Counterfactual Training: Teaching Models Plausible and Actionable Explanations

Author information scrubbed for double-blind reviewing

No Institute Given

Abstract. We propose a novel training regime termed Counterfactual Training that leverages counterfactual explanations to increase the explanatory capacity of models. Counterfactual explanations have emerged as a popular post-hoc explanation method for opaque machine learning models: they inform how factual inputs would need to change in order for a model to produce some desired output. To be useful in real-word decision-making systems, counterfactuals should be plausible with respect to the underlying data and actionable with respect to the stakeholder requirements. 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 maintaining high predictive performance.

Keywords: Counterfactual Training · Counterfactual Explanations · Algorithmic Recourse · Explainable AI · Representation Learning

1 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 revolution, [**?**] showed that artificial neural networks (ANN) are sensitive to adversarial examples: 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 (DL) 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 [?] and generalizability [?]. Similarly, [?] showed that state-of-the-art neural networks are so expressive that they can 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 related challenges. Firstly, there is no dependable way to verify if such complex representations correspond to meaningful and 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 [?,?]. 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"). 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** and deemed **plausible** by humans (see Def. 1). 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 also 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 robustness, generalization, and generative capacity [?,?,?]. As further discussed in Section 2, our work consolidates these findings within a single objective.

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

Specifically, we introduce counterfactual training: a novel training regime explicitly meant to align learned representations with plausible explanations that comply with user requirements. Our contributions are as follows:

- We present a novel methodological framework that leverages adversarial examples and faithful counterfactual explanations during the training phase to improve the explanatory capacity and robustness of machine learning models (Section 3).

- We propose a method to enforce global actionability constraints by preventing models from assigning importance to immutable features, i.e., ones over which decision subjects have no control (Section 3).

- Through extensive experiments we demonstrate that counterfactual training promotes explainability while preserving high predictive performance. We run ablation studies and grid searches to understand how the underlying model components and hyperparameters affect outcomes. (Section 4).

Despite some limitations discussed in Section 5, we conclude in Section 6 that counterfactual training provides a useful framework for researchers and practitioners interested in making opaque models more trustworthy. We also believe that this work serves as an opportunity for XAI researchers to re-evaluate the trend of improving XAI tools without improving the underlying models.

2 Related Literature

To the best of our knowledge, the proposed framework for counterfactual training represents the first attempt to use counterfactual explanations during training to improve model explainability. In high-level terms, we define model explainability as the extent to which valid explanations derived for an opaque model are also deemed plausible with respect to the underlying data and stakeholder requirements; the former means that the counterfactuals should comply with the distribution of the factual data, the latter means that they should respect arbitrary (global) actionability constraints. To make the desiderata for our framework more concrete, we follow [**?**] in tying the concept of explainability to the quality of counterfactual explanations that we can generate for a given model. The authors show that CEs—understood here as minimal input perturbations that yield some desired model prediction—are generally 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., adversarial examples). Thus, by inducing models to "unlearn" susceptiblity to such examples, AT can effectively remove implausible explanations from the solution space.

2.1 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. [?] demonstrate that using counterfactual images during classifier training improves model robustness. Similarly, [?] 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, demonstrate that augmenting the training data of image classifiers can improve generalization. Finally, [?] propose 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, counterfactuals have similarly been used to improve models through data augmentation. [?] propose *Polyjuice*, a general-purpose counterfactual generator for language models. They demonstrate empirically that the augmentation of training data through *Polyjuice* counterfactuals improves robustness in a number of NLP tasks. [?] similarly use *Polyjuice* to augment NLP datasets through diverse counterfactuals and show that classifier robustness improves by up to 20%. Finally, [?] introduce Counterfactual Adversarial Training (CAT), which also aims at improving generalization and robustness of language models through a three-step procedure. First, the authors identify training samples that are subject to high predictive uncertainty. Second, they generate counterfactual explanations for those samples. Finally, they fine-tune the given language model on the augmented dataset that includes the generated counterfactuals.

There have also been several attempts at formalizing the relationship between counterfactual explanations and adversarial examples (AE). 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, [?] show that CEs and AEs are equivalent under certain conditions and derive theoretical upper bounds on distances between them.

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, [?] propose 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 (e.g., *age*, *gender*) are immutable. Secondly, [?] are the first to propose an end-to-end training pipeline that includes counterfactual explanations 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 neural network architecture at all.

2.2 Beyond Robustness

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, [?] show 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 by [?] similarly found that generating plausible ("interpretable") counterfactual explanations 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 proposed counterfactual training objective, as we explain in Section 3.

3 Counterfactual Training

Counterfactual training (CT) combines ideas from adversarial training, energy-based modelling and counterfactuals explanations with the explicit goal of aligning representations with plausible explanations that comply with user requirements. In the context of CEs, 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. For AR recommendations to be actionable, they need to not only result in plausible counterfactuals but 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. CT aims to improve model explainability by aligning models with counterfactuals that meet both desiderata: plausibility and actionability. Formally, we define explainability as follows:

Definition 1 (Model Explainability). *Let* $\mathbf{M} \quad X \quad Y$ *denote a supervised classification model that maps from the* D*-dimensional input space* X *to representations* $(\mathbf{x};\ )$ *and finally to the* K*-dimensional output space* Y*. Assume that for any given input-output pair* $\{\mathbf{x},\mathbf{y}\}_i$ *there exists a counterfactual* $\mathbf{x} = \mathbf{x} + \quad \mathbf{M}(\mathbf{x}) = \mathbf{y}^+ \quad \mathbf{y} = \mathbf{M}(\mathbf{x})$ *where* $\arg\max_y \mathbf{y}^+ = y^+$ *and* y^+ *denotes the index of the target class.*

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

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

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

The characterization of faithfulness and plausibility in Def. 1 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. 1 vary and depend on the context in which $\mathbf{M}$ 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 5.

3.1 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, : $S(\mathbf{M}(\mathbf{x}))[y^+]$, where S is the softmax function.[2]

To train models with high explainability as defined in Def. 1, we propose to leverage counterfactuals in the following objective:

$$\begin{aligned} \min \text{yloss}(\mathbf{M}(\mathbf{x}),\mathbf{y}) + {}_{\text{div}}\text{div}(\mathbf{x},\mathbf{x}_T,y;\ ) + {}_{\text{adv}}\text{advloss}(\mathbf{M}(\mathbf{x}_{tT}),\mathbf{y}) \\ + {}_{\text{reg}}\text{ridge}(\mathbf{x},\mathbf{x}_T,y;\ ) \end{aligned} \tag{1}$$

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

where yloss() is a classification loss that induces discriminative performance (e.g., cross-entropy). The second and third terms in Equation 1 are explained in detail below. For now, they can be sufficiently described as inducing explainability directly and indirectly by penalizing: (1) the contrastive divergence, div(), between mature counterfactuals $\mathbf{x}_T$ and observed samples x and, (2) the adversarial loss, advloss(.), with respect to nascent counterfactuals $\mathbf{x}_{tT}$. Finally, ridge() denotes a Ridge penalty ($_2$-norm) that regularizes the magnitude of the energy terms involved in div() [**?**]. The trade-off between the components can be governed by adjusting the strengths of the penalties $_{\mathrm{div}}$, $_{\mathrm{adv}}$ and $_{\mathrm{reg}}$.

3.2 Directly Inducing Explainability with Contrastive Divergence

[**?**] observe that 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 $p(\mathbf{x})$. The authors show that JEMs can be trained to perform well at both tasks by directly maximizing the joint log-likelihood factorized as $\log p(\mathbf{x}, \mathbf{y}) = \log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x})$. The first term can be optimized using conventional cross-entropy as in Equation 1. Then, to optimize $\log p(\mathbf{x})$ [**?**] minimize the contrastive divergence between these observed samples from $p(\mathbf{x})$ and generated samples from $p(\mathbf{x})$.

A key empirical finding in [**?**] was that JEMs tend to do well with respect to the plausibility objective in Def. 1. This follows directly if we consider samples drawn from $p(\mathbf{x})$ as counterfactuals because the JEM objective effectively minimizes the divergence between the conditional posterior and $p(\mathbf{x}|\mathbf{y}^+)$. To generate samples, [**?**] rely 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:

$$\mathrm{div}(\mathbf{x}, \mathbf{x}_T, y;) = E(\mathbf{x}, y) \quad E(\mathbf{x}_T, y) \tag{2}$$

where $E()$ denotes the energy function. We set $E(\mathbf{x}, \mathbf{y}) = \mathbf{M}(\mathbf{x}^+)[y^+]$ where y^+ denotes the index of the randomly drawn target class, $y^+ \quad p(y)$, and $\mathbf{x}^+$ denotes an observed sample from target domain: $\mathbf{X}^+ = \{\mathbf{x} \quad y = y^+\}$. 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 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) [**?**]. As the counterfactuals get more plausible (Def. 1) during training, these opposing effects gradually balance each other out [**?**].

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 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.

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 [**?**,**?**] or AEs [**?**,**?**]. This leaves some flexibility with respect to the choice for advloss() in Equation 1. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

$$
\begin{aligned}
\text{advloss}(\mathbf{M}(\mathbf{x}_{tT}), \mathbf{y};) &= \text{yloss}(\mathbf{M}(\mathbf{x}_{t}), \mathbf{y}) \\
t &= \max_{t}\{t \ \ ||_{t}|| < \}
\end{aligned}
\tag{3}
$$

Under this choice, we consider nascent counterfactuals $\mathbf{x}_{tT}$ as AEs as long as the magnitude of the perturbation to any single feature is at most . This is closely aligned with [**?**] who define an adversarial attack as an "imperceptible non-random perturbation". Thus, we choose to work with a different distinction between CE and AE than [**?**] who consider misclassification as the key distinguishing feature of AE. One of the key observations in this work is that we can leverage CEs during training and get adversarial examples essentially for free.

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 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 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. 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 4. Under certain conditions, this results holds theoretically:[3]

Proposition 1 (Protecting Immutable Features). *Let $f(\mathbf{x}) = S(\mathbf{M}(\mathbf{x})) = S(\mathbf{x})$ denote a linear classifier with softmax activation S where $y \ \ \{1, ..., K\} = K$ and $\mathbf{x} \ \ R^D$. If we assume multivariate Gaussian class densities with common diagonal covariance matrix $_k =$ for all $k \ \ K$, 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 Prp.~1 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 5.

3.5 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 ($y \ \ \{1 = \text{default}, 2 = \text{no default}\}$) 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 two independent features and Gaussian class-conditional densities in Figure 1. The four panels in Figure 1 show the outcomes for different training procedures using the same model architecture each time (a linear classifier). In each case, we show the decision boundary (in green) and the training data colored according to their ground-truth label: orange points belong to the target class, $y^+ = 2$, blue points belong to the non-target class, $y = 1$. Stars indicate counterfactuals in the target class generated at test time using generic gradient descent until convergence.

In panel (a), we have trained our model conventionally, and we do not impose mutability constraints at test time. The generated counterfactuals are all valid, but not plausible: they do not comply with the distribution of the factual samples in the target class to the point where they are clearly distinguishable from the ground-truth data. In panel (b), we have trained our model with counterfactual training, once again without any mutability constraints. We observe that the counterfactuals are highly plausible, meeting the first objective of Def. 1.

In panel (c), we have used conventional training again, this time imposing the mutability constraint on *age* at test time. Counterfactuals are valid but involve some substantial reductions in *debt* for some individuals (very young applicants).

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

By comparison, counterfactual paths are shorter on average in panel (d), where we have used counterfactual training and protected the immutable feature as described in Section 3.4. We observe that due to the classifier's lower sensitivity to *age*, recourse recommendations with respect to *debt* are much more homogenous and do not disproportionately punish younger individuals. The 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. 1.

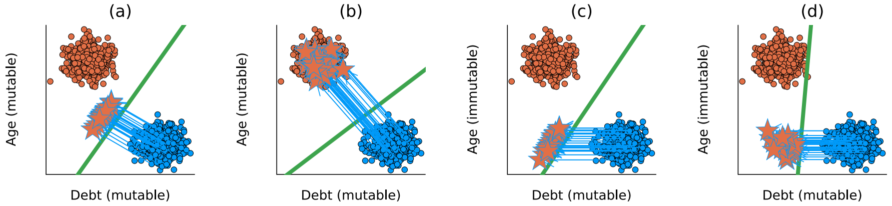

Fig. 1: Illustration of how CT improves model explainability.

4 Experiments

In this section, we present experiments that we have conducted in order to answer the following research questions:

1. To what extent does our proposed counterfactual training objective (Equation 1) induce models to learn plausible explanations?
2. To what extent does our proposed counterfactual training objective (Equation 1) yield more favorable algorithmic recourse outcomes in the presence of actionability constraints?
3. What are the effects of hyperparameter selection with respect to Equation 1?

4.1 Experimental Setup

Evaluation Our key outcome of interest is how well do models perform with respect to explainability (Def. 1). To this end, we focus primarily on the plausibility and cost of faithfully generated counterfactuals at test time. To measure the cost of counterfactuals, we follow the standard convention of using distances ($_1$-norm) between factuals and counterfactuals as a proxy. For plausibility, we assess how similar counterfactuals are to observed samples in the target domain. We rely on the distance-based metric used by [**?**],

$$\mathrm{IP}(\mathbf{x}, \mathbf{X}^{+}) = \frac{1}{|\mathbf{X}^{+}|} {}_{\mathbf{x} \mathbf{X}^{+}} \mathrm{dist}(\mathbf{x}, \mathbf{x}) \tag{4}$$

and introduce a novel divergence metric,

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

where $\mathbf{X}$ denotes a set of multiple counterfactuals and MMD() is an unbiased estimate of the squared population maximum mean discrepancy [?]. The metric in Equation 5 is equal to zero iff the two distributions are the same, $\mathbf{X} = \mathbf{X}^{+}$.

In addition to cost and plausibility, we also compute other standard metrics to evaluate counterfactuals at test time including validity and redundancy. Finally, we also assess the predictive performance of models using standard metrics.

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 CEs.

4.2 Experimental Results

Plausibility Table 1 presents our main empirical findings. The top five rows show the percentage reduction in implausibility according to Equation 4 for varying degrees of the energy penalty used for *ECCo* at test time. The following row shows the reduction in implausibility as measured by Equation 5 and aggregated across all test specifications of *ECCo*. The final two rows show the test accuracies for the model trained with CT and conventionally trained models ("vanilla").

We observe that for all datasets except *OL* and across all test settings, the average distance of counterfactuals from observed samples in the target class is reduced, indicating improved plausibility. The magnitude of improvements varies by dataset: 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 are generally smaller but still substantial in many cases with around 10-15% for *CH*, 11-28% for *GMSC*, 7-8% for *Cred* and around 3% for *Adult*. For our only 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 looking at the absolute outcome. Improvements for other datasets range from 28% (*Moon*) to 78% (*GMSC*). For *OL* the reduction is negative, consistent with the distance-based metric. The only dataset, for which our proposed metric disagrees with the distance-based metric is *MNIST*.

These broad and substantial improvements in plausibility generally do not come at the cost of decreased 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 reduction in test accuracy of 2, 5 and 15 percentage points, respectively. We note in this context, that we have not optimized our models for predictive performance at all and worked with very small networks. In summary, we find that CT can substantially improve the quality of explanations learned by models without generally sacrificing predictive accuracy.

Table 1: Key plausibility and predictive performance metrics for all datasets. The top five rows show the percentage reduction in implausibility according to Equation 4 for varying degrees of the energy penalty used for *ECCo* at test time. The following row shows the reduction in implausibility as measured by Equation 5 and aggregated across all test specifications of *ECCo*. The final two rows show the test accuracies for the model trained with CT and conventionally trained models ("vanilla").

Measure	$_{\text{egy}}$	**Adult**	**CH**	**Circ**	**Cred**	**GMSC**	**LS**	**MNIST**	**Moon**	**OL**
IP (%)	0.1	2.93	9.59	56.5	6.7	11	27.1	9.11	20.4	-6.72
IP (%)	0.5	3.4	9.26	57.1	6.18	13.4	26.7	8.26	21.4	-6.19
IP (%)	1	3.53	10.4	56.5	7.19	13.4	26.6	8.07	21.6	-6.1
IP (%)	5	2.88	11.9	58.5	7.01	21.4	27.1	6.1	19	-2.77
IP (%)	10	3.15	14.6	49.3	7.78	27.9	38.6	3.53	19.8	-1.44
IP (%)	(agg.)	34.8	66.6	93.4	51.6	77.9	54.5	-2.28	27.6	-25.5
Acc. (CT)		0.848	0.794	0.997	0.712	0.608	1	0.902	0.999	0.918
Acc. (vanilla)		0.854	0.85	0.999	0.706	0.751	1	0.922	1	0.914

Actionability

Impact of hyperparameter settings We extensively test the impact of three types of hyperparameters on the proposed training regime. Our complete results are available in the technical appendix; this section focuses on the main findings.

Hyperparameters of the CE generators. First, we observe that CT is highly sensitive to hyperparameter settings 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. Second, we note that the choice of a CE generator has a major impact on the results. For example, *REVISE* tends to perform the worst, most likely because it uses a surrogate VAE to generate counterfactuals which impedes faithfulness [**?**]. Third, increasing T, the maximum number of steps, generally yields better outcomes because more CEs can mature in each training epoch. Fourth, the impact of , the required decision threshold is more difficult to predict. On "harder" datasets it may be difficult to satisfy high for any given sample (i.e., also factuals) and so increasing this threshold does not seem to correlate with better outcomes. In fact, we have generally found that a choice of $= 0.5$ leads to optimal results because it is associated with high proportions of mature counterfactuals.

Hyperparameters for penalties. We find that the strength of the energy regularization, $_{\text{reg}}$ is highly impactful; energy must be sufficiently regularized to avoid poor performance in terms of decreased plausibility and increased costs. The sensitivity with respect to $_{\text{div}}$ and $_{\text{adv}}$ is much less evident. While high values of $_{\text{reg}}$ may increase the variability in outcomes when combined with high values of $_{\text{div}}$ or $_{\text{adv}}$, this effect is not very pronounced.

Other hyperparameters. We observe that 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 find that it is not necessary to employ CT during the entire training phase to achieve the desired improvements in explainability. When training models conventionally during the first 50% of epochs before switching to CT for the next 50% of epochs, we observed positive results. Put differently, CT may be a way to improve the explainability of models in a fine-tuning manner.

5 Discussion

We first address the direct extensions of the counterfactual training approach in Section 5.1. Then, we look at its limitations and challenges in Section 5.2.

5.1 Future research

CT is defined only for 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 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 counterfactuals.

CT is subject to training instabilities. Joint energy-based models are susceptible to instabilities during training [?] and even though we depart from the SGLD-based sampling, we still encounter major variability in the outcomes. CT is exposed to two potential sources of instabilities: (1) the energy-based contrastive divergence term in Equation 2, and (2) the underlying counterfactual explainers. For example, [?] recognize this to be a challenge for *ECCCo* and so it may have downstream impacts on our proposed method. Still, we find that training instabilities can be successfully mitigated by regularizing energy ($_{\text{reg}}$), generating a sufficiently large number of counterfactuals during each training epoch, and including only mature counterfactuals for contrastive divergence.

CT is sensitive to hyperparameter selection. Our method benefits from tuning certain key hyperparameters (see Section 4.2). In this work, we have relied exclusively on grid search for this task. Future work on CT could benefit from investigating more sophisticated approaches towards hyperparameter tuning. Notably, CT is iterative which makes a variety of methods applicable, including Bayesian [?] or gradient-based [?] optimization.

5.2 Limitations and challenges

CT increases the training time of models. Counterfactual training promotes explainability through CEs and robustness through AEs at the cost of longer training times compared to conventional training regimes. While higher numbers of iterations and counterfactuals per iteration positively impact the quality of found solutions, they also increase the required amount of computations. We find that 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]. However, there are three factors that attenuate the impact of this limitation. First, CT provides counterfactual explanations for the training samples essentially for free, which may be beneficial in many ADM systems. Second, we find that CT can retain its value when used as a “fine-tuning” training regime for conventionally-trained models. Third, in principle, CT 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 with sufficient domain knowledge to protect all immutable features, there may exist proxies that are theoretically mutable (and hence should not be protected) but preserve enough information about the principals to hinder the protections. As an example, consider the Adult dataset used in our experiments where the mutable education status is a proxy for the immutable age, in that the attainment of degrees is correlated with age. Delineating actionability is a major undecided challenge in the AR literature [**?**] impacting the capacity of CT to increase the explainability of the model.

Interventions on features may impact fairness downstream. Related to the point above, 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 individuals subject to automated decisions. As protecting a set of features leads the model to assign higher relative importance to unprotected features, model owners could misuse our solution by enforcing explanations based on features that are more difficult to modify by some (group of) individuals. For example, consider again the Adult dataset where features such as workclass or education may be more difficult to change for underpriviledged groups. When applied irresponsibly, counterfactual training could result in an unfairly assigned burden of recourse [**?**], threatening the equality of opportunity in the system [**?**] and potentially reinforcing social segregation [**?**]. Still, as the referenced publications indicate, such phenomena are not specific to CT; all types of ADM solutions without strong external protections have been recognized to promote harmful power dynamics [**?**].

[4] See supplementary appendix for computational details.

6 Conclusion

State-of-the-art machine learning models are prone to learning complex representations that cannot be interpreted by humans. Although post-hoc explainability approaches have attracted major research interest, these cannot guarantee that the explanations agree with the opaque 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 counterfactual training satisfies its objectives while preserving the predictive performance of the trained models. We also find that our approach can be used to fine-tune conventionally-trained models and achieve similar gains in explainability. Finally, this work showcases that it is practical to improve models *and* their explanations at the same time.