Counterfactual Training: Teaching Models Plausible and Actionable Explanations

A Preprint

Patrick Altmeyer
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

p.altmeyer@tudelft.nl

Aleksander Buszydlik
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Arie van Deursen
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Cynthia C. S. Liem
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

March 13, 2025

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.

***K*eywords** 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 (I. Goodfellow, Bengio, and Courville 2016). 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, Szegedy et al. (2013) 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** (I. J. Goodfellow, Shlens, and Szegedy 2014)—truly robust deep learning (DL) 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 provide room for many solutions that are locally optimal with respect to narrow objectives (Wilson 2020).[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 (Frankle and Carbin 2019) and generalizability (Morcos et al. 2019). Similarly, Zhang et al. (2021) 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 (O'Neil 2016; McGregor 2021). 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 Wachter, Mittelstadt, and Russell (2017) 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 Definition 3.1). 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 (Altmeyer et al. 2024). 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 (Schut et al. 2021; Augustin, Meinke, and Hein 2020; Altmeyer et al. 2024). As further discussed in Section 2, our work consolidates these findings within a single objective.

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).

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

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 Augustin, Meinke, and Hein (2020) 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. 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, demonstrate that augmenting the training data of image classifiers can improve generalization. Finally, Teney, Abbasnedjad, and Hengel (2020) 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. Wu et al. (2021) 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. Balashankar et al. (2023) similarly use *Polyjuice* to augment NLP datasets through diverse counterfactuals and show that classifier robustness improves by up to 20%. Finally, Luu and Inoue (2023) 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, 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 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, Ross, Lakkaraju, and Bastani (2024) 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, Guo, Nguyen, and Yadav (2023) 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 Guo, Nguyen, and Yadav (2023), 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, Altmeyer et al. (2024) 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 (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 most called for when models are underspecified by the available data (Wilson 2020). 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 Schut et al. (2021) 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 (Grathwohl et al. 2020). 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 (Poyiadzi et al. 2020; Guidotti 2022; Altmeyer et al. 2024). 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 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 $\arg\max_y \mathbf{y}^+ = y^+$ 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 (i.e., consistent with the data) and actionable. Formally, 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.

We consider counterfactuals as faithful to the extent that they are consistent with what the model has learned about the input data. Let $p_\theta(\mathbf{x}|\mathbf{y}^+)$ denote the conditional posterior over inputs, then formally:

3. (Faithfulness) $\int^A p_\theta(\mathbf{x}'|\mathbf{y}^+)d\mathbf{x} \rightarrow 1$ where A is defined as above.

The characterization of faithfulness and plausibility in Definition 3.1 is the same as in Altmeyer et al. (2024), with adapted notation. Actionability constraints in Definition 3.1 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 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 Wachter, Mittelstadt, and Russell (2017). For our purposes, we let T vary and consider the counterfactual search to be 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 Definition 3.1, we propose to leverage counterfactuals in the following objective:

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

where yloss$(\cdot)$ is a classification loss that induces discriminative performance (e.g., cross-entropy). The second and third terms 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$(\cdot)$, between mature counterfactuals $\mathbf{x}'_T$ and observed samples x 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$(\cdot)$ (Du and Mordatch 2020). The trade-off between the components can be governed by adjusting the strengths of the penalties 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 (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 1. Then, to optimize $\log p_\theta(\mathbf{x})$ Grathwohl et al. (2020) minimize the contrastive divergence between these observed samples from $p(\mathbf{x})$ and generated samples from $p_\theta(\mathbf{x})$.

A key empirical finding in Altmeyer et al. (2024) was that JEMs tend to do well with respect to the plausibility objective in Definition 3.1. 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, Grathwohl et al. (2020) 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:

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

where $\mathcal{E}_\theta(\cdot)$ denotes the energy function. We set $\mathcal{E}_\theta(\mathbf{x}, \mathbf{y}) = -\mathbf{M}_\theta(\mathbf{x}^+)[y^+]$ where y^+ denotes the index of the randomly drawn target class, $y^+ \sim p(y)$, and $\mathbf{x}^+$ denotes an observed sample from target domain: $\mathbf{X}^+ = \{\mathbf{x} : y = y^+\}$. 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 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 (Definition 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 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 (Murphy 2022), 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 (Luu and Inoue 2023; Balashankar et al. 2023) or AEs (Freiesleben 2022; Pawelczyk et al. 2022). This leaves some flexibility with respect to the choice for advloss$(\cdot)$ in Equation 1. An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

[2] For detailed background information on gradient-based counterfactual search and convergence see **?@sec-app-ce**.

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

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. (2013) 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 Freiesleben (2022) 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 (Altmeyer, Deursen, et al. 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. 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 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 $\mathcal{S}$ (i.e.,* multinomial logistic regression*) 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 (Equation 2) 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 3.1 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: influence of counterfactual training on explainability

To better convey the intuition underlying our proposed method, we illustrate different model outcomes in Example 3.1.

Example 3.1 (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 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.

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

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 Definition 3.1.

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). 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 Definition 3.1.

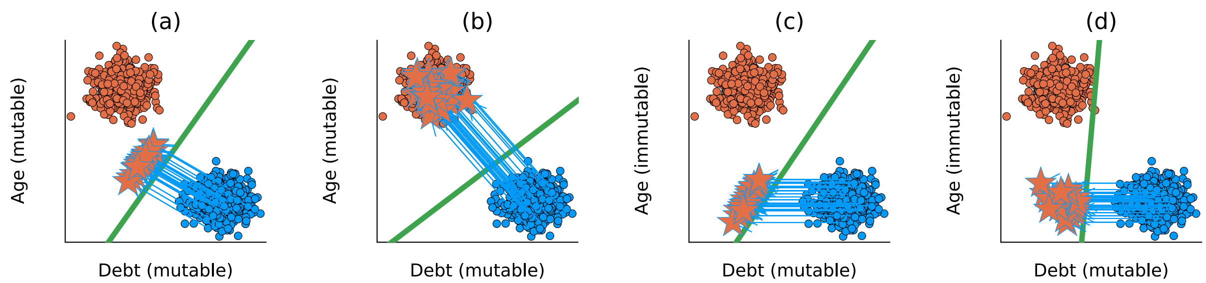

Figure 1: Illustration for Example 3.1 of how CT improves model explainability.

4 Experiments

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

Research Question 4.1 (Plausibility)**.** *To what extent does our proposed counterfactual training objective (Equation 1) induce models to learn plausible explanations?*

Research Question 4.2 (Actionability)**.** *To what extent does our proposed counterfactual training objective (Equation 1) yield more favorable algorithmic recourse outcomes in the presence of actionability constraints?*

Beyond this, we are also interested in understanding how robust our answers to RQ 4.1 and RQ 4.2 are:

Research Question 4.3 (Hyperparameters)**.** *What are the effects of different hyperparameter choices wrt. Equation 1?*

4.1 Experimental Setup

4.1.1 Evaluation

Our key outcome of interest is how well models perform with respect to explainability (Definition 3.1): 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 (ℓ_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 Altmeyer et al. (2024),

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

and introduce a novel divergence metric,

$$\text{implaus}_{\text{div}}(\mathbf{X}', \mathbf{X}^+) = \text{MMD}(\mathbf{X}', \mathbf{X}^+) \tag{5}$$

where $\mathbf{X}'$ denotes a set of multiple counterfactuals and $\text{MMD}(\cdot)$ is an unbiased estimate of the squared population maximum mean discrepancy (Gretton et al. 2012). The metric in Equation 5 is equal to zero iff $\mathbf{X}' = \mathbf{X}^+$.

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 CE 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. (2023) but without the conformal prediction component—as a method that directly targets both faithfulness and plausibility of the CEs.

4.2 Experimental Results

4.2.1 Plausibility

4.2.2 Actionability

4.2.3 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 (Altmeyer et al. 2024). 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 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, we have generally found that a choice of $\tau = 0.5$ 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, λ_{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 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.

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 begin the discussion by addressing the direct extensions of the counterfactual training approach in Section 5.1. Then, we look at its broader 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 by the user, CT requires the output space $\mathcal{Y}$ to be discrete. Thus, it applies to binary and multi-class classification but it is not well-defined for other ML tasks where the change in outcome with respect to a decision threshold τ cannot be readily quantified. In fact, this is a common restriction in research on CEs and AR that predominantly focuses on classification models. Although other settings have attracted some interest (e.g., regression in Spooner et al. 2021; Zhao, Broelemann, and Kasneci 2023), there is still no consensus on what constitutes a counterfactual in such settings.

CT is subject to training instabilities. Joint energy-based models are susceptible to instabilities during training (Grathwohl et al. 2020) 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, Altmeyer et al. (2023) 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 (λ_{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. As discussed in Section 4.2.3, our method benefits from tuning certain key hyperparameters. 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 (e.g., Snoek, Larochelle, and Adams 2012) or gradient-based (e.g., Franceschi et al. 2017) 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 (see details in **?@sec-hardware**). 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. In Proposition 3.1 we define an approach to protect immutable features and thus increase the actionability of the generated counterfactuals. However, this approach 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—ones that cannot be changed at all and ones that cannot be reasonably expected to change—there may exist proxies that are theoretically mutable (and hence should not be protected) but preserve enough information about the principals to counteract the protections. As one 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 (see, e.g., Venkatasubramanian and Alfano 2020) impacting the capacity of CT to increase the explainability of the model.

Interventions on features may have downstream impacts on fairness. 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 (e.g., Sharma, Henderson, and Ghosh 2020), threatening the equality of opportunity in the system (Bell et al. 2024) and potentially reinforcing social segregation (Gao and Lakkaraju 2023). 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 (Maas 2023).

6 Conclusion

References

Abbasnejad, Ehsan, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. "Counterfactual Vision and Language Learning." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10041–51. https://doi.org/10.1109/CVPR42600.2020.01006.

Altmeyer, Patrick, Arie van Deursen, et al. 2023. "Explaining Black-Box Models Through Counterfactuals." In *Proceedings of the JuliaCon Conferences*, 1:130. 1.

Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Faithful Model Explanations Through Energy-Constrained Conformal Counterfactuals." https://arxiv.org/abs/2312.10648.

Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia CS Liem. 2024. "Faithful Model Explanations Through Energy-Constrained Conformal Counterfactuals." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:10829–37. 10.

Augustin, Maximilian, Alexander Meinke, and Matthias Hein. 2020. "Adversarial Robustness on in-and Out-Distribution Improves Explainability." In *European Conference on Computer Vision*, 228–45. Springer.

Balashankar, Ananth, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. 2023. "Improving Classifier Robustness Through Active Generative Counterfactual Data Augmentation." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 127–39.

Bell, Andrew, Joao Fonseca, Carlo Abrate, Francesco Bonchi, and Julia Stoyanovich. 2024. "Fairness in Algorithmic Recourse Through the Lens of Substantive Equality of Opportunity." https://arxiv.org/abs/2401.16088.

Du, Yilun, and Igor Mordatch. 2020. "Implicit Generation and Generalization in Energy-Based Models." https://arxiv.org/abs/1903.08689.

Franceschi, Luca, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. 2017. "Forward and Reverse Gradient-Based Hyperparameter Optimization." In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:1165–73. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v70/franceschi17a.html.

Frankle, Jonathan, and Michael Carbin. 2019. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJl-b3RcF7.

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

Gao, Ruijiang, and Himabindu Lakkaraju. 2023. "On the Impact of Algorithmic Recourse on Social Segregation." In *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." https://arxiv.org/abs/1412.6572.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. "Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One." In *International Conference on Learning Representations*.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. "A Kernel Two-Sample Test." *The Journal of Machine Learning Research* 13 (1): 723–73.

Guidotti, Riccardo. 2022. "Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking." *Data Mining and Knowledge Discovery*, 1–55.

Guo, Hangzhi, Thanh H. Nguyen, and Amulya Yadav. 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*, 577–89. KDD '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3580305.3599290.

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." https://arxiv.org/abs/1907.09615.

Kolter, Zico. 2023."Keynote Addresses: SaTML 2023 ." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, xvi–. Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/SaTML54575.2023.00009.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *Advances in Neural Information Processing Systems* 30.

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

Luu, Hoai Linh, and Naoya Inoue. 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*, 881–88.

Maas, Jonne. 2023. "Machine Learning and Power Relations." *AI & SOCIETY* 38 (4): 1493–1500.

McGregor, Sean. 2021. "Preventing repeated real world AI failures by cataloging incidents: The AI incident database." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:15458–63. 17.

Morcos, Ari S., Haonan Yu, Michela Paganini, and Yuandong Tian. 2019. "One Ticket to Win Them All: Generalizing Lottery Ticket Initializations Across Datasets and Optimizers." In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.

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

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

Pawelczyk, Martin, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. "Exploring Counterfactual Explanations Through the Lens of Adversarial Examples: A Theoretical and Empirical Analysis." In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, edited by Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, 151:4574–94. Proceedings of Machine Learning Research. PMLR. https://proceedings.mlr.press/v151/pawelczyk22a.html.

Poyiadzi, Rafael, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. "FACE: Feasible and Actionable Counterfactual Explanations." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–50.

Ross, Alexis, Himabindu Lakkaraju, and Osbert Bastani. 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.

Sauer, Axel, and Andreas Geiger. 2021. "Counterfactual Generative Networks." https://arxiv.org/abs/2101.06046.

Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. "Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties." In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. 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*, 166–72. AIES '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375812.

Snoek, Jasper, Hugo Larochelle, and Ryan P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms." In *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Spooner, Thomas, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. 2021. "Counterfactual Explanations for Arbitrary Regression Models." *CoRR* abs/2106.15212. https://arxiv.org/abs/2106.15212.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. "Intriguing Properties of Neural Networks." https://arxiv.org/abs/1312.6199.

Teney, Damien, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. "Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part x 16*, 580–99. Springer.

Venkatasubramanian, Suresh, and Mark Alfano. 2020. "The Philosophical Basis of Algorithmic Recourse." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–93. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372876.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harv. JL & Tech.* 31: 841. https://doi.org/10.2139/ssrn.3063289.

Wilson, Andrew Gordon. 2020. "The Case for Bayesian Deep Learning." https://arxiv.org/abs/2001.10995.

Wu, Tongshuang, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 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)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, 6707–23. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.523.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Commun. ACM* 64 (3): 107–15. https://doi.org/10.1145/3446776.

Zhao, Xuan, Klaus Broelemann, and Gjergji Kasneci. 2023. "Counterfactual Explanation for Regression via Disentanglement in Latent Space." In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 976–84. Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/ICDMW60847.2023.00130.