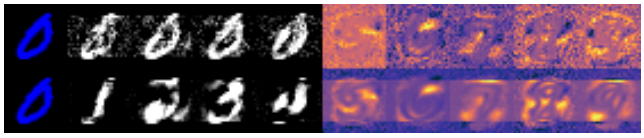

Figure 2: Visual explanations for *MNIST* for BL (top) and CT (bottom). **Plausibility**: col. 1 is a random factual 0 (blue); cols. 2-5 are corresponding *ECCCo* counterfactuals in target classes 1 to 4. **Actionability**: cols. 6-10 show integrated gradients averaged over test images in classes 5 to 9.

Actionability (RQ2). *CT tends to improve actionability in the presence of immutable features, but this is not guaranteed if the assumptions in Proposition 3.1 are violated.*

For synthetic datasets, we always protect the first feature; for all real-world tabular datasets we could identify and protect an *age* variable; for *MNIST*, we protect the five upper and lower pixel rows of the full image. Statistically significant reductions in costs overwhelmingly point in the expected positive direction reaching up to around 60% for *GMSC*. Only in the case of *Cred*, average costs increase, likely because any potential benefits from protecting the *age* are outweighed by the increase in costs required for greater plausibility. The findings for *Adult* and *MNIST* are not significant. A qualitative inspection of the class-conditional integrated gradients in Figure 2 (columns 6-10) suggests that CT still has the expected effect: the model (bottom) is insensitive (blue) to the protected rows of pixels; details of this experiment are reported in the supplementary appendix.

Predictive Performance (RQ3). *Models trained with CT are substantially more robust to gradient-based adversarial attacks than conventionally-trained baselines.*

Test accuracies on adversarially perturbed data are shown in Figure 3. The perturbations size, $\varepsilon \in [0, 0.1]$, increases along the horizontal axis and includes zero, corresponding to standard test accuracy for non-perturbed data. For all synthetic datasets, predictive performance of CT is virtually identical to the baseline and unaffected by perturbations. For all real-world datasets, we find that CT substantially improves robustness: while in some cases baseline accuracies drop to essentially zero for large enough perturbation sizes, accuracies of CT models remain remarkably robust.

Hyperparameter settings (RQ4). *CT is highly sensitive to the choice of a CE generator and its hyperparameters but (1) we observe manageable patterns, and (2) we can usually identify settings that improve either plausibility or actionability, and typically both of them at the same time.*

We evaluate the impacts of three types of hyperparameters on CT. In this section we focus on the highlights and make the full results available in the supplementary appendix.

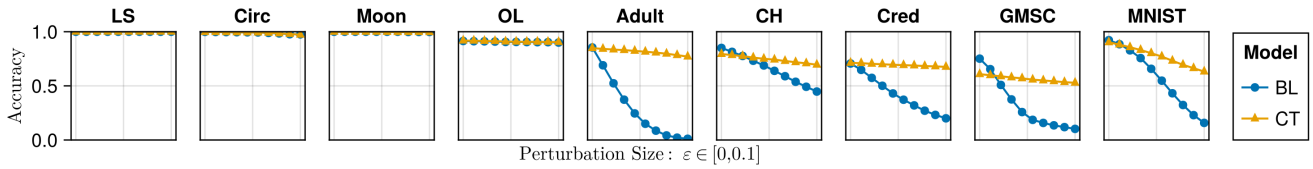

Figure 3: Test accuracies on adversarially perturbed data with varying perturbation sizes for all non-synthetic datasets.

Firstly, we find that optimal results are generally obtained when using *ECCCo* to generate counterfactuals. Conversely, using a generator that may inhibit faithfulness (*REVISE*), tends to yield poor results. Concerning hyperparameters that guide the gradient-based counterfactual search, we find that increasing T, the maximum number of steps, generally yields better outcomes because more CEs can mature. Relatedly, we also find that the effectiveness and stability of CT is positively associated with the total number of counterfactuals generated during each training epoch. The impact of τ, the decision threshold, is more difficult to predict. On "harder" datasets it may be difficult to satisfy high τ for any given sample (i.e., also factuals) and so increasing this threshold does not seem to correlate with better outcomes. In fact, $\tau = 0.5$ generally leads to optimal results as it is associated with high proportions of mature counterfactuals.

Secondly, the strength of the energy regularization, λ_{reg} is highly impactful and should be set sufficiently high to avoid common problems associated with exploding gradients. The sensitivity with respect to λ_{div} and λ_{adv} is much less evident. While high values of λ_{reg} may increase the variability in outcomes when combined with high values of λ_{div} or λ_{adv}, this effect is not particularly pronounced.

Finally, we also observe desired improvements when CT was combined with conventional training and applied only for the final 50% of epochs of the complete training process. Put differently, CT can improve the explainability of models in a post-hoc, fine-tuning manner.

5 Discussion

As our results indicate, counterfactual training produces models that are more explainable. Nonetheless, these advantages come at the cost of two important limitations.

Interventions on features have implications for fairness. We provide a method to modify the sensitivity of a model to certain features, which can be misused by enforcing explanations based on features that are more difficult to modify by a (group of) decision subjects. Such abuse could result in an unfairly assigned burden of recourse (Sharma, Henderson, and Ghosh 2020), threatening the equality of opportunity (Bell et al. 2024). Also, even if all immutable features are protected, there may exist proxies that are theoretically mutable, but preserve sufficient information about the principals to hinder these protections. Indeed, deciding on the actionability of features remains a major open challenge in the AR literature (Venkatasubramanian and Alfano 2020).

Plausibility is costly. As noted by Altmeyer et al. (2024), more plausible counterfactuals are inevitably more costly. CT improves plausibility and robustness, but it can impact average costs and validity when cheap, implausible and adversarial explanations are removed from the solution space.

CT increases the training times. Just like contrastive and robust learning, CT is more resource-intensive than conventional regimes. Three factors mitigate this effect: (1) CT yields itself to parallel execution; (2) it amortizes the cost of CEs for the training samples; and (3) it can be used to fine-tune conventionally-trained models.

We also highlight three key directions for future research. Firstly, it is an interesting challenge to extend CT beyond classification settings. Our formulation relies on the distinction between non-target class(es) and target class(es), requiring the output space to be discrete. Thus, it does not apply to ML tasks where the change in outcome cannot be readily discretized. Focus on classification is a common choice in research on CEs and AR; other settings have attracted some interest, e.g., regression (Spooner et al. 2021), but there is little consensus how to robustly extend the notion of CEs.

Secondly, our analysis covers CE generators with different characteristics, but it is interesting to extend it to more algorithms, including ones that do not rely on computationally costly gradient-based optimization. This should reduce training costs while possibly preserving the benefits of CT.

Finally, we believe that it is possible to considerably improve hyperparameter selection procedures, and thus performance. We have relied exclusively on grid searches, but future work could benefit from more sophisticated approaches.

6 Conclusion

State-of-the-art machine learning models are prone to learning complex representations that cannot be interpreted by humans. Existing explainability solutions cannot guarantee that explanations agree with these learned representation. As a step towards addressing this challenge, we introduce counterfactual training, a novel training regime that integrates recent advances in contrastive learning, adversarial robustness, and counterfactual explanations to incentivize highly-explainable models. Through extensive experiments, we demonstrate that CT satisfies this goal while preserving the predictive performance and promoting robustness of models. Explanations generated from CT-based models are both more plausible (compliant with the underlying data-generating process) and more actionable (compliant with user-specified mutability constraints), and thus meaningful to their recipients. In turn, our work highlights the value of simultaneously improving models and their explanations.

References

Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; and van den Hengel, A. 2020. Counterfactual Vision and Language Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10041–10051.

Altmeyer, P.; Farmanbar, M.; van Deursen, A.; and Liem, C. C. S. 2024. Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38, 10829–10837.

Altmeyer, P.; van Deursen, A.; and Liem, C. C. S. 2023. Explaining Black-Box Models through Counterfactuals. In *Proceedings of the JuliaCon Conferences*, volume 1, 130.

Augustin, M.; Meinke, A.; and Hein, M. 2020. Adversarial Robustness on In- and Out-Distribution Improves Explainability. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 228–245. Cham: Springer. ISBN 978-3-030-58574-7.

Balashankar, A.; Wang, X.; Qin, Y.; Packer, B.; Thain, N.; Chi, E.; Chen, J.; and Beutel, A. 2023. Improving Classifier Robustness through Active Generative Counterfactual Data Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 127–139. ACL.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Bell, A.; Fonseca, J.; Abrate, C.; Bonchi, F.; and Stoyanovich, J. 2024. Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity. ArXiv:2401.16088, arXiv:2401.16088.

Du, Y.; and Mordatch, I. 2020. Implicit Generation and Generalization in Energy-Based Models. ArXiv:1903.08689, arXiv:1903.08689.

Freiesleben, T. 2022. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*, 32(1): 77–109.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. ArXiv:1412.6572, arXiv:1412.6572.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *The Journal of Machine Learning Research*, 13(1): 723–773.

Guo, H.; Nguyen, T. H.; and Yadav, A. 2023. CounterNet: End-to-End Training of Prediction Aware Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 577—589. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.

Joshi, S.; Koyejo, O.; Vijitbenjaronk, W.; Kim, B.; and Ghosh, J. 2019. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. ArXiv:1907.09615, arXiv:1907.09615.

Kaggle. 2011. Give Me Some Credit, Improve on the State of the Art in Credit Scoring by Predicting the Probability That Somebody Will Experience Financial Distress in the next Two Years. https://www.kaggle.com/c/GiveMeSomeCredit. Accessed: 2023-12-14.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2021. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6405–6416. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510860964.

LeCun, Y. 1998. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

Lippe, P. 2024. UvA Deep Learning Tutorials. https://uvadlc-notebooks.readthedocs.io/en/latest/.

Luu, H. L.; and Inoue, N. 2023. Counterfactual Adversarial Training for Improving Robustness of Pre-trained Language Models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 881–888. ACL.

Molnar, C. 2022. *Interpretable Machine Learning*. 2 edition.

Murphy, K. P. 2022. *Probabilistic Machine Learning: An Introduction*. MIT Press.

Pace, R. K.; and Barry, R. 1997. Sparse Spatial Autoregressions. *Statistics & Probability Letters*, 33(3): 291–297.

Pawelczyk, M.; Agarwal, C.; Joshi, S.; Upadhyay, S.; and Lakkaraju, H. 2022. Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis. In Camps-Valls, G.; Ruiz, F. J. R.; and Valera, I., eds., *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 4574–4594. PMLR.

Ross, A.; Lakkaraju, H.; and Bastani, O. 2024. Learning Models for Actionable Recourse. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.

Sauer, A.; and Geiger, A. 2021. Counterfactual Generative Networks. ArXiv:2101.06046, arXiv:2101.06046.

Schut, L.; Key, O.; McGrath, R.; Costabello, L.; Sacaleanu, B.; Gal, Y.; et al. 2021. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, 1756–1764. PMLR.

Sharma, S.; Henderson, J.; and Ghosh, J. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models.

In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 166–172. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.

Spooner, T.; Dervovic, D.; Long, J.; Shepard, J.; Chen, J.; and Magazzeni, D. 2021. Counterfactual Explanations for Arbitrary Regression Models. ArXiv:2106.15212, arXiv:2106.15212.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. arXiv:1703.01365.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. ArXiv:1312.6199, arXiv:1312.6199.

Teh, Y. W.; Welling, M.; Osindero, S.; and Hinton, G. E. 2003. Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.*, 4(null): 1235–1260.

Teney, D.; Abbasnedjad, E.; and van den Hengel, A. 2020. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. In *Computer Vision - ECCV 2020*, 580–599. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58606-5.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.

Venkatasubramanian, S.; and Alfano, M. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 284–293. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Verma, S.; Boonsanong, V.; Hoang, M.; Hines, K. E.; Dickerson, J. P.; and Shah, C. 2022. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. arXiv:2010.10596.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wilson, A. G. 2020. The Case for Bayesian Deep Learning. ArXiv:2001.10995, arXiv:2001.10995.

Wu, T.; Ribeiro, M. T.; Heer, J.; and Weld, D. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6707–6723. Online: ACL.

Yeh, I.-C. 2016. Default of Credit Card Clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.