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.

Introduction

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

The trend towards complexity for the sake of performance has come under serious scrutiny in recent years. At the very cusp of the deep learning (DL) revolution, **?** showed that artificial neural networks (ANN) are sensitive to adversarial examples (AEs): perturbed versions of data instances that yield vastly different model predictions despite being "imperceptible" in that they are semantically indifferent from their factual counterparts. Even though some partially effective mitigation strategies have been proposed—most notably **adversarial training** (**?**)—truly robust deep learning remains unattainable even for models that are considered "shallow" by today's standards (**?**).

Part of the problem is that the high degrees of freedom provide room for many solutions that are locally optimal with respect to narrow objectives (**?**).[1] Indeed, recent work on the so-called "lottery ticket hypothesis" suggests that modern neural networks can be pruned by up to 90% while preserving their predictive performance (**?**). Similarly, **?** showed that state-of-the-art neural networks are expressive enough to fit randomly labeled data. Thus, looking at the predictive performance alone, the solutions may seem to provide compelling explanations for the data, when in fact they are based on purely associative, semantically meaningless patterns. This poses two challenges. Firstly, there is no dependable way to verify if representations correspond to meaningful, plausible explanations. Secondly, even if we could resolve the first challenge, it remains undecided how to ensure that models can *only* learn valuable explanations.

The first challenge has attracted an abundance of research on **explainable AI** (XAI), a paradigm that focuses on the development of tools to derive (post-hoc) explanations from complex model representations. Such explanations should mitigate a scenario in which practitioners deploy opaque models and blindly rely on their predictions. On countless occasions, this has happened in practice and caused real harms to people who were adversely and unfairly affected by automated decision-making (ADM) systems involving opaque models (see, e.g., (**?**)). Effective XAI tools can aid us in monitoring models and providing recourse to individuals to turn negative outcomes (e.g., "loan application rejected") into positive ones (e.g., "application accepted"). In line with this, our work builds upon **counterfactual explanations** (CE) proposed by **?** as an effective approach to achieve this goal. CEs prescribe minimal changes for factual inputs that, if implemented, would prompt some fitted model to produce a desired output.

To our surprise, the second challenge has not yet attracted major research interest. Specifically, there has been no concerted effort towards improving the "explanatory capacity" of models, i.e., the degree to which learned representations correspond to explanations that are **interpretable**

[1] We follow the standard ML convention, where "degrees of freedom" refer to the number of parameters estimated from data.

and deemed **plausible** by humans (see Def. **??**). Instead, the choice has generally been to improve the ability of XAI tools to identify the subset of explanations that are both plausible and valid for any given model, independent of whether the learned representations are in fact compatible with plausible explanations (**?**). Fortunately, recent findings indicate that improved explanatory capacity can arise as a consequence of regularization techniques aimed at other training objectives such as generative capacity, generalization, or robustness (**???**). As further discussed in Section **??**, our work consolidates these findings within a single objective.

Specifically, we introduce Counterfactual Training (CT): a novel training regime explicitly geared towards improving the explainability of models. In high-level terms, we define this concept as as the extent to which valid explanations derived for an opaque model are also deemed plausible with respect to the underlying data and the global actionability constraints. To the best of our knowledge, our framework represents the first attempt to address this challenge by employing counterfactual explanations already in the training phase.

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

Related Literature

To make the desiderata for our framework more concrete, we follow **?** in tying the concept of 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 when looking 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.

Adversarial Examples are Counterfactual Explanations

This interpretation of the link between explainability through counterfactuals on one side and robustness to adversarial examples on the other is backed by empirical evidence. **?** demonstrates that using counterfactual images during classifier training improves model robustness. Similarly, **?** argues 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, demonstrate that augmenting the training data of image classifiers can improve generalization. Finally, **?** proposes an approach using counterfactuals in training that does not rely on data augmentation: they argue that counterfactual pairs typically already exist in training datasets. Specifically, their approach relies on identifying similar input samples with different annotations and ensuring that the gradient of the classifier aligns with the vector between such pairs of counterfactual inputs using the cosine distance as the loss function.

In the natural language processing (NLP) domain, CEs have also been used to improve models through data augmentation. **?** proposes *Polyjuice*, a general-purpose counterfactual generator for language models. The authors empirically demonstrate that the augmentation of training data with their method improves robustness in a number of NLP tasks. **?** similarly uses *Polyjuice* to augment NLP datasets through diverse counterfactuals and show that classifier robustness improves by up to 20%. Finally, **?** introduces Counterfactual Adversarial Training (CAT) that also aims to improve generalization and robustness of language models through a three-step procedure: the authors identify training samples that are subject to high predictive uncertainty, generate CEs for them, and fine-tune the language model on a dataset augmented with the CEs.

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, **?** 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 (**?**). Similarly, **?** shows that CEs and AEs are equivalent under certain conditions.

Two recent 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, **?** proposes a way to train models that guarantee individual recourse to some positive target class with high probability. Their approach builds on adversarial training by explicitly inducing susceptibility to targeted adversarial examples for the positive class. Additionally, the proposed method allows for imposing a set of actionability constraints ex-ante. For example, users can specify that certain features are immutable. Secondly, **?** is the first to propose an end-to-end training pipeline that includes CEs as part of the training procedure. In particular, they propose a specific network architecture that includes a predictor and CE generator network, where the parameters of the CE generator network are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor network. In contrast to **?**, we impose no restrictions on the ANN architecture at all.

Aligning Representations with Plausible Explanations

Improving the adversarial robustness of models is not the only path towards aligning representations with plausible explanations. In a work closely related to this one, **?** shows that explainability can be improved through model averaging and refined model objectives. The authors 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 (**?**) to minimize the divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by the model. Their proposed 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 (**?**) and joint energy-based models (JEMs) (**?**) 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 most called for when models are underspecified by the available data (**?**). Averaging across solutions mitigates the aforementioned risk of relying on a single locally optimal representations that corresponds to semantically meaningless explanations for the data. Previous work of **?** similarly found that generating plausible ("interpretable") CEs is almost trivial for deep ensembles that have also undergone adversarial training. The case for JEMs is even clearer: they involve a hybrid objective that induces both high predictive performance and generative capacity (**?**). This is closely related to the idea of aligning models with plausible explanations and has inspired our CT objective.

Counterfactual Training

In this section we propose our novel counterfactual training objective. In CT, we combine ideas from adversarial training, counterfactual explanations, and energy-based modelling with the explicit goal of aligning representations with plausible explanations that comply with user-defined actionability constraints.

In the context of counterfactual explanations, plausibility has broadly been defined as the degree to which counterfactuals comply with the underlying data-generating process (**???**). Plausibility is a necessary but insufficient condition for using CEs to provide algorithmic recourse (AR) to individuals (negatively) affected by opaque models. To be actionable, AR recommendations must also be attainable. 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 were 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 0.1 (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 $\arg\max_y \mathbf{y}^+ = y^+$ and y^+ denotes the index of the target class.

We say that $\mathbf{M}_\theta$ is **explainable** to the extent that faithfully generated counterfactuals are plausible and actionable. We define these properties as follows:

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

where $p_\theta(\mathbf{x}|\mathbf{y}^+)$ denotes the conditional posterior over inputs.

The characterization of faithfulness and plausibility in Def. **??** is the same as in **?**, with adapted notation. Intuitively, plausible counterfactuals are consistent with the data and faithful counterfactuals are consistent with what the model has learned about input data. Actionability constraints in Def. **??** vary and depend on the context in which $\mathbf{M}_\theta$ is deployed. In this work, we focus on domain and mutability constraints for individual features x_d for $d = 1, ..., D$. We limit ourselves to classification tasks for reasons discussed in Section **??**.

Our Proposed Objective

Let $\mathbf{x}'_t$ for $t = 0, ..., T$ denote a counterfactual explanation generated through gradient descent over T iterations as initially proposed by **?**. For our purposes, we let T vary and consider the counterfactual search as converged as soon as the predicted probability for the target class has reached a pre-determined threshold, τ: $\mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau$, where $\mathcal{S}$ is the softmax function.[2]

To train models with high explainability as defined in Def. **??**, we propose to leverage counterfactuals in the following objective:

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

where yloss$(\cdot)$ is a 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, div$(\cdot)$,

[2]For detailed background information on gradient-based counterfactual search and convergence see supplementary appendix.

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, advloss$(.)$, with respect to 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$(\cdot)$ (**?**). The trade-off between the components can governed through penalties λ_{div}, λ_{adv} and λ_{reg}.

Directly Inducing Explainability with Contrastive Divergence

As observed by **?**, any classifier can be re-interpreted as a joint energy-based model (JEM) 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 factorized as $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$. The first term can be optimized using conventional cross-entropy as in Equation **??**. To optimize $\log p_\theta(\mathbf{x})$, **?** minimizes the contrastive divergence between the observed samples from $p(\mathbf{x})$ and generated samples from $p_\theta(\mathbf{x})$.

A key empirical finding in **?** was that JEMs tend to do well with respect to the plausibility objective in Def. **??**. This follows directly if we consider samples drawn from $p_\theta(\mathbf{x})$ as counterfactuals because the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|\mathbf{y}^+)$. To generate samples, **?** relies on Stochastic Gradient Langevin Dynamics (SGLD) using an uninformative prior for initialization but we depart from their methodology. Instead of SGLD, we propose to use counterfactual explainers to generate counterfactuals of observed training samples. Specifically, we have:

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

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 **??** decreases the energy of observed training samples (positive samples) while increasing the energy of counterfactuals (negative samples) (**?**). As the counterfactuals get more plausible (Def. **??**) during training, these opposing effects gradually balance each other out (**?**).

The departure from SGLD of (**?**) 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 facilitate the imposition of domain and mutability constraints. In principle, any existing approach for generating counterfactual explanations is viable, so long as it does not violate the faithfulness condition. Like JEMs (**?**), CT can be considered a form of contrastive representation learning.

Indirectly Inducing Explainability with Adversarial Robustness

Based on our analysis in Section **??**, counterfactuals $\mathbf{x}'$ can be repurposed as additional training samples (**??**) or adversarial examples (**??**). This leaves some flexibility wrt. the choice for advloss$(\cdot)$ in Equation **??**. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

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

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 **?** that defines an adversarial attack as an "imperceptible non-random perturbation". Thus, we choose to work with a different distinction between CE and AE than **?** that considers misclassification as the key distinguishing feature of AE. One of the key observations of this work is that we can leverage CEs during training and get adversarial examples essentially for free, which can be used to reap the aforementioned benefits of adversarial training.

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 counterfactual explainer that relies on gradient descent in the feature space for optimization (**?**). 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. Mutability constraints can similarly be enforced by setting partial derivatives to zero to ensure that features are only perturbed in the allowed direction, if at all.

Since such 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 **??**. 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. We have $\mathbf{x}^+[d] > \mathbf{x}'[d]$, but feature d can only be decreased in practice.
3. We have $\mathbf{x}^+[d] < \mathbf{x}'[d]$, but feature 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 $p(\mathbf{x}^+)$ with respect to feature d. Intuitively, we think of this simply in terms ignoring implausibility costs with respect to immutable features, which effectively forces the model to instead seek plausibility with respect to the remaining features. This in turn results in lower overall sensitivity to immutable features, which we demonstrate empirically for different classifiers in Section **??**. Under certain conditions, this result holds theoretically:[3]

Proposition 0.1 (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 $\mathbf{x} \in \mathbb{R}^D$. If we 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.*

It is worth highlighting that Proposition **??** assumes independence of features. This raises a valid concern about the effect of protecting immutable features in the presence of proxies that remain unprotected. We address this in Section **??**.

Example (Prediction of Consumer Credit Default)

Suppose we are interested in predicting the likelihood that loan applicants default on their credit. We have access to historical data on previous loan takers comprised of a binary outcome variable ($y \in \{1 = \text{default}, 2 = \text{no default}\}$) with two input features: (1) the subjects' *age*, which we define as immutable, and (2) the subjects' existing level of *debt*, which we define as mutable.

We have simulated this scenario using synthetic data with independent *age* and *debt* features, and Gaussian class-conditional densities in Figure **??**. The four panels show the outcomes for different training procedures using the same model architectures (a linear classifier). In panels (a) and (c) we have trained the models conventionally, while in panels (b) and (d) we used CT.

In all cases, all counterfactuals (stars) are valid—they have crossed the decision boundary (green)—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 distinguishable cluster. In panel (b), they are highly plausible, meeting the first objective of Def. **??**. In panel (c), 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 **??**. 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 counterfactuals are also plausible with respect to the mutable feature. Thus, we consider the model in panel (d) as the most explainable according to Def. **??**.

Experiments

In our experiments we seek to answer the following three research questions:

[3] For the proof, see the supplementary appendix.

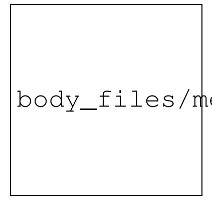

Figure 1: Illustration of how CT improves model explainability: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable. The linear decision boundary is shown in green along with training data colored according to their ground-truth label: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

1. To what extent does the counterfactual training objective as it is defined in Equation **??** induce models to learn plausible explanations?
2. To what extent does the CT objective produce more favorable algorithmic recourse outcomes in the presence of actionability constraints?
3. What are the effects of hyperparameter selection wrt. the CT objective?

Experimental Setup

Our key outcome of interest is improvement in explainability (Def. **??**). To this end, we focus primarily on the plausibility and cost of faithfully generated counterfactuals at test time. To measure the cost, we follow the standard convention of using distances (ℓ_1-norm) between factuals and counterfactuals as a proxy. 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 distance-based metric used in **?**,

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

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 an unbiased estimate of the squared population maximum mean discrepancy (**?**). The metric in Equation **??** is equal to zero iff the two distributions are the same, $\mathbf{X}' = \mathbf{X}^+$.

In addition to cost and plausibility, we compute other standard metrics to evaluate counterfactuals including validity and redundancy. Finally, we also assess the predictive performance of models using standard metrics, including robust accuracy estimated on adversarially perturbed data using FGSM (**?**).

We run the experiments with three gradient-based generators: *Generic* of **?** as a simple baseline approach, *REVISE* (**?**) that aims to generate plausible counterfactuals using a surrogate Variational Autoencoder (VAE), and *ECCo*—the generator of **?** but 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*). These datasets are generated using the library of (**?**) and we present them in the supplementary appendix. Next, we have four real-world binary tabular datasets from the domain of economics: *Adult* (a.k.a. Census data) of (**?**), California housing (*CH*) of (**?**), Default of Credit Card Clients (*Cred*) of (**?**), and Give Me Some Credit (*GMSC*) of (**?**). Finally, for the convenience of illustration, we use of the 10-class *MNIST* vision dataset (**?**).

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 how assessment is done in the related literature (**???**).

Experimental Results

Plausibility Table **??** presents our main empirical findings. For all datasets except *OL* and across all test settings, the average distance of CEs from observed samples in the target class is reduced, indicating improved plausibility. The magnitude of improvements varies. For the simple synthetic datasets, distance reductions range from around 20-40% (*LS*, *Moon*) to almost 60% (*Circ*). For the real-world tabular datasets, improvements tend to be smaller but still substantial, with around 10-15% for *CH*, 11-28% for *GMSC*, 7-8% for *Cred*, and around 3% for *Adult*. For the vision dataset (*MNIST*), distances are reduced by up to 9%. The results for our proposed divergence metric are qualitatively similar, but generally even more pronounced: for the *Circ* dataset, implausibility is reduced by almost 94% to virtually zero as we verified by the absolute outcome. Improvements for other datasets range from 28% (*Moon*) up to 78% (*GMSC*). For *OL* the reduction is negative, consistent with the distance-based metric. *MNIST* is the only dataset for which the distance and divergence metrics disagree. Upon visual inspection of the image counterfactuals we find that CT clearly improves plausibility (see supplementary appendix for images).

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 find that CT improves adversarial robustness on all datasets.

Actionability In Section **??**, we show that our proposed way for encoding mutability constraints leads to lower classifier sensitivity wrt. immutable features for linear models, tilting the decision boundary in favour of mutable features instead. For binding constraints at test time, this leads to shorter counterfactual paths and hence smaller average costs (ℓ_1-norm) to individuals. To extend this to the non-linear case, we test the effect of imposing mutability constraints

Table 1: Key performance metrics across all datasets (column 1). **Plausibility**: Columns 2-6 show the percentage reduction in implausibility (IP) for varying degrees of the energy penalty λ_{egy} used for *ECCo* at test time; column 7 shows the reduction in IP* (MMD), aggregated across all test specifications. **Accuracy** (columns 8-11): test accuracies and robust accuracies (Acc*) for CT and the baseline (BL). **Actionability** (column 12): average reduction in costs when imposing mutability constraints reported for the four datasets for which we could identify meaningful features to protect.

empirically for our synthetic data using the same evaluation scheme as above. The final row in Table **??** reports the average reduction in costs for CT compared to the "vanilla" baseline, when imposing that either the first or the second feature is immutable. In all cases, costs are reduced substantially, indicating that classifiers trained with CT are indeed more sensitive to mutable features.

Impact of hyperparameter settings. We test the impact of three types of hyperparameters; our complete results are in the supplementary appendix.

We note that CT is highly sensitive to the choice of a CE generator and its hyperparameters but (a) there are manageable patterns and (b) we can typically identify settings that improve either plausibility or cost, and commonly 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 (**?**). 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, the choice of $\tau = 0.5$ generally leads to optimal results because it is associated with high proportions of mature counterfactuals.

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.

The effectiveness and stability of CT is positively associated with the number of counterfactuals generated during each training epoch. We also confirm that a higher number of training epochs is beneficial. Interestingly, we observed 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 fine-tuning manner.

Conclusions

As our results indicate, counterfactual training produces models that are more explainable. Nonetheless, it brings about three important limitations.

CT increases the training time of models. CT can be more time-consuming than conventional training regimes. While higher numbers of CEs per iteration positively impact the quality of solutions, they also increase the amount of computations. Relatively small grids with 270 settings can take almost four hours for more demanding datasets on a high-performance computing cluster with 34 2GB CPUs.[4] Three factors attenuate this effect: (1) CT amortizes the cost of CEs for the training samples; (2) it can retain its value when used as a "fine-tuning" technique for conventionally-trained models; and (3) it yields itself to parallel execution, which we have leveraged for our own experiments.

Immutable features may have proxies. We propose an approach to protect immutable features and thus increase the actionability of the generated CEs. However, it requires that model owners define the mutability constraints for (all) 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 enough information about the principals to hinder the protections. Delineating actionability is a major open challenge in the AR literature (see, e.g., (**?**)) impacting the capacity of CT to fulfill its intended goal.

Interventions on features may impact fairness. We provide a tool that allows practitioners to modify the sensitivity of a model with respect to certain features, which may have implication for the fair and equitable treatment of decision subjects. Model owners could misuse this solution by enforcing explanations based on features that are more difficult to modify by some (group of) individuals. For example, consider the *Adult* dataset used in our experiments, where *workclass* or *education* may be more difficult to change for underpriviledged groups. When applied irresponsibly, CT could result in an unfairly assigned burden of recourse (**?**), threatening the equality of opportunity in the system (**?**). Nonetheless, these phenomena are not specific to CT.

We also highlight several 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 **??**. 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 quantified. Focus on classification models is a common restriction in research on CEs and AR. Other settings have attracted some interest (e.g., regression in (**?**)), but there is little consensus how to robustly extend the notion of CEs.

Secondly, our approach is susceptible to training instabilities. This problem has been recognized for JEMs (**?**) and even though we depart from the SGLD-based sampling, we still encounter considerable variability in the outcomes. CT is exposed to two potential sources of instabilities: (1) the energy-based contrastive divergence term in Equation **??**, and (2) the underlying counterfactual 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 for contrastive divergence.

Finally, we believe that it is possible to substantially improve hyperparameter selection procedures. Our method benefits from the tuning of certain key hyperparameters (see Section **??**). In this work, we have relied exclusively on grid search for this task. Future work on CT could benefit from investigating more sophisticated approaches. Notably, CT is iterative which makes methods such as Bayesian or gradient-based optimization applicable (see, e.g., (**?**)).

To conclude, state-of-the-art machine learning models are prone to learning complex representations that cannot be interpreted by humans and existing post-hoc explainability approaches cannot guarantee that the 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 demonstrate that CT satisfies its objective while promoting robustness and preserving the predictive performance of the models. It can also be used to fine-tune conventionally-trained models and achieve similar gains. Finally, this work showcases that it is practical to improve models *and* their explanations at the same time.

[4] See supplementary appendix for computational details.