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 (1) plausible with respect to the underlying data and (2) actionable with respect to the user-defined 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 ad-hoc 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 explanations while promoting robustness and preserving high predictive performance.

1 Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by **representation learning**: instead of relying on features and rules hand-crafted by humans, modern machine learning (ML) models are tasked with learning representations directly from the data, guided by narrow objectives such as predictive accuracy (Goodfellow, Bengio, and Courville 2016). Advances in computing have made it possible to provide these models with ever-growing degrees of freedom to achieve this task, which often allows them to outperform traditionally parsimonious models. Unfortunately, in doing so, models learn increasingly complex, sensitive representations that humans can no longer easily interpret.

The trend towards complexity for the sake of performance has come under scrutiny in recent years. At the very cusp of the deep learning (DL) revolution, Szegedy et al. (2014) showed that artificial neural networks (ANN) are susceptible to adversarial examples (AEs): perturbed versions of data instances that yield vastly different model predictions despite being semantically indistinguishable from their factual counterparts. Some partial mitigation strategies have been proposed—most notably **adversarial training** (Goodfellow, Shlens, and Szegedy 2015)—but truly robust deep learning remains unattainable even for models that are considered "shallow" by today's standards (Kolter 2023).

Part of the problem is that the high degrees of freedom—high number of parameters estimated from data—provide room for many solutions that are locally optimal with respect to narrow objectives (Wilson 2020). As one example, research on the "lottery ticket hypothesis" suggests that modern neural networks can be pruned by up to 90% without losing predictive performance (Frankle and Carbin 2019). Thus, looking at the predictive performance alone, found solutions may seem to provide compelling explanations for the data, when in fact they are based on purely associative and semantically meaningless patterns. This poses two related challenges. Firstly, there is no dependable way to verify if learned representations correspond to meaningful, plausible explanations. Secondly, even if we resolve this challenge, it remains undecided how to ensure that machine learning models can *only* learn valuable explanations.

The first challenge has attracted an abundance of work on **explainable AI** (XAI), a paradigm that focuses on the development of tools to derive (post-hoc) explanations from complex model representations, aiming to mitigate scenarios in which practitioners deploy opaque models and have to blindly rely on their predictions. On many occasions, this has happened in practice, causing harms to people who were adversely and unfairly affected by automated decision-making (ADM) systems involving opaque models; see, e.g., O'Neil (2016). Effective XAI tools can also aid in monitoring models and providing recourse, empowering people to turn negative outcomes (e.g., "loan application rejected") into positive ones (e.g., "loan application accepted"). In line with this, our work builds upon **counterfactual explanations** (CE) 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.

To our surprise, the second challenge has not yet attracted major research interest. In particular, there has been no concerted effort towards improving the degree to which learned representations promote explanations that are both **interpretable** to and deemed **plausible** by humans. Instead, the typical choice has been to improve the ability of XAI tools to identify the subset of explanations that are plausible and valid for any given model, independent of whether these explanations are compatible with the learned representations (Altmeyer et al. 2024). Fortunately, recent findings indicate that improved "explanatory capacity" of a model can arise as a consequence of regularization techniques aimed at other training objectives such as generative capacity, generalization, or robustness (Altmeyer et al. 2024; Augustin, Meinke, and Hein 2020; Schut et al. 2021). Our contribution consolidates these findings within a unified framework.

Specifically, **we propose Counterfactual Training (CT)**: a novel training regime explicitly geared towards improving the explanatory capacity of models that, in high-level terms, we define as the extent to which valid explanations derived for a model can be deemed plausible with respect to the underlying data and global actionability constraints (we refine this notion in Def. 3.1). For simplicity, we refer to models with high explanatory capacity as *explainable*. To the best of our knowledge, Counterfactual Training represents the first attempt to achieve more explainable models by employing counterfactual explanations already in the training phase.

The remainder of this manuscript is structured as follows. Section 2 presents related work, focusing on the link between AEs and CEs. Then follow our two principal contributions. In Section 3, we introduce our methodological framework and show theoretically that it can be employed to enforce global actionability constraints. In Section 4, through extensive experiments, we empirically demonstrate that CT substantially improves explainability and positively contributes to the robustness of trained models without sacrificing predictive performance. Finally, in Section 5, we discuss open challenges and conclude that CT is a promising approach towards making opaque models more trustworthy.

2 Related Literature

To make the desiderata for our framework more concrete, we follow Augustin, Meinke, and Hein (2020) in tying explainability to the quality of CEs that can be generated for a given model. The authors show that CEs (understood as minimal input perturbations that yield some desired model prediction) tend to be more meaningful if the underlying model is more robust to adversarial examples. 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. As argued before, learned representations may be sensitive to producing implausible explanations and mispredicting for worst-case counterfactuals (i.e., AEs). Thus, by inducing models to "unlearn" susceptiblity to such examples, adversarial training can effectively remove implausible explanations from the solution space.

2.1 Adversarial Examples are Counterfactuals

The 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 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.

2.2 Aligning Representations with Explanations

Improving the adversarial robustness of models is not the only path towards aligning representations with plausible explanations. In a closely related work, Altmeyer et al. (2024) show that explainability can be improved through model averaging and refined model objectives. They propose a way to generate counterfactuals that are maximally faithful to the model in that they are consistent with what the model has learned about the underlying data. Formally, they rely on tools from energy-based modelling (Teh et al. 2003) to minimize the divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by the model. Their counterfactual explainer, *ECCCo*, yields plausible 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) tend to do well in this regard.

Once again 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 aforementioned risk of overrelying on a single locally optimal representation that corresponds to semantically meaningless explanations for the data. 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 bears resemblance to the idea of aligning models with plausible explanations and has inspired our CT objective.

3 Counterfactual Training

This section introduces the Counterfactual Training framework. CT combines ideas from adversarial training, counterfactual explanations, and energy-based modelling with the explicit goal of producing models whose learned representations align with plausible explanations that further comply with user-defined actionability constraints.

In the context of counterfactual explanations, plausibility has broadly been defined as the degree to which generated CEs comply with the underlying data-generating process (Altmeyer et al. 2024; Guidotti 2022; Poyiadzi et al. 2020). Plausibility is a necessary but insufficient condition for using CEs to provide algorithmic recourse (AR) to individuals (negatively) affected by opaque models. An AR recommendations must also be actionable, i.e., possible to attain by the recipient. A plausible CE for a rejected 20-year-old loan applicant, for example, might reveal that their application would have been accepted, if only they had been 20 years older. Ignoring all other features, this would comply with the definition of plausibility if 40-year-old individuals were in fact more credit-worthy on average than young adults. But of course this CE does not qualify for providing actionable recourse to the applicant since *age* is not a (directly) mutable feature. Counterfactual training aims to improve model explainability by aligning models with counterfactuals that meet both desiderata: plausibility and actionability. 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 counterfactuals are also plausible and actionable. We define these properties as follows:

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

and $p_\theta(\mathbf{x}|\mathbf{y}^+)$ 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 is the same as in 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. Actionability 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.

3.1 Proposed Objective

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). In broad terms, searching for CEs using gradient descent entails optimizing some form of an objective that balances (1) the classification loss $\text{yloss}(\mathbf{M}_\theta(\mathbf{x}),\mathbf{y})$, and (2) one or more penalty terms $\lambda_i \text{cost}_i(\cdot)$. The exact specification of these penalties induces various properties in the counterfactual outcomes, and tends to be the key feature that distinguishes various gradient-based "generators" or "explainers" in the literature (Altmeyer, van Deursen, and Liem 2023), including all generators used in our experiments. We refer the reader to the supplementary appendix for details.

CT adopts gradient-based CE search during training to generate on-the-fly model explanations $\mathbf{x}'$ for training samples. We use the term *nascent* to denote counterfactuals $\mathbf{x}'_{t \leq T}$ that are not yet valid where t indicates the last iteration before the label is flipped. We store and use these interim counterfactuals as adversarial examples. Conversely, we consider counterfactuals $\mathbf{x}'_T$ as *mature* explanations if they have either exhausted all T iterations or converged in terms 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:

$$\begin{aligned}\min_\theta \text{yloss}(\mathbf{M}_\theta(\mathbf{x}),\mathbf{y}) + \lambda_{\text{div}}\text{div}(\mathbf{x}^+,\mathbf{x}'_T,y;\theta) \\ +\lambda_{\text{adv}}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t \leq T}),\mathbf{y}) + \lambda_{\text{reg}}\text{ridge}(\mathbf{x}^+,\mathbf{x}'_T,y;\theta)\end{aligned} \tag{1}$$

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 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 the adversarial loss, advloss(.), wrt. nascent counterfactuals $\mathbf{x}'_{t \leq T}$. Finally, ridge$(\cdot)$ denotes a Ridge penalty (ℓ_2-norm) that regularizes the magnitude of the energy terms involved in div$(\cdot)$ (Du and Mordatch 2020). The trade-off between the components are governed through λ_{div}, λ_{adv} and λ_{reg}.

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 1. To optimize $\log p_\theta(\mathbf{x})$, they minimize the contrastive divergence between the observed samples from $p(\mathbf{x})$ and the generated samples from $p_\theta(\mathbf{x})$.

A key empirical finding of Altmeyer et al. (2024) was that JEMs perform well on the plausibility objective in Def. 3.1. This follows directly if we consider samples drawn from $p_\theta(\mathbf{x})$ as counterfactuals — the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|\mathbf{y}^+)$. 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. Instead we propose to leverage counterfactual explainers to generate counterfactuals of observed training samples. Specifically, we have:

$$\text{div}(\mathbf{x}^+, \mathbf{x}'_T, y; \theta) = \mathcal{E}_\theta(\mathbf{x}^+, y) - \mathcal{E}_\theta(\mathbf{x}'_T, y) \tag{2}$$

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. Mature counterfactuals are ones that have either reached convergence wrt. the decision threshold τ or exhausted T.

Intuitively, the gradient of Equation 2 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).

The departure from SGLD 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 advloss$(\cdot)$ term in Equation 1. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

$$\begin{aligned}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{t\leq T}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t\{t : ||\Delta_t||_\infty < \varepsilon\}\end{aligned} \tag{3}$$

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 out-of-the-box. 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 should 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 2. 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 lower overall sensitivity to immutable features, which we investigate empirically in 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 three research questions:

1. To what extent does the CT objective in Equation 1 induce models to learn plausible explanations?
2. To what extent does CT lead to more favorable AR outcomes in the presence of actionability constraints?
3. What are the effects of hyperparameter selection on CT?

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 the observed samples in the target domain, $\mathbf{X}^+ \subset \mathcal{X}^+$. We rely on the metric used by Altmeyer et al. (2024),

$$\text{IP}(\mathbf{x}', \mathbf{X}^+) = \frac{1}{|\mathbf{X}^+|} \sum_{\mathbf{x} \in \mathbf{X}^+} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{4}$$

and introduce a novel divergence metric,

$$\text{IP}^*(\mathbf{X}', \mathbf{X}^+) = \text{MMD}(\mathbf{X}', \mathbf{X}^+) \tag{5}$$

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 5 is equal to zero if and only if the two distributions are exactly the same, $\mathbf{X}' = \mathbf{X}^+$.

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 *ECCo*—the generator of Altmeyer et al. (2024) without the conformal prediction component—as a method that directly targets both faithfulness and plausibility of the counterfactuals.

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*). They are generated using the library of Altmeyer, van Deursen, and Liem (2023) and we present them in the supplementary appendix. 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).

4.2 Experimental Results

Our main quantitative results for *MLP* models are summarised in Table 1, which presents average outcomes along with bootstrapped two standard errors. The following example motivates CT and illustrates how to read Table 1.

Synthetic Example: Prediction of Credit Card Defaults. Figure 1 presents results for a linear classifier fitted to *LS* that complies with the data assumptions in Proposition 3.1. The four panels show the outcomes for different training procedures: in panels (a) and (c) we have trained the models conventionally, while in panels (b) and (d) we have applied CT. For illustrative purposes, suppose the first feature represents *debt* (mutable) and the second feature represents *age* (immutable) of loan applicants seeking counterfactual explanations for moving to the target class: loan provided (orange).

In all four cases, it is possible to generate valid counterfactuals (stars) for unsuccessful applicants (blue). They cross the decision boundary (green) into the target class, but their quality differs. In panel (a), they are not plausible: they do not comply with the distribution of the factuals in y^+ to the point where they form a clearly discernible cluster. In panel (b), they are highly plausible, meeting the first objective of Def. 3.1. This difference in outcomes is quantified for the non-linear MLP in the first two columns of Table 1 as the %-reduction in implausibility: it is substantial and statistically significant for *LS* across both metrics, the distance-based IP (29%) and divergence-based IP^* (55%).

In panel (c) of Figure 1, the CEs involve substantial reductions in *debt* for younger applicants. By comparison, counterfactual paths are shorter on average in panel (d) where we have protected the immutable *age* as described in Section 3.4. Due to the classifier's lower sensitivity to *age*, recommendations with respect to *debt* are much more homogenous and do not unfairly punish younger individuals. These CEs are also plausible with respect to the mutable feature, despite requiring smaller debt reductions on average, resulting in smaller costs to individuals. This result is quantified for the non-linear case in column three of Table 1, which shows the %-reduction in costs averaged across valid counterfactuals. Once again, the impact of CT is statistically significant and substantial (14%). Thus, we consider the model in panel (d) as the most explainable according to Def. 3.1. Next, we present the results for all remaining datasets.

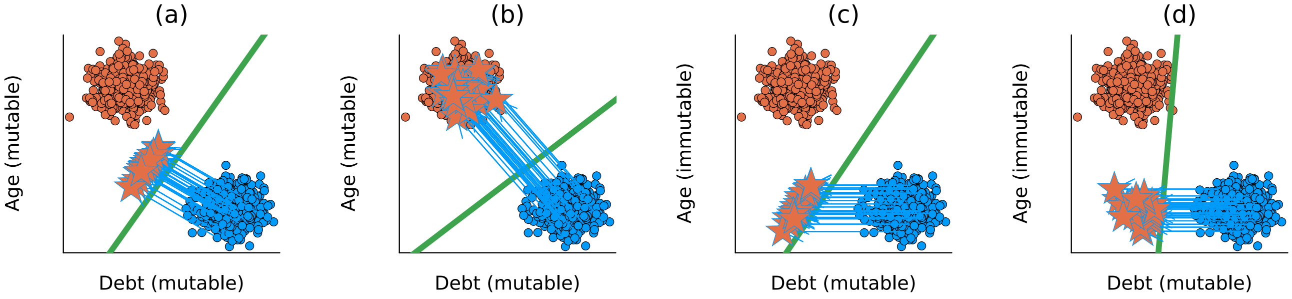

Figure 1: Illustration of how CT improves model explainability: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable. The linear decision boundary is shown in green along with training data colored according to ground-truth labels: $y^- = 1$ (blue) and $y^+ = 2$ (orange). Stars indicate counterfactuals in the target class.

Plausibility. We find that CT generally leads to 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 both *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: average reductions range from around 20% (*Moons*) to almost 90% (*Circ*). The only negative findings for *OL* and *MNIST* are statistically insignificant and, for MNIST, do not align with qualitative findings, which are much more plausible for CT (Figure 2).

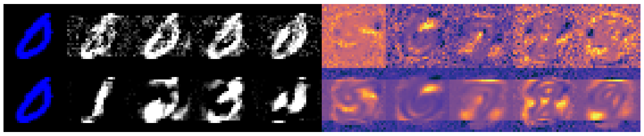

Figure 2: Sample explanations for *MNIST* for BL (top) and CT (bottom). First column is a random factual 0 (blue). Columns 2 to 5 are corresponding *ECCo* counterfactuals in target classes 1 to 4. Columns 6 to 10 show integrated gradients averaged over all test images in classes 5 to 9.

Actionability. We also find that CT can reduce sensitivity to immutable, protected features and thus lead to less costly counterfactual outcomes as shown in Figure 1. In column three of Table 1, we impose mutability constraints on selected features and compute the reduction in average costs of CEs associated with CT compared to the baseline: 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 top and bottom rows of digits. Reductions in costs are overwhelmingly positive and significant of up to nearly 60% for *GMSC*. While the estimated cost reductions for *Adult* and *MNIST* are not significant, Figure 2 (columns 6-10) demonstrates that CT does have the expected effect: sensitivity to protected features as per integrated gradients is drastically reduced; details of this experiment are reported in the supplementary appendix. 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.

Predictive Performance. Test accuracy for CT is virtually identical to the baseline for *Adult*, *Circ*, *LS*, *Moon*, and *OL*, and even slightly improved for *Cred*. Exceptions to this general pattern are *MNIST*, *CH*, and *GMSC*, for which we observe a reduction in test accuracy of 2, 5, and 15 percentage points respectively. When looking at robust test accuracies (Acc.*) for these datasets in particular, we find that CT strongly outperforms the baseline. In fact, we observe that CT improves adversarial robustness on all datasets.

Hyperparameter settings. We test the impact of three types of hyperparameters. Here we focus on the highlights; full results are available in the supplementary appendix.

First, we note that CT is highly sensitive to the choice of a CE generator and its hyperparameters but (1) there are manageable patterns, and (2) we can usually identify settings that improve either plausibility or cost, and often both of them at the same time. For example, *REVISE* tends to perform the worst, most likely because it uses a surrogate VAE to generate counterfactuals which impedes faithfulness (Altmeyer et al. 2024). Increasing T, the maximum number of steps, generally yields better outcomes because more CEs can mature in each training epoch. The impact of τ, the required 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.

Second, the strength of the energy regularization, λ_{reg} is highly impactful and leads to poor performance in terms of decreased plausibility and increased costs if insufficiently high. 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 very pronounced.

Third, the effectiveness and stability of CT is positively associated with the number of counterfactuals generated during each training epoch. A higher number of training epochs is also beneficial. Interestingly, we observed desired improvements when CT was combined with conventional

Table 1: Key performance metrics and bootstrapped *two* standard errors for all datasets. **Plausibility** (columns 1-2): percentage reduction in implausibility for IP and IP^*, respectively; **Actionability** (column 3): percentage reduction in costs with protected features. **Accuracy** (columns 4-7): test accuracies and robust accuracies (Acc*) for CT and the baseline (BL). Counterfactual outcomes in columns 1-3 are aggregated across bootstrap samples and varying degrees of the energy penalty λ_{egy} used for *ECCo* at test time. Standard errors for accuracy are bootstrapped from the test set.

	IP $(-\%)$	**IP*** $(-\%)$	**Cost** $(-\%)$	**Acc. (CT)**	**Acc. (BL)**	**Acc.*** **(CT)**	**Acc.*** **(BL)**
Adult	0.77 ± 2.69	32.29 ± 13.74	-2.82 ± 9.77	0.85 ± 0.01	0.85 ± 0.01	0.83 ± 0.01	0.41 ± 0.01
CH	12.05 ± 2.82	70.27 ± 7.43	40.71 ± 3.09	0.79 ± 0.01	0.85 ± 0.01	0.76 ± 0.01	0.74 ± 0.01
Circ	56.29 ± 0.89	89.38 ± 18.60	45.55 ± 1.52	1.0	1.0	0.99 ± 0.01	1.0
Cred	12.31 ± 3.67	54.89 ± 22.41	-17.43 ± 10.34	0.71 ± 0.02	0.71 ± 0.02	0.70 ± 0.02	0.52 ± 0.02
GMSC	23.44 ± 3.99	73.31 ± 9.65	62.64 ± 4.08	0.61 ± 0.02	0.75 ± 0.02	0.58 ± 0.02	0.42 ± 0.02
LS	29.05 ± 1.34	55.33 ± 4.05	14.07 ± 1.19	1.0	1.0	1.0	1.0
MNIST	7.05 ± 3.61	-25.09 ± 218.10	-12.34 ± 13.04	0.90 ± 0.01	0.92 ± 0.01	0.84 ± 0.01	0.78 ± 0.01
Moon	20.62 ± 1.38	19.26 ± 16.25	2.86 ± 2.06	1.0	1.0	1.0	1.0
OL	-1.13 ± 1.75	-24.52 ± 29.03	38.39 ± 4.41	0.92 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02

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 fine-tuning manner.

5 Conclusions

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 tool that allows practitioners to modify the sensitivity of a model with respect to certain features. Model owners can use our solution to support the fair and equitable treatment of decision subjects, but they could also misuse it by enforcing explanations based on features that are more difficult to modify by some (group of) individuals. When used irresponsibly, CT could result in an unfairly assigned burden of recourse (Sharma, Henderson, and Ghosh 2020), threatening the equality of opportunity (Bell et al. 2024). Additionally, CT requires mutability constraints for the features considered by the model. Even if all immutable features are protected, there may exist proxies that are mutable, and hence should not be protected, but preserve sufficient information about the principals to hinder these protections. Deciding on actionability is still a major open challenge in the AR literature (Venkatasubramanian and Alfano 2020) impacting the capacity of CT to fulfill its intended goal.

CT increases the training times. Like adversarial training, CT is more resource-intensive than conventional regimes. Higher numbers of CEs improve the quality of learned representations but they also increase the number of computations. As our codebase is not performance optimized, grids of 270 settings for the largest datasets in our experiments took up to four hours using 34 2GB CPUs (see supplementary appendix). Other than optimization, 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 important 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) y^- and target class(es) y^+ to generate counterfactuals through Equation 1. While y^- and y^+ can be arbitrarily defined, CT requires the output space $\mathcal{Y}$ 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, CT is susceptible to training instabilities. This problem has been recognized for JEMs (Grathwohl et al. 2020) and even though we depart from the SGLD sampling, we still encounter variability in outcomes. CT is exposed to two potential sources of instabilities: (1) the energy-based contrastive divergence term, $\text{div}(\cdot)$, in Equation 2, and (2) the underlying explainers. We find several promising ways to mitigate this problem: regularizing energy (λ_{reg}), generating sufficiently many counterfactuals during each epoch, and including only mature counterfactuals in $\text{div}(\cdot)$.

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 research could benefit from more sophisticated approaches.

To conclude, 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 the model's learned representation of data. As a step towards addressing this challenge, we introduced counterfactual training, a novel training regime that incentivizes highly-explainable models. Our approach leads to explanations that are both plausible (compliant with the underlying data-generating process) and actionable (compliant with user-specified mutability constraints), and thus meaningful to their recipients. Through extensive experiments, we demonstrated that CT satisfies its objective while promoting robustness and preserving the predictive performance of models. It can also be used to fine-tune conventionally-trained models and achieve similar gains. Lastly, 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.

Frankle, J.; and Carbin, M. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations*.

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.

Guidotti, R. 2022. Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking. *Data Mining and Knowledge Discovery*, 38(5): 2770–2824.

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.

Kolter, Z. 2023. Keynote Addresses: SaTML 2023 . In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Los Alamitos, CA, USA: IEEE Computer Society.

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.

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

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

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.

Poyiadzi, R.; Sokol, K.; Santos-Rodriguez, R.; De Bie, T.; and Flach, P. 2020. FACE: Feasible and Actionable Counterfactual Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

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.

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.

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.

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 *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. ACL.

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