Counterfactual Training: Teaching Models Plausible and Actionable Explanations

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***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-world decision-making systems, counterfactuals should be plausible with respect to the underlying data and actionable with respect to the feature mutability constraints. Much existing research has therefore focused on developing post-hoc methods to generate counterfactuals that meet these desiderata. In this work, we instead hold models directly accountable for the desired end goal: counterfactual training employs counterfactuals during the training phase to minimize the divergence between learned representations and plausible, actionable explanations. We demonstrate empirically and theoretically that our proposed method facilitates training models that deliver inherently desirable counterfactual explanations and additionally exhibit improved adversarial robustness.**

***Index Terms*—explainable AI, representation learning, contrastive learning, adversarial machine learning**

I. Introduction

Today's prominence of artificial intelligence (AI) has largely been driven by the success of representation learning with high degrees of freedom: instead of relying on features and rules hand-crafted by humans, modern machine learning (ML) models are tasked with learning highly complex representations directly from the data, guided by narrow objectives such as predictive accuracy [1]. These models tend to be so complex that humans cannot easily interpret their decision logic.

Counterfactual explanations (CE) have become a key part of the broader explainable AI (XAI) toolkit [2] that can be applied to make sense of this complexity. Originally proposed in [3], CEs prescribe minimal changes for factual inputs that, if implemented, would prompt some fitted model to produce an alternative, more desirable output. This is useful and necessary to not only understand how opaque models make their predictions, but also to provide algorithmic recourse to individuals subjected to them: a retail bank, for example, could use CE to provide meaningful feedback to unsuccessful loan applicants that were rejected based on an opaque automated decision-making (ADM) system (Fig. 1).

For such feedback to be meaningful, counterfactual explanations need to fulfill certain desiderata [4], [5]—they should be faithful to the model [6], plausible [7], and actionable [8]. Plausibility is typically understood as counterfactuals being *in-domain*: unsuccessful loan applicants that implement the provided recourse should end up with credit profiles that are genuinely similar to that of individuals who have successfully repaid their loans in the past. Actionable explanations further comply with practical constraints: a young, unsuccessful loan applicant cannot increase their age, in an instance.

Existing state-of-the-art (SOTA) approaches in the field have largely focused on designing model-agnostic CE methods that identify subsets of counterfactuals, which comply with specific desiderata. This is problematic, because the narrow focus on any specific desideratum can adversely affect others: it is possible, for example, to generate plausible counterfactuals for models that are also highly vulnerable to implausible, possibly adversarial counterfactuals [6]. Indeed, existing approaches generally fail to guarantee that the representations learned by a model are compatible with truly meaningful explanations.

In this work, we propose an approach to bridge this gap, embracing the paradigm that models (as opposed to explanation methods) should be held accountable for explanations that are plausible and actionable. While previous work has shown that at least plausibility can be indirectly achieved through existing techniques aimed at models' generative capacity, generalization and robustness [6], [9], [10], we directly incorporate both plausibility and actionability in the training objective of models to improve their overall explanatory capacity.

Specifically, we introduce **counterfactual training (CT)**: a novel training regime that leverages counterfactual expla-

nations on-the-fly to ensure that differentiable models learn plausible and actionable explanations for the underlying data, while at the same time being more robust to adversarial examples (AE). Fig. 1 illustrates the outcomes of CT compared to a conventionally trained model. First, in panel (a), faithful and valid counterfactuals end up near the decision boundary forming a clearly distinguishable cluster in the target class (orange). In panel (b), CT is applied to the same underlying linear classifier architecture resulting in much more plausible counterfactuals. In panel (c), the classifier is again trained conventionally and we have introduced a mutability constraint on the *age* feature at test time—counterfactuals are valid but the classifier is roughly equally sensitive to both features. By contrast, the decision boundary in panel (d) has tilted, making the model trained with CT relatively less sensitive to the immutable *age* feature. To achieve these outcomes, CT draws inspiration from the literature on contrastive and robust learning: we contrast faithful CEs with ground-truth data while protecting immutable features, and capitalize on methodological links between CE and AE by penalizing the model's adversarial loss on interim (*nascent*) counterfactuals. To the best of our knowledge, CT represents the first venture in this direction with promising empirical and theoretical results.

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

II. Related Literature

To make the desiderata for CT more concrete, we follow previous work, tying the explanatory capacity of models to the quality of CEs that can be generated for them [6], [9].

A. Explanatory Capacity and Contrastive Learning

A closely related work, [6], shows that model averaging and, in particular, contrastive model objectives can produce models that have a higher explanatory capacity, and hence ones that are more trustworthy. The authors propose a way to generate counterfactuals that are maximally faithful in that they are consistent with what models have learned about the underlying data. Formally, they rely on tools from energy-based modelling [11] to minimize the contrastive divergence between the distribution of counterfactuals and the conditional posterior over inputs learned by a model. Their algorithm, *ECCCo*, yields plausible counterfactual explanations if and only if the underlying model has learned representations that align with them. The authors find that both deep ensembles [12] and joint energy-based models (JEMs) [13], a form of contrastive learning, do well in this regard.

It helps to look at these findings through the lens of representation learning with high degrees of freedom. Deep ensembles are approximate Bayesian model averages, which are particularly effective when models are underspecified by the available data [14]. Averaging across solutions mitigates the risk of overrelying on a single locally optimal representation that corresponds to semantically meaningless explanations. Likewise, [10] demonstrates that generating plausible ("interpretable") CEs is almost trivial for deep ensembles that have undergone adversarial training. The case for JEMs is even clearer: they optimize a hybrid objective that induces both high predictive performance and strong generative capacity [13], resembling the idea of aligning models with plausible explanations. This was an inspiration for CT.

B. Explanatory Capacity and Robust Learning

The authors of [9] show that counterfactual explanations tend to be more meaningful ("explainable") if the underlying model is more robust to adversarial examples. Once again, we can make intuitive sense of this finding if we look at adversarial training (AT) through the lens of representation learning with high degrees of freedom: highly complex and flexible models may learn representations that make them sensitive to implausible or even adversarial examples [15]. Thus, by inducing models to "unlearn" susceptibility to such examples, adversarial training can effectively remove implausible explanations from the solution space.

This interpretation of the link between explanatory capacity through counterfactuals on the one side, and robustness to adversarial examples on the other is backed by empirical evidence. Firstly, [16] demonstrates that using counterfactual images during classifier training improves model robustness. Similarly, [17] argues that counterfactuals represent potentially useful training data in machine learning tasks, especially in supervised settings where inputs may be reasonably mapped to multiple outputs. They, too, show that augmenting the training data of (image) classifiers can improve generalization performance. Finally, [18] argues that counterfactual pairs tend to exist in training data. Hence, the proposed approach aims to identify similar inputs with different annotations and ensure that the gradient of the classifier aligns with the vector between such pairs of inputs using a cosine distance loss function.

CEs have also been used to improve models in the natural language processing domain. For example, [19] proposes *Polyjuice*, a general-purpose CE generator for language models, and demonstrates that the augmentation of training data with *Polyjuice* improves robustness in a number of tasks. Also, [20] introduces the *Counterfactual Adversarial Training* (CAT) framework that aims to improve generalization and robustness of language models by generating counterfactuals for training samples that are subject to high predictive uncertainty.

There have also been several attempts at formalizing the relationship between counterfactual explanations and adversarial examples. Pointing to clear similarities in how CEs and

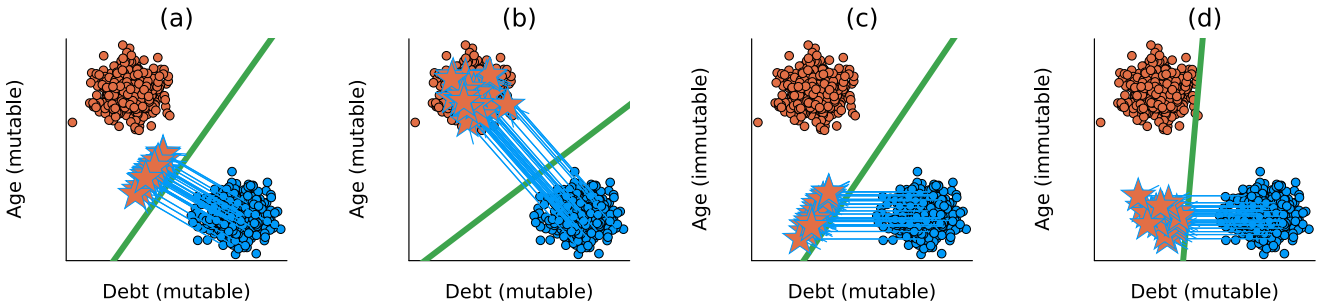

Fig. 1. Counterfactual explanations (stars) for linear classifiers trained under different regimes on synthetic data: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable. The linear decision boundary is shown in green along with training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

AEs are generated, [21] 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 [21]. Similarly, [22] shows that CEs and AEs are equivalent under certain conditions.

Two other works are closely related to ours in that they use counterfactuals during training with the explicit goal of affecting certain properties of the post-hoc counterfactual explanations. Firstly, [23] proposes a way to train models that guarantee recourse to a positive target class with high probability. The approach builds on adversarial training by explicitly inducing susceptibility to targeted AEs for the positive class. Additionally, the method allows for imposing a set of actionability constraints ex-ante. For example, users can specify that certain features are immutable. Secondly, [24] is the first to propose an end-to-end training pipeline that includes CEs as part of the training procedure; the *CounterNet* network architecture includes a predictor and a CE generator, where the parameters of the CE generator are learnable. Counterfactuals are generated during each training iteration and fed back to the predictor. In contrast, we impose no restrictions on the ANN architecture at all.

III. Counterfactual Training

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

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

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

where $\mathbf{M}_\theta : \mathcal{X} \mapsto \mathcal{Y}$ denotes a classifier, $\mathbf{x}'$ denotes the counterfactual with D features and $\mathbf{y}^+ \in \mathcal{Y}$ denotes some target class. The yloss$(\cdot)$ function quantifies the discrepancy between current model predictions for $\mathbf{x}'$ and the target class (a conventional choice is cross-entropy). Finally, we use reg$(\cdot)$ to denote any form of regularization used to induce certain properties on the counterfactual. The seminal CE paper, [3], proposes regularizing the distance between counterfactuals and their original factual values to ensure that individuals seeking recourse through CE face minimal costs in terms of feature changes. Different variations of equation (1) have been proposed in the literature to address many desiderata including the ones discussed above (faithfulness, plausibility and actionability). As in [3], most of these approaches rely on gradient descent to optimize equation (1), and this holds true for all approaches tested in this work. We introduce them briefly in Section IV-A, but refer the reader to the supplementary appendix for details. In the following, we describe how counterfactuals are generated and used in CT.

A. Proposed Training Objective

The goal of CT is to improve the explanatory capacity of models by aligning the learned representations with faithful explanations that are plausible and actionable. For simplicity, we refer to models with high explanatory capacity as **explainable** in this manuscript. We define explainability as follows:

Definition III.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 $\mathcal{Y}$. Let $\mathbf{x}'_0$ denote a factual input and assume that for any given input-output pair $\{\mathbf{x}'_0, \mathbf{y}\}_i$ there exists a counterfactual $\mathbf{x}' = \mathbf{x}'_0 + \Delta : \mathbf{M}_\theta(\mathbf{x}') = \mathbf{y}^+ \neq \mathbf{y} = \mathbf{M}_\theta(\mathbf{x})$, where $\arg\max_y \mathbf{y}^+ = y^+$ is the index of the target class.

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

- (Faithfulness) $P(\mathbf{x}' \in \mathcal{X}_\theta | \mathbf{y}^+) = 1 - \delta$, where δ is some small value, and $\mathcal{X}_\theta | \mathbf{y}^+$ is the conditional posterior distribution over inputs (adapted from [6], Def. 4.1).
- (Plausibility) $P(\mathbf{x}' \in \mathcal{X} | \mathbf{y}^+) = 1 - \delta$, where δ is some small value, and $\mathcal{X} | \mathbf{y}^+$ is the conditional distribution of inputs in the target class (adapted from [6], Def. 2.1).
- (Actionability) Perturbations Δ may be subject to some actionability constraints.

Intuitively, plausible counterfactuals are consistent with the data, and faithful counterfactuals are consistent with what

the model has learned about the input data. Actionability constraints in Def. III.1 depend on the context in which $\mathbf{M}_\theta$ is deployed (e.g., specified by end-users or model owners). We consider two types of actionability constraints: on the domain of features and on their mutability. The former naturally arise in automated decision-making systems whenever a feature can only take a specific range of values. For example, *age* is lower bounded by zero and upper bounded by the maximum human lifespan. Specifying such domain constraints can also help address training instabilities commonly associated with energy-based modelling [13]. The latter arise when a feature cannot be freely modified. Continuing the example, *age* of a person can only increase, but it may even be considered as an immutable feature: waiting many years for an improved outcome is hardly feasible for individuals affected by algorithmic decisions. We choose to only consider domain and mutability constraints for individual features x_d for $d = 1, ..., D$. Of course, this is a simplification since feature values may correlate, e.g., higher *age* may be associated with higher *level of completed education*. We address this challenge in Section V, where we also explain why we restrict this work to classification settings.

Let $\mathbf{x}'_t$ for $t = 0, ..., T$ denote a counterfactual generated through gradient descent over T iterations as originally proposed in [3]. CT adopts gradient-based CE search in training to generate on-the-fly model explanations $\mathbf{x}'$ for the training samples. We use the term *nascent* to denote interim counterfactuals $\mathbf{x}'_{\text{AE}}$ that have not yet converged. As we explain below, these nascent counterfactuals can be stored and repurposed as adversarial examples. Conversely, we consider counterfactuals $\mathbf{x}'_{\text{CE}}$ as *mature* explanations if they have converged within the T iterations by reaching a pre-specified threshold, τ, for the predicted probability of the target class: $\mathcal{S}(\mathbf{M}_\theta(\mathbf{x}'))[y^+] \geq \tau$, where $\mathcal{S}$ is the softmax function.

Formally, we propose the following counterfactual training objective to train explainable (as in Def. III.1) models,

$$\begin{aligned} &\min_\theta \text{yloss}(\mathbf{M}_\theta(\mathbf{x}), \mathbf{y}) + \lambda_{\text{div}}\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y^+; \theta) \\ &+\lambda_{\text{adv}}\text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{\text{AE}}), \mathbf{y}_{\text{AE}}) + \lambda_{\text{reg}}\text{ridge}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta) \end{aligned} \tag{2}$$

where yloss$(\cdot)$ is any classification loss that induces discriminative performance (e.g., cross-entropy). The second and third terms are explained in detail in the following subsections. For now, they can be summarized as inducing explainability directly and indirectly by penalizing (1) the contrastive divergence, div$(\cdot)$, between mature counterfactuals $\mathbf{x}'_{\text{CE}}$ and observed samples $\mathbf{x}^+ \in \mathcal{X}^+ = \{\mathbf{x} : y = y^+\}$ in the target class y^+, and (2) the adversarial loss, advloss$(.)$, wrt. nascent counterfactuals $\mathbf{x}'_{\text{AE}}$ and their corresponding labels $\mathbf{y}_{\text{AE}}$. Finally, ridge$(\cdot)$ denotes a Ridge penalty (squared ℓ_2-norm) that regularizes the magnitude of the energy terms involved in the contrastive divergence, div$(\cdot)$, term [25]:

$$\frac{1}{n_{\text{CE}}} \sum_{i=1}^{n_{\text{CE}}} \left(\mathcal{E}_\theta(\mathbf{x}^+, y^+)^2 + \mathcal{E}_\theta(\mathbf{x}'_{\text{CE}}, y^+)^2 \right) \tag{3}$$

The trade-offs between these components are adjusted through penalties λ_{div}, λ_{adv}, and λ_{reg}.

The full counterfactual training regime is sketched out in Fig. 2. During each iteration, we do the following steps. Firstly, we randomly draw a subset of $n_{\text{CE}} \leq n$ factuals $\mathbf{x}'_0$ from $\mathbf{X}$ of size n, for which we uniformly draw a target class y^+ (ensuring that it does not coincide with the class currently predicted for $\mathbf{x}'_0$) and a corresponding training sample from the target class, $\mathbf{x}^+ \sim \mathbf{X}^+ = \{\mathbf{x} \in \mathbf{X} : y = y^+\}$. Secondly, we conduct the counterfactual search by solving (1) through gradient descent. Thirdly, we sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from the training data set $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ for conventional training and distribute the tuples composed of counterfactuals, their target labels and corresponding training samples, as well as adversarial examples and corresponding labels, $(\mathbf{x}'_{\text{CE}i}, {y^+}_i, \mathbf{x}'_{\text{AE}i}, \mathbf{y}_{\text{AE}i}, {\mathbf{x}^+}_i)_{i=1}^{n_{\text{CE}}}$, across the mini-batches. Finally, we backpropagate through (2).

Require: Training dataset $\mathcal{D}$, initialize model $\mathbf{M}_\theta$
while not converged **do**
  Sample $\mathbf{x}'_0 \sim \mathbf{X}$, $y^+ \sim \mathcal{U}(\mathcal{Y})$ and $\mathbf{x}^+ \sim \mathbf{X}^+$.
  for $t = 1$ to T **do**
    Backpropagate $\nabla_{\mathbf{x}'}$ through equation (1). Store $\mathbf{x}'_{\text{CE}}, \mathbf{x}'_{\text{AE}}, \mathbf{y}_{\text{AE}}$.
  end for
  Sample mini-batches $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{n_b}$ from dataset $\mathcal{D}$.
  Distribute $(\mathbf{x}'_{\text{CE}i}, {y^+}_i, \mathbf{x}'_{\text{AE}i}, \mathbf{y}_{\text{AE}i}, {\mathbf{x}^+}_i)_{i=1}^{n_{\text{CE}}}$.
  for each batch **do**
    Backpropagate ∇_θ through equation (2).
  end for
end while
return $\mathbf{M}_\theta$

Fig. 2. Pseudo-Code for Counterfactual Training

By limiting ourselves to a subset of n_{CE} counterfactuals, we reduce runtimes; this approach has previously been shown to improve efficiency in the context of adversarial training [26], [27]. To improve runtimes even more, we choose to first generate counterfactuals and then distribute them across mini-batches to benefit from greater degrees of parallelization during the counterfactual search. Alternatively, it is possible to generate counterfactuals separately for each mini-batch.[1]

B. Directly Inducing Explainability: Contrastive Divergence

As observed in [13], any classifier can be re-interpreted as a joint energy-based model that learns to discriminate output classes conditional on the observed (training) samples from $p(\mathbf{x})$ and the generated samples from $p_\theta(\mathbf{x})$. The authors show that JEMs can be trained to perform well at both tasks by directly maximizing the joint log-likelihood: $\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x})$, where the first term can be optimized using cross-entropy as in equation (2). To optimize $\log p_\theta(\mathbf{x})$, they minimize the contrastive divergence between the observed samples from $p(\mathbf{x})$ and samples generated from $p_\theta(\mathbf{x})$.

[1]During initial prototyping of CT we also tested an implementation that relies on generating counterfactuals and adversarial examples at the batch level with no discernible difference in outcomes, but increased training times.

To generate samples, [13] suggests Stochastic Gradient Langevin Dynamics (SGLD) with an uninformative prior for initialization but we depart from this methodology: we propose to leverage counterfactual explainers to generate counterfactuals of observed training samples. Specifically, we have:

$$\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y^+; \theta) = \mathcal{E}_\theta(\mathbf{x}^+, y^+) - \mathcal{E}_\theta(\mathbf{x}'_{\text{CE}}, y^+) \quad (4)$$

where $\mathcal{E}_\theta(\cdot)$ denotes the energy function defined as $\mathcal{E}_\theta(\mathbf{x}, y^+) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$, with y^+ denoting the index of the randomly drawn target class, $y^+ \sim p(y)$. Conditional on the target class y^+, $\mathbf{x}'_{\text{CE}}$ denotes a mature counterfactual for a randomly sampled factual from a non-target class generated with a gradient-based CE generator for up to T iterations. Intuitively, the gradient of equation (4) decreases the energy of observed training samples (positive samples) while increasing the energy of counterfactuals (negative samples) [25]. As the counterfactuals get more plausible (Def. III.1) during training, these opposing effects gradually balance each other out [28].

Since the maturity of counterfactuals in terms of a probability threshold is often reached before T, this form of sampling is not only more closely aligned with Def. III.1., but can also speed up training times compared to SGLD. The departure from SGLD also allows us to tap into the vast repertoire of explainers that have been proposed in the literature to meet different desiderata. For example, many methods support domain and mutability constraints. In principle, any approach for generating CEs is viable, so long as it does not violate the faithfulness condition. Like JEMs [29], counterfactual training can be viewed as a form of contrastive representation learning.

C. Indirectly Inducing Explainability: Adversarial Robustness

Based on our analysis in Section II, counterfactuals $\mathbf{x}'$ can be repurposed as additional training samples [20], [30] or adversarial examples [21], [22]. This leaves some flexibility with regards to the choice for the advloss$(\cdot)$ term in equation (2). An intuitive functional form, but likely not the only sensible choice, is inspired by adversarial training:

$$\begin{aligned} \text{advloss}(\mathbf{M}_\theta(\mathbf{x}'_{\text{AE}}), \mathbf{y}; \varepsilon) &= \text{yloss}(\mathbf{M}_\theta(\mathbf{x}'_{t_\varepsilon}), \mathbf{y}) \\ t_\varepsilon &= \max_t \{t : ||\Delta_t||_\infty < \varepsilon\} \end{aligned} \quad (5)$$

Under this choice, we consider nascent counterfactuals $\mathbf{x}'_{\text{AE}}$ as AEs as long as the magnitude of the perturbation at time t (Δ_t) to any single feature is at most ε. The most strongly perturbed counterfactual $\mathbf{x}'_{t_\varepsilon}$ that still satisfies the condition is used as an adversarial example $\mathbf{x}'_{\text{AE}}$. This formalization is closely aligned with [15] who define an adversarial attack as an "imperceptible non-random perturbation". Thus, we work with a different distinction between CE and AE than [21], which considers misclassification as the distinguishing feature of adversarial examples. One of the key observations of our work is that we can leverage CEs during training and get AEs essentially for free to reap the benefits of adversarial training, leading to improved adversarial robustness and plausibility.

D. Encoding Actionability Constraints

Many existing counterfactual explainers support domain and mutability constraints. In fact, both types of constraints can be implemented for any explainer that relies on gradient descent in the feature space for optimization [31]. In this context, domain constraints can be imposed by simply projecting counterfactuals back to the specified domain; if the previous gradient step resulted in updated feature values that were out-of-domain. Similarly, mutability constraints can be enforced by setting partial derivatives to zero to ensure that features are only perturbed in the allowed direction, if at all.

As actionability constraints are binding at test time, we must 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 (4). It is crucial that we avoid penalizing implausibility that arises from 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 defined for d. Specifically, we set $\mathbf{x}^+[d] := \mathbf{x}'[d]$ if:

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

From a Bayesian perspective, setting $\mathbf{x}^+[d] := \mathbf{x}'[d]$ can be understood as assuming a point mass prior for $p(\mathbf{x}^+)$ with respect to feature d, i.e., we can model this as absolute certainty that the value $\mathbf{x}^+[d]$ remains the same as in the neighbor, $\mathbf{x}'[d]$, but it could be equivalently seen as masking changes to feature d. Intuitively, we can think of this as ignoring implausibility costs of immutable features, which effectively forces the model to instead seek plausibility through the remaining features. This can be expected to produce a classifier with relatively lower sensitivity to immutable features, and the higher relative sensitivity to mutable features should make mutability-constrained recourse less costly (see Section IV). Under certain conditions, this result also holds theoretically (for the proof, see the supplementary appendix):

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

IV. Experiments

We start by introducing the experimental setup, including performance metrics, datasets, algorithms, and explain our approach to evaluation in Section IV-A. Then, we address

the research questions. Two questions relating to the principal goals of counterfactual training are presented in Section IV-B:

(RQ1) To what extent does the CT objective in equation (2) induce models to learn plausible explanations?

(RQ2) To what extent does CT result in more favorable algorithmic recourse outcomes in the presence of actionability constraints

Next, in Section IV-C we consider the performance of models trained with CT, focusing on their adversarial robustness but also commenting on the validity of generated CEs.

(RQ3) To what extent does CT influence the adversarial robustness of trained models?

Finally, in Section IV-D we perform an ablation of the CT objective and evaluate its sensitivity to hyperparameters:

(RQ4) How does the CT objective depends on its individual components? (*ablation*)

(RQ5) What are the effects of hyperparameter selection on counterfactual training?

A. Experimental Setup

Our focus is the improvement in explainability (Def. III.1). Thus, we mainly look at the plausibility and cost of faithfully generated counterfactuals at test time, but several other metrics are covered in the supplementary appendix. To measure the cost, we follow the standard proxy of distances (ℓ_1-norm) between factuals and counterfactuals. For plausibility, we assess how similar CEs are to observed samples in the target domain, $\mathbf{X}^+ \subset \mathcal{X}^+$. For the evaluation, we rely on the metric used by [6] with ℓ_1-norm for distances,

$$\text{IP}(\mathbf{x}', \mathbf{X}^+) = \frac{1}{|\mathbf{X}^+|} \sum_{\mathbf{x} \in \mathbf{X}^+} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{6}$$

and introduce a novel divergence-based adaptation,

$$\text{IP}^*(\mathbf{X}', \mathbf{X}^+) = \text{MMD}(\mathbf{X}', \mathbf{X}^+) \tag{7}$$

where $\mathbf{X}'$ denotes a collection of counterfactuals and $\text{MMD}(\cdot)$ is the unbiased estimate of the squared population maximum mean discrepancy, proposed in [32]:

$$\begin{aligned} \text{MMD}(\mathbf{X}', \mathbf{X}^+) = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(\tilde{x}_i, \tilde{x}_j) \\ &- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, \tilde{x}_j) \end{aligned} \tag{8}$$

with a kernel function $k(\cdot, \cdot)$. We use a characteristic Gaussian kernel with a constant length-scale parameter of 0.5, which means that the metric in equation (7) is equal to zero if and only if the two distributions are exactly the same, $\mathbf{X}' = \mathbf{X}^+$.

To assess outcomes with respect to actionability for non-linear models, we look at the costs of (just) valid counterfactuals in terms of their distances from factual starting points with $\tau = 0.5$. While this is an imperfect proxy of sensitivity, we hypothesize that CT can reduce these costs by teaching models to seek plausibility with respect to mutable features, much like we observe in Fig. 1 in panel (d) compared to (c). We supplement this analysis with estimates using integrated gradients (IG) [33]. To evaluate predictive performance, we use standard metrics, such as robust accuracy estimated on adversarially perturbed data using the fast gradient sign method (FGSM) [34] and projected gradient descent (PGD) [35].

We make use of nine classification datasets common in the CE/AR literature. Four of them are synthetic with two classes and different characteristics: linearly separable Gaussian clusters (*LS*), overlapping clusters (*OL*), concentric circles (*Circ*), and interlocking moons (*Moon*). Next, we have four real-world binary tabular datasets: *Adult* (Census data) of [36], California housing (*CH*) of [37], Default of Credit Card Clients (*Cred*) of [38], and Give Me Some Credit (*GMSC*) from [39]. Finally, for convenient illustration, we use the 10-class *MNIST* [40].

We run experiments with three gradient-based generators: *Generic* of [3] as a simple baseline; *REVISE* [7] that aims to generate plausible counterfactuals using a surrogate Variational Autoencoder (VAE); and *ECCCo* [6], targeting faithfulness. In all cases, we use standard logit cross-entropy loss for $\text{yloss}(\cdot)$ and all generators penalize the distance (ℓ_1-norm) of counterfactuals from their original factual state. *Generic* and *ECCCo* search for counterfactuals directly in the feature space; *REVISE* traverses the latent space of a variational autoencoder (VAE) fitted to the training data, so its outputs depend on the quality of the surrogate model. In addition to the distance penalty, *ECCCo* uses a penalty that regularizes the energy associated with the counterfactual, $\mathbf{x}'$ [6]. We omit the conformal set size penalty proposed in the original paper, since the authors found that faithfulness primarily depends on the energy penalty, freeing us from one additional hyperparameter.

Our method does not aim to be agnostic to the underlying CE generator and, as explained in Section III-B, the selection of the CE generator can impact the explainability of models. To evaluate the specific value of counterfactual training, we extensively test the method using the three above-mentioned CE generators, which are characterized by varying complexity and desiderata, and we present the complete results in the supplementary appendix. Indeed, we observe that *ECCCo* outclasses the other two generators as the backbone of CT, generally leading to the highest reduction in implausibility. This is not surprising; the goals of *ECCCo* most closely align with the objectives of CT: maximally faithful explanations should also be the most useful for feedback. Conversely, we cannot expect the model to learn much from counterfactual explanations that largely depend on the quality of the surrogate model that is trained for *REVISE*. Similarly, *Generic* is a very simple baseline that optimizes only for minimal changes of features (measured in [3] using median absolute deviation).

Thus, while counterfactual training can be used with any gradient-based CE generator to improve the explainability of the resulting model, in Section IV-B we mainly discuss its effectiveness with *ECCCo*, the strongest identified generator,

allowing us to optimize the quality of the models. This constitutes our treatment method, but we still present the complete results for all generators in the supplementary appendix.

To assess the effects of CT, we investigate the improvements in performance metrics when using it on top of a weak baseline (BL), a naively (conventionally) trained multilayer perceptron (*MLP*), as the control method. As we hold all other things constant, this is the best way to get a clear picture of the improvement in explainability that can be directly attributed to CT. It is also consistent with the evaluation practices in the related literature [18], [23], [34].

We also note that counterfactual training involves multiple objectives but our principal goal is high explainability as in Def. III.1, while improved robustness is a welcome byproduct. We neither aim to outperform state-of-the-art approaches that target any single one of these objectives, nor do we claim that CT can achieve this. Specifically, we do not aim to beat JEMs with respect to their generative capacity, SOTA robust neural networks with respect to (adversarial) robustness, or (quasi-)Bayesian neural networks with respect to uncertainty quantification. As we have already explained in Section II, existing literature has shown that all of these objectives tend to correlate (explaining some of our positive findings), but we situate counterfactual training squarely in the context of (counterfactual) explainability and algorithmic recourse, where it tackles an important shortcoming of existing approaches.

In terms of computing resources, all of our experiments were executed on a high-performance cluster. We have relied on distributed computing across multiple central processing units (CPU); for example, the hyperparameter grid searches were carried out on 34 CPUs with 2GB memory each. Graphical processing units (GPU) were *not* used. All computations were performed in the Julia Programming Language [41]; our codebase (algorithms and experimental settings) has been anonymized and is available to reviewers.[2] We explain more about the hardware, software, and reproducibility considerations in the supplementary appendix. Details will be disclosed upon publication to avoid revealing information that might interfere with the double-blind review process.

B. Main Results

Our main results for plausibility and actionability for *MLP* models are summarised in Table I that presents counterfactual outcomes grouped by dataset along with standard errors averaged across bootstrap samples. Asterisks (*) are used when the bootstrapped 99%-confidence interval of differences in mean outcomes does *not* include zero, so the observed effects are statistically significant at the 0.01 level. As our experimental procedure is (by virtue of the proposed method) relatively complex, we choose to work at this stringent alpha level to demonstrate the high reliability of counterfactual training.

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

[2]https://anonymous.4open.science/r/CounterfactualTraining/README.md.

TABLE I

KEY EVALUATION METRICS FOR VALID COUNTERFACTUAL ALONG WITH BOOTSTRAPPED STANDARD ERRORS FOR ALL DATASETS. **PLAUSIBILITY** (COLUMNS 1-2): PERCENTAGE REDUCTION IN IMPLAUSIBILITY FOR IP AND IP*, RESPECTIVELY; **COST / ACTIONABILITY** (COLUMN 3): PERCENTAGE REDUCTION IN COSTS WHEN SELECTED FEATURES ARE PROTECTED. OUTCOMES ARE AGGREGATED ACROSS BOOTSTRAP SAMPLES (100 ROUNDS) AND VARYING DEGREES OF THE ENERGY PENALTY λ_{EGY} USED FOR ECCCO AT TEST TIME. ASTERISKS (*) INDICATE THAT THE BOOTSTRAPPED 99%-CONFIDENCE INTERVAL OF DIFFERENCES IN MEAN OUTCOMES DOES **NOT** INCLUDE ZERO.

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

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

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

Average reductions in IP range from around 7% for *MNIST* to almost 60% for *Circ*. For the real-world tabular datasets they are around 12% for *CH* and *Cred* and almost 25% for *GMSC*; for *Adult* and *OL* we find no significant impact of CT on IP. The former is subject to a large proportion of categorical features, which inhibits the generation of large numbers of valid counterfactuals during training and may therefore explain this finding.

Reductions in IP* are even more substantial and generally statistically significant, although the average degree of uncertainty is higher than for IP: reductions range from around 20% (*Moon*) to almost 90% (*Circ*). The only negative findings are for *OL* and *MNIST*, but they are insignificant. A qualitative inspection of the counterfactuals in Fig. 3 suggests recognizable digits for the model trained with CT (bottom row), unlike the baseline (top row).

Actionability (RQ2): CT tends to improve actionability in the presence of immutable features, but this is not guaranteed if the assumptions in Proposition III.1 are violated.

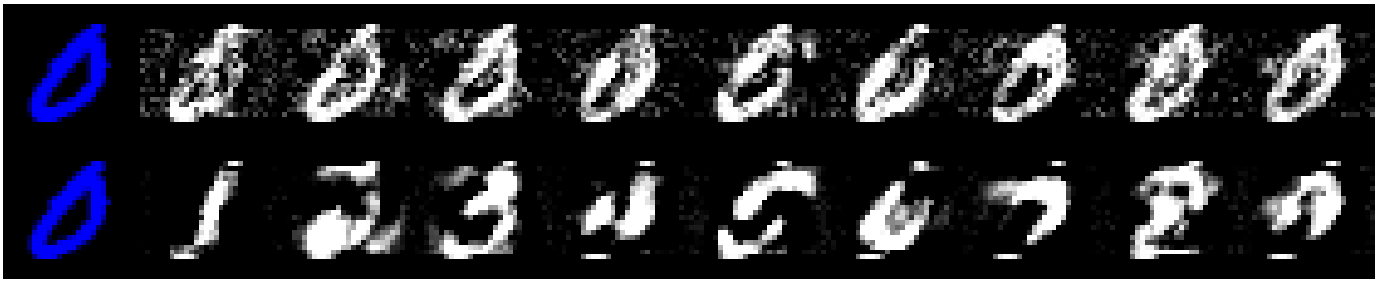

Fig. 3. *Plausibility*: BL (top row) vs CT using the *ECCCo* generator (bottom row) counterfactuals for a randomly selected factual from class "0" (in blue). CT produces more plausible counterfactuals than BL.

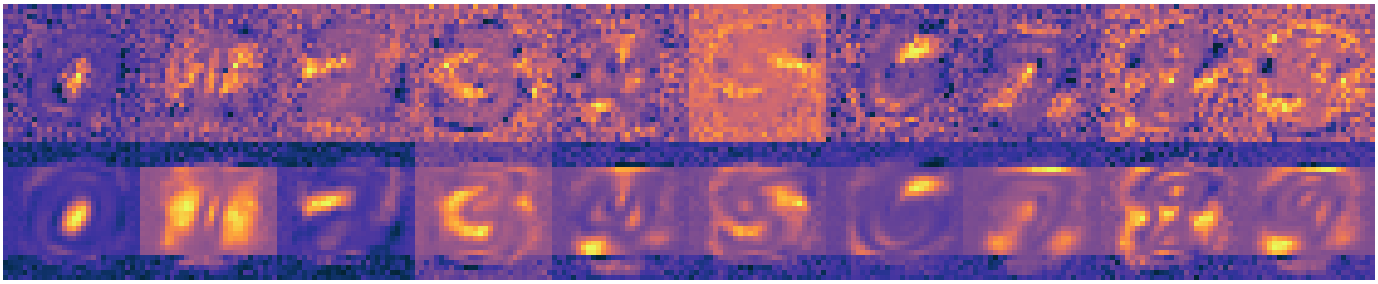

Fig. 4. Sample visual explanations for all classes in the *MNIST* dataset. Top and bottom rows of images show the results for BL and CT, respectively. Mutability constraints are imposed on the five top and five bottom rows of pixels. CT is less sensitive to protected features.

For synthetic datasets, we always protect the first feature; for all real-world tabular datasets we could identify and protect an *age* variable; for *MNIST*, we protect the five top and five bottom rows of pixels of the full image. Statistically significant reductions in costs overwhelmingly point in the positive direction reaching up to around 60% for *GMSC* data. Only in the case of *Cred*, average costs increase, most likely because any benefits from protecting *age* are outweighed by an increase in costs required for greater plausibility. The findings for *Adult* and *MNIST* are insignificant.

To empirically evaluate the feature protection mechanism of CT beyond linear models covered in Proposition III.1, we make use of integrated gradients (IG) as proposed in [33]. IG calculates the contribution of each input feature towards a specific prediction by approximating the integral of the model output with respect to its input, using a set of samples that linearly interpolate between a test instance and some baseline instance. This process produces a vector of real numbers, one per input feature, which informs about the contribution of each feature to the prediction. The selection of an appropriate baseline is an important design decision [33]; to remain consistent in our evaluations, we use a baseline drawn at random from the uniform distribution $\mathcal{U}(-1, 1)$ for all datasets, which aligns with standard evaluation practices for IG. As the outputs are not bounded (i.e., they are real numbers), we standardize the integrated gradients across features to allow for a meaningful comparison of the results for different models.

Qualitatively, the class-conditional integrated gradients in Fig. 4 suggest that CT has the expected effect even for non-linear models: the model trained with CT (bottom row) is less sensitive (blue) to the five top and five bottom rows of pixels that were protected. Quantitatively, we observe substantial improvements for seven out of nine datasets, and inconclusive results for the remaining two datasets. Table II shows the average sensitivity for protected features as per standardized integrated gradients for CT and BL: for the synthetic datasets we observe strong reductions in sensitivity to the protected features for *LS*, *OL* and *OL*, in line with expectations. The only dataset negatively impacted by CT—albeit with highly variable results—is *Moon*; in this case, the underlying data clearly violates the assumptions in Proposition 3.1, but we did observe above that costs for recourse are still reduced when the constrained feature is protected. For the real-world datasets, the reductions in sensitivity to the protected *age* variable are over two-fold for *Adult* and *CH* and almost three-fold for protected pixels in *MNIST*, mirroring the qualitative findings in Fig. 4. In case of *Cred* we fully prevent the CT model from considering *age* as a factor in classification, with sensitivity reduced to zero. Only for *GMSC* we observe negative impacts of CT, which we believe is due to any or all of the following: a) data assumptions are violated; b) the impact of other components of the CT objective outweighs expected effects of feature protection; or c) the baseline choice applied consistently to all data sets is not appropriate for *GMSC*.

C. Predictive Performance

Adversarial Robustness (RQ3): Models trained with CT are much more robust to gradient-based adversarial attacks than conventionally-trained (weak) baselines.

Test accuracies on clean and adversarially perturbed test data are shown in Fig. 5. The perturbation size, $\varepsilon \in [0, 0.1]$, increases along the horizontal axis, where the case of $\varepsilon = 0$

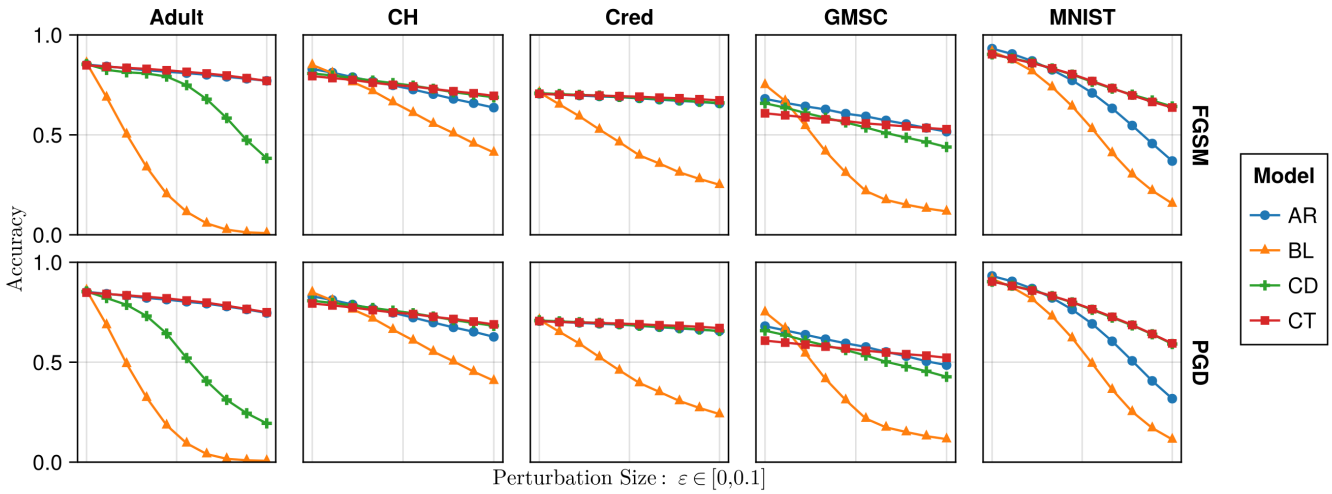

Fig. 5. Test accuracies on adversarially perturbed data with varying perturbation sizes for the non-synthetic data sets. Different training objectives are distinguished by color and shape: (1) BL–the weak baseline; (2) CT–the full CT objective; (3) AR–a partial CT objective without contrastive divergence; (4) CD–a partial CT objective without adversarial loss. Top and bottom rows show the results for FGSM and PGD (40 steps at step size $\eta = 0.01$), respectively.

TABLE II
AVERAGE SENSITIVITY ± BOOTSTRAPPED STANDARD ERRORS OF PROTECTED FEATURES AS PER STANDARDIZED INTEGRATED GRADIENTS.

Data	CT	BL
LS	0.03	10.24 ± 2.40
Circ	3.20 ± 0.67	149.76±842.75
Moon	60.84 ± 128.51	0.55 ± 0.06
OL	0.78 ± 0.12	4.81 ± 1.08
Adult	0.43 ± 0.01	1.0
CH	0.08 ± 0.01	0.23 ± 0.01
Cred	0.0	0.43 ± 0.01
GMSC	1.0	0.21 ± 0.03
MNIST	0.18 ± 0.01	0.41 ± 0.01

TABLE III
AVERAGE VALIDITY OF COUNTERFACTUALS FOR CT VS BL. FIRST TWO COLUMNS CORRESPOND TO NO MUTABILITY CONSTRAINTS IMPOSED ON THE FEATURES; LAST TWO COLUMNS INVOLVE MUTABILITY CONSTRAINTS IMPOSED ON THE SPECIFIED FEATURES.

Data	CT mut.	BL mut.	CT constr.	BL constr.
LS	1.0	1.0	1.0	1.0
Circ	0.97	0.52	0.67	0.49
Moon	1.0	1.0	0.99	0.98
OL	0.87	0.98	0.37	0.57
Adult	0.61	0.99	0.56	0.99
CH	0.96	1.0	0.96	1.0
Cred	0.7	1.0	0.67	1.0
GMSC	0.63	1.0	0.38	1.0
MNIST	1.0	1.0	1.0	1.0
Avg.	0.86	0.94	0.73	0.89

corresponds to standard test accuracy for non-perturbed data. For synthetic datasets, predictive performance is virtually unaffected by perturbations for all models; those results are therefore omitted from Fig. 5 in favor of better illustrations for the real-world data.

We find that standard test accuracy is largely unaffected by CT, while robustness against both types of attacks (FGSM and PGD) is greatly improved: while in some cases robust accuracies for the weak baseline drop to virtually zero (worse than random guessing) for large enough perturbation sizes, accuracies of CT models remain remarkably robust, even though robustness is not the primary objective of counterfactual training. In the only case where standard accuracy on unperturbed test data is substantially reduced for CT (*GSMC*), we note that robust accuracy decreases particularly fast for the weak baseline as the perturbation size increases. This seems to indicate that the standard accuracy for the weak baseline is inflated by sensitivity to meaningless associations in the data.

We also look at the validity of generated counterfactuals, or the proportion of counterfactuals that attain the target class, as presented in Table III. We find that in many cases CT leads to substantial reductions in average validity, but this effect does not seem to be influenced by the imposed mutability constraints (columns 1-2 vs columns 3-4). This result does not surprise us: by design, CT shrinks the solution space for valid counterfactual explanations, thus making it "harder" (and yet not "more costly") to reach validity compared to the baseline model. As further discussed in the supplementary appendix, this should not be seen as a shortcoming of the method for a number of reasons: validity rates can be increased with longer searches; costs of found solutions still generally decrease, as we observe in our experiments; and achieving high validity does not entail that explanations are practical for the recipients (e.g., valid solutions may still be extremely costly) [42].

D. Ablation and Hyperparameter settings

In this subsection, we use ablation studies to investigate how the different components of the counterfactual training objective in equation (2) affect outcomes. Beyond this, we also interested in understanding how CT depends on various other hyperparameters. To this end, we present the results from extensive grid searches run across all synthetic datasets.

Ablation (RQ4): All components of the CT objective affect outcomes, even independently, but the full objective achieves the most consistent improvements wrt. our goals.

We ablate the effect of both (1) the contrastive divergence component and (2) the adversarial loss included in the full CT objective in equation (2). In the following, we refer to the resulting partial objectives as adversarial robustness (AR) and contrastive divergence (CD), respectively. We note that AR corresponds to a form of adversarial training and the CD objective is similar to that of a joint energy-based model. Therefore, the ablation also serves as a comparison of counterfactual training to stronger baselines, although we emphasize again that we do not seek to outperform SOTA methods in the domains of generative or robust machine learning, focusing CT squarely on models with high explainability and actionability in the context of algorithmic recourse.

Firstly, we find that both components play an important role in shaping final outcomes. Both AR and CD can independently improve the plausibility and adversarial robustness of models.

Concerning plausibility, Fig. 6 shows the percentage reductions in implausibility for the partial and full objectives compared to the weak baseline. The results for IP and IP* are shown in the top and bottom graphs, respectively, and the data sets are differentiated by color. We find that in the best identified hyperparameter settings, the final results for the full objective are predominantly affected by the contrastive divergence component, but the inclusion of adversarial loss leads to additional improvements for some data sets (*Adult*, *GMSC*). We penalize contrastive divergence twice as strongly as adversarial loss, which may explain why this component dominates. The outcome for *Adult*, in particular, demonstrates the benefit of including both components: as noted earlier, the large proportion of categorical features in this data set seems to inhibit the generation of valid counterfactuals, which in turn appears to diminish the effect of the contrastive divergence component.

Looking at AR alone, we find that it produces mixed results, with strong positive results nonetheless dominating, reflecting previous findings from the related literature. In particular, for real-world tabular datasets, adversarial robustness seems to substantially benefit plausibility. In these cases, the inclusion of the component in the full objective also helps to substantially improve outcomes in relation to the partial CD objective: improvements in plausibility for the *Adult* and *GMSC* datasets are notably higher for full CT. In summary, the full CT objective leads to the most consistent improvements with respect to plausibility.

Regarding adversarial robustness, we also find that the full CT objective outperforms the partial objectives, which both independently yield improvements. Consistent with the existing literature on JEMs [13], CD yields substantially more robust models than the weak baseline at varying perturbation sizes (Fig. 5). Similarly, AR yields consistent improvements in robustness, as expected. Still, we observe that in cases where either CD or AR show signs of degrading robust accuracy at higher perturbation sizes, the full CT objective maintains robustness. Much like in the context of plausibility, full CT benefits from both components, highlighting the effectiveness of our approach to reusing nascent counterfactuals as AEs.

Hyperparameter settings (RQ5): CT is quite sensitive to the choice of a CE generator and its hyperparameters but (1) we observe manageable patterns, and (2) we can usually identify settings that improve either plausibility or actionability, and typically both of them at the same time.

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

Firstly, we find that optimal results are generally obtained when using *ECCCo* to generate counterfactuals. Conversely, using a generator that may inhibit faithfulness (*REVISE*), regularly yields smaller improvements in plausibility and is more likely to even increase implausibility. The results of the grid search for *REVISE* also exhibit higher variability than the results for *ECCCo* and *Generic*. As argued above, this finding confirms our intuition that maximally faithful explanations are most suitable for counterfactual training.

Concerning hyperparameters that guide the gradient-based counterfactual search, we find that increasing T, the maximum number of steps, generally yields better outcomes because more CEs can mature. Relatedly, we also find that the effectiveness and stability of CT is positively associated with the total number of counterfactuals generated during each training epoch. The impact of τ, the decision threshold, is more difficult to predict. On "harder" datasets it may be difficult to satisfy high τ for any given sample (i.e., also factuals) and so increasing this threshold does not seem to correlate with better outcomes. In fact, $\tau = 0.5$ generally leads to optimal results as it is associated with high proportions of mature counterfactuals. This is likely because the special case of $\tau = 0.5$ corresponds to equal class probabilities, so a counterfactual is considered mature when the logit for the target class is higher than the logits for all other classes.

Secondly, the strength of the energy regularization, λ_{reg}, is highly impactful and should be set sufficiently high to avoid common problems associated with exploding gradients. The sensitivity with respect to λ_{div} and λ_{adv} is much less evident. While high values of λ_{reg} may increase the variability in outcomes when combined with high values of λ_{div} or λ_{adv}, this effect is not particularly pronounced. These results mirror our observations from the ablation studies and lend further weight to the argument that CT benefits from both components.

Finally, we also observe desired improvements when CT was combined with conventional training and employed only for the final 50% of epochs of the complete training process.

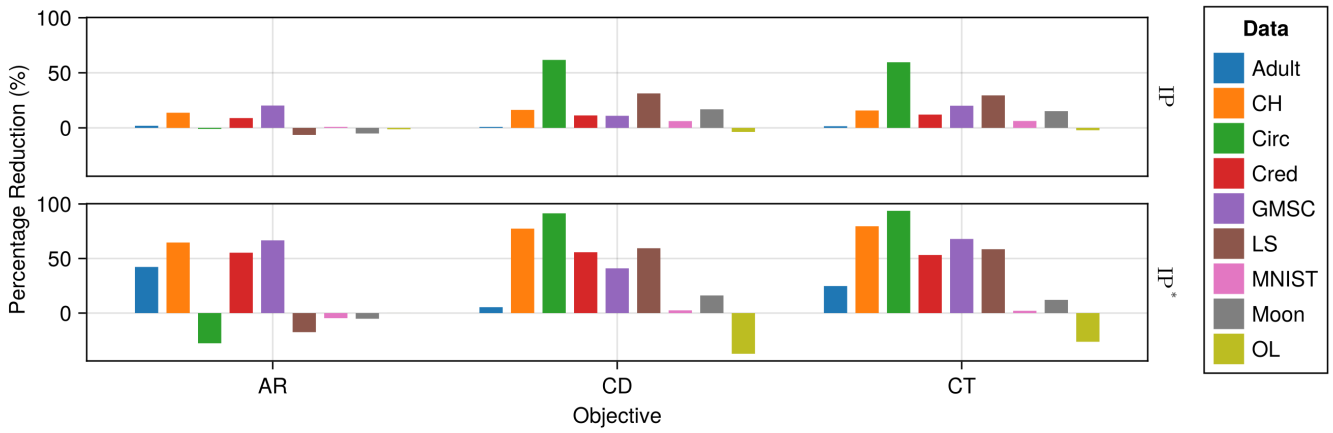

Fig. 6. Percentage reductions in implausibility for the partial (AR, CD) and full (CT) objectives compared to the weak baseline. The results for IP and IP* are shown in the top and bottom graphs, respectively, and the data sets are differentiated by color.

Put differently, CT can improve the explainability of models in a post-hoc, fine-tuning manner.

V. Discussion

As our results indicate, counterfactual training achieves its objective of producing models that are more explainable. Nonetheless, these advantages come with certain limitations.

Immutable features may have proxies. We propose a method to modify the sensitivity of a model to certain features, and thus increase the actionability of the generated CEs. However, it requires that model owners define the mutability constraints for (all) features considered by the model. Even if all immutable features are protected, there may exist proxies that are theoretically mutable (and hence should not be protected) but preserve enough information about the principals to hinder these protections. Delineating actionability is a major open challenge in the AR literature (see, e.g., [42]) impacting the capacity of CT to fulfill its intended goal.

Interventions on features may have implications for fairness. Modifying the sensitivity of a model to certain features may also have implications for the fair and equitable treatment of decision subjects. Model owners could misuse this solution by enforcing explanations based on features that are more difficult to modify by some (group of) decision subjects. For example, consider the *Adult* dataset used in our experiments, where *workclass* or *education* may be more difficult to change for underprivileged groups. When applied irresponsibly, CT could result in an unfairly assigned burden of recourse [43], threatening the equality of opportunity in the system [44]. Nonetheless, these phenomena are not specific to CT.

Plausibility is costly. As noted by [6], more plausible counterfactuals are inevitably more costly. CT improves plausibility and robustness, but this can negatively affect average costs and validity whenever cheap, implausible, and adversarial explanations are removed from the solution space.

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

We also highlight three key directions for future research. Firstly, it is an interesting challenge to extend CT beyond classification settings. Our formulation relies on the distinction between target and non-target classes, requiring the output space to be discrete. Thus, it does not apply to ML tasks where the change in outcome cannot be readily discretized. Classification remains the focus of CE and algorithmic recourse research; other settings have attracted some interest (e.g., regression [45]), but there is little consensus on how to extend the notion of CEs.

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

Finally, we believe that it is possible to considerably improve hyperparameter selection procedures. Our method benefits from the tuning of certain key hyperparameters but we have relied exclusively on grid searches. Future work on CT could benefit from more sophisticated approaches. Notably, CT is iterative, which makes methods such as Bayesian or gradient-based optimization applicable (see, e.g., [46]).

VI. Conclusion

State-of-the-art machine learning models are prone to learning complex representations that cannot be interpreted by humans. Existing work on counterfactual explanations has largely focused on designing tools to generate plausible and actionable explanations for any model. In this work, we instead hold models accountable for delivering such explanations. We introduce counterfactual training: a novel training regime that integrates recent advances in contrastive learning, adversarial robustness, and CE to incentivize highly explainable mod-

els. Through theoretical results and extensive experiments, we demonstrate that CT satisfies this goal while promoting adversarial robustness of models. Explanations generated from CT-based models are both more plausible (compliant with the underlying data-generating process) and more actionable (compliant with user-specified mutability constraints), and thus meaningful to recipients. In turn, our work highlights the value of simultaneously improving models and their explanations.

REFERENCES

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

[2] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Christoph Molnar, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[3] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[4] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," 2022.

[5] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: definitions, formulations, solutions, and prospects," 2021.

[6] P. Altmeyer, M. Farmanbar, A. van Deursen, and C. C. S. Liem, "Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 10 829–10 837.

[7] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," 2019, arXiv:1907.09615.

[8] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[9] M. Augustin, A. Meinke, and M. Hein, "Adversarial robustness on in- and out-distribution improves explainability," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer, 2020, pp. 228–245.

[10] L. Schut, O. Key, R. McGrath, L. Costabello, B. Sacaleanu, Y. Gal *et al.*, "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1756–1764.

[11] Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton, "Energy-based models for sparse overcomplete representations," *J. Mach. Learn. Res.*, vol. 4, no. null, pp. 1235–1260, Dec. 2003.

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

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

[14] A. G. Wilson, "The case for bayesian deep learning," 2020, arXiv:2001.10995.

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

[16] A. Sauer and A. Geiger, "Counterfactual generative networks," 2021, arXiv:2101.06046.

[17] E. Abbasnejad, D. Teney, A. Parvaneh, J. Shi, and A. van den Hengel, "Counterfactual vision and language learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 041–10 051.

[18] D. Teney, E. Abbasnedjad, and A. van den Hengel, "Learning what makes a difference from counterfactual examples and gradient supervision," in *Computer Vision - ECCV 2020*. Berlin, Heidelberg: Springer-Verlag, 2020, pp. 580–599.

[19] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "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)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: ACL, Aug. 2021, pp. 6707–6723.

[20] H. L. Luu and N. Inoue, "Counterfactual adversarial training for improving robustness of pre-trained language models," in *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. ACL, 2023, pp. 881–888. [Online]. Available: https://aclanthology.org/2023.paclic-1.88/

[21] T. Freiesleben, "The intriguing relation between counterfactual explanations and adversarial examples," *Minds and Machines*, vol. 32, no. 1, pp. 77–109, 2022.

[22] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, and H. Lakkaraju, "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*, ser. Proceedings of Machine Learning Research, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., vol. 151. PMLR, 28–30 Mar 2022, pp. 4574–4594. [Online]. Available: https://proceedings.mlr.press/v151/pawelczyk22a.html

[23] A. Ross, H. Lakkaraju, and O. Bastani, "Learning models for actionable recourse," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2024.

[24] H. Guo, T. H. Nguyen, and A. Yadav, "CounterNet: End-to-end training of prediction aware counterfactual explanations," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 577—589.

[25] Y. Du and I. Mordatch, "Implicit generation and generalization in energy-based models," 2020, arXiv:1903.08689.

[26] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2017. [Online]. Available: https://arxiv.org/abs/1611.01236

[27] M. Kaufmann, Y. Zhao, I. Shumailov, R. Mullins, and N. Papernot, "Efficient adversarial training with data pruning," *arXiv preprint arXiv:2207.00694*, 2022.

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

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

[30] A. Balashankar, X. Wang, Y. Qin, B. Packer, N. Thain, E. Chi, J. Chen, and A. Beutel, "Improving classifier robustness through active generative counterfactual data augmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*. ACL, 2023, pp. 127–139.

[31] P. Altmeyer, A. van Deursen, and C. C. S. Liem, "Explaining black-box models through counterfactuals," in *Proceedings of the JuliaCon Conferences*, vol. 1, 2023, p. 130.

[32] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017. [Online]. Available: https://arxiv.org/abs/1703.01365

[34] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015, arXiv:1412.6572.

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

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

[37] R. K. Pace and R. Barry, "Sparse spatial autoregressions," *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997.

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

[39] Kaggle, "Give me some credit, Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years." https://www.kaggle.com/c/GiveMeSomeCredit, 2011, accessed: 2023-12-14. [Online]. Available: https://www.kaggle.com/c/GiveMeSomeCredit

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

[41] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM review*, vol. 59, no. 1, pp. 65–98, 2017. [Online]. Available: https://doi.org/10.1137/141000671
[42] S. Venkatasubramanian and M. Alfano, "The philosophical basis of algorithmic recourse," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 284–293.
[43] S. Sharma, J. Henderson, and J. Ghosh, "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*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 166–172.
[44] A. Bell, J. Fonseca, C. Abrate, F. Bonchi, and J. Stoyanovich, "Fairness in algorithmic recourse through the lens of substantive equality of opportunity," 2024, arXiv:2401.16088.
[45] T. Spooner, D. Dervovic, J. Long, J. Shepard, J. Chen, and D. Magazzeni, "Counterfactual explanations for arbitrary regression models," 2021, arXiv:2106.15212.
[46] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, and M. Lindauer, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1484, 2023.

APPENDIX A
SUPPLEMENTARY APPENDIX AND CODE

Due to its length, we make the supplementary appendix available to reviewers as a separate document, instead of including it below in addition to the previous reviews we have received. Specifically, the appendix can be found in the anonymised code repository at `paper/preprint/appendix.pdf`[3].

APPENDIX B
PREVIOUS REVIEWS

Below we include reviews we received recently for this work. Please note that the first review (xuqy) evidently does not discuss our paper but some other work. We contacted the Ethics Chair of the other venue as soon as we noticed this, but unfortunately we have not received a response so far. For the sake of completeness, we still include the review (xuqy) here, but it can be safely ignored. We also want to point out the final review listed below was AI-generated and we have clearly marked it as such.

We begin by summarizing the changes we made to the paper following the reviews and then list each review verbatim.

A. Revisions

In response to the reviews we received, we have:

- Added an ablation study for the final experiments. We previously had reported grid searches for the penalty strengths in the appendix, but we agree with the reviewer bsoG and AI review that a full ablation in the paper helps.
- Added PGD on top of FGSM to evaluate adversarial robustness (suggestion from AI review).
- Addressed all notation inconsistencies pointed out by the reviewers and the AI review that we deemed relevant.
- Adjusted our narrative to avoid confusion about our objectives and claims (reviewer Nb3w). Specifically, we emphasize that we situate our work squarely in the context of explainability and actionability and that improved adversarial robustness is a welcome byproduct. Furthermore, we highlight that CT inherently targets multiple objectives and that we therefore neither aim nor claim to beat SOTA methods that target any one of the objectives (generative capacity, adversarial robustness, ...).
- Made it clearer now in the main paper that we do indeed test different counterfactual generators (reviewer Nb3w) and report detailed results in the appendix. However, also our goal is not to propose a generator-agnostic training regime, but rather to ensure that models learn from explanations that are as faithful as possible.
- Since we had some additional space for this submission, we have also moved and/or summarized parts of the appendix, which (1) we deemed relevant or (2) previous reviewers seem to have missed, into the main paper.
- Fairness analysis (AI review): we added more detail here that we had previously removed due to space constraints.

B. Reviews

1) Review by xuqy:

a) Summary: The paper addresses the problem of ensuring robustness in neural network verification by introducing a novel verification framework. The authors propose a method that combines symbolic reasoning with approximation strategies to compute certified robustness bounds for deep networks under various perturbation models. The work includes theoretical guarantees, algorithmic innovations, and experiments on benchmark datasets such as MNIST, CIFAR-10, and ImageNet-scale networks. Results suggest that the proposed approach achieves tighter bounds and faster runtimes compared to existing baselines. Supplementary material provides additional proofs, algorithmic details, and extended experimental results.

b) Evaluation: Quality: The technical quality is strong, with clear formal definitions, theoretical results, and well-structured proofs in the supplementary material. Experiments cover standard datasets and compare against several state-of-the-art verifiers. However, while the improvements are noticeable, they are not consistently dominant across all tasks, and some baselines appear underexplored.

Clarity: The paper is generally well-written, with a logical flow and clear problem statement. Figures (e.g., runtime vs. bound tightness plots) help illustrate improvements. However, some notation-heavy sections (e.g., the proofs in the appendix) are difficult to parse, and the explanations of experimental results could be more detailed, particularly regarding why certain models benefit more than others.

Originality: The paper introduces an interesting combination of symbolic reasoning and approximation, representing a moderate step forward in neural network verification. The main ideas build upon existing frameworks but are extended in a way that broadens applicability. The originality is present but not transformative.

[3]https://anonymous.4open.science/r/CounterfactualTraining/paper/preprint/appendix.pdf

Significance: The significance is moderate. Verification of neural networks is an important and active area, and improvements in efficiency and bound tightness are valuable. However, the practical impact may be somewhat limited, as scalability to very large modern architectures remains uncertain, and the experimental validation is restricted to relatively modest datasets compared to current industry-scale benchmarks.

c) Strengths: The paper provides precise formal definitions, clear assumptions, and rigorous proofs (with details in the supplementary material). The proposed integration of symbolic reasoning with approximation is well-motivated, directly addressing the balance between tightness of bounds and computational efficiency. Experiments on MNIST, CIFAR-10, and selected ImageNet-scale models demonstrate that the method often achieves tighter certified bounds or faster runtimes than baselines. While gains are not uniform, the results consistently show competitiveness. Reporting of runtimes and scalability considerations adds practical value. The introduction effectively situates the work in the verification literature, highlighting limitations of prior methods. The extensive appendix (with pseudocode, proofs, and extra results) and the inclusion of an anonymous code/data link improve transparency and reproducibility, strengthening the credibility of the contribution.

d) Weaknesses: While the paper provides a useful extension, it builds closely on existing verification frameworks and does not introduce a fundamentally new paradigm. The improvements are more evolutionary than revolutionary, which may limit its perceived novelty at a top-tier venue like AAAI. Most experiments are on relatively small-scale benchmarks (e.g., MNIST, CIFAR-10). The claims of scalability to larger, more complex architectures are not strongly demonstrated, with ImageNet-scale experiments being limited in scope. This leaves open the question of whether the method can truly handle modern deep networks used in practice. In some experimental settings, the proposed method underperforms or shows only marginal improvements compared to baselines. The paper does not provide a deep analysis of why this occurs, missing an opportunity to characterize the limitations of the approach and guide future extensions. Several sections, particularly those with proofs and heavy symbolic notation, are difficult to follow for readers outside the formal verification community. While rigor is a strength, the paper could have done more to balance accessibility with technical detail, for instance by adding more intuition, diagrams, or worked examples. Although the paper reports numerical improvements in runtime and bound tightness, it does not sufficiently connect these gains to practical impact. For instance, it is unclear how much these improvements would matter in safety-critical deployments or adversarial robustness certification. A stronger discussion of practical significance would improve the paper's overall contribution.

e) Questions for the Authors: How well does the method scale to very large and modern architectures such as ResNets with hundreds of layers, Transformers, or vision-language models? Are there theoretical bottlenecks (e.g., in the symbolic component) or practical runtime/memory constraints that might prevent application at that scale? In experiments where your method underperforms or shows only marginal gains compared to baselines, what are the key contributing factors? For example, is performance more sensitive to network architecture type, dataset complexity, or perturbation size? A deeper error analysis would help characterize when your method is most effective. Could your framework be adapted to perturbations beyond the ℓ_p-norm settings tested, such as semantic perturbations (e.g., lighting, rotation, occlusion) or distributional shifts? If so, what challenges would arise in extending your symbolic-approximate approach to such settings? How sensitive are your results to hyperparameter choices in the approximation stage (e.g., step sizes, relaxation levels, stopping criteria)? Did you observe significant variability in runtime or bound tightness depending on these choices, and could you provide guidance for practitioners on tuning them? Do you envision this framework being applied in safety-critical settings such as autonomous driving, healthcare, or finance? If so, what adaptations would be required for deployment (e.g., guarantees on worst-case runtime, handling of real-world data distributions, integration with certification pipelines)?

Rating: 5: Marginally below acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

2) Review by bsoG:

a) Review: The paper introduces counterfactual training, using on-the-fly generated counterfactual explanations as some sort of data augmentation providing positive and negative examples (for the CF target class and for the original class depending on success). The paper shows that the introduction of the CF training improves CF explanations and adversarial robustness.

b) Pros:

- Interesting approach and sensible evaluation
- The paper is generally easy to follow

c) Cons:

- Limited novelty, unclear how much of the benefits stem from simple adversarial training. An ablation would help.
- Some notation inconsistencies.

d) Questions & Comments:

- Eq1 seems a bit incorrect. Shouldn't this be something like the argmin over the changes to x (or x') that make it look like the target class y+ with the regulariser taking delta x? (closer to the form in 3.1)
- It seems the notation is a bit inconsistent. E.g. in eq3 should y be y+ or rather in E(x, y) = -M(x)[y+] shouldn't this read y... y or y+ ... y+?
- How do you ensure 3.2 and 3.3 do not interfere? How do you ensure you don't use the same samples for both loss terms? How sensitive is the integration into the loss from 3.2 as this is basically bootstrapped and could suffer for highly imbalanced datasets or datasets with spurious correlations? Is this resolved because of the adversarial loss?

- How do you set τ in practice? Why 0.5? Shouldn't this depend on the number of classes? Or be based on entropy?

Rating: 5: Marginally below acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

3) Review by Nb3w:

a) Summary of the Paper: This study introduces a training paradigm, referred to as counterfactual training, which aims to train a classification model so that it can generate plausible and actionable counterfactual explanations (CEs) at test time.

b) Strengths: The proposed approach—training models to improve the quality of counterfactual explanations—is interesting and potentially valuable if the problem to solve is clearly identified. It appears to have improved plausibility and actionability for at least one CE method, ECCCo, even though the experimental description is vague.

c) Weakness 1: Unclear Contributions and Limitations: The primary limitation of this paper is that it introduces a new method without first identifying specific shortcomings of the current state-of-the-art CE methods. Consequently, the motivation for developing the proposed approach remains unclear. Unlike conventional test-time CE methods, the proposed method requires intervention during the training process, which could potentially influence the model's predictive performance. It is therefore important for the paper to explicitly justify this training-time intervention by clearly identifying the limitations that cannot be addressed by test-time CE methods.

In particular, the paper should address:

What specific problems exist in the state-of-the-art CE methods; and Why these problems necessitate intervention during training rather than solely at test time. The paper claims improvements in two desiderata—plausibility and actionability—but the rationale and scope of these improvements are not fully clear.

Regarding plausibility, it remains unclear whether the proposed method can enhance plausibility for any CE method or primarily for non-surrogate-based methods such as ECCCo. As discussed in the ECCCo paper (Altmeyer et al., 2024), most conventional plausibility-aware CE methods (e.g., REVISE) are surrogate-based; they rely on a surrogate generative model (e.g., a VAE) to ensure plausibility, meaning that plausibility depends largely on the surrogate rather than the classification model. A few methods such as ECCCo are non-surrogate-based; they enforce plausibility directly from the classification model. Thus, the proposed method, which trains the classification model to improve plausibility, may only benefit non-surrogate-based methods such as ECCCo. More specifically, the ideal output in Figure 1(b) can be achieved by just applying a surrogate-based method such as REVISE, instead of ECCCo.

The paper should clarify:

Whether the proposed method improves plausibility for any CE methods or is specific to non-surrogate-based methods; and If the paper claims that the proposed method is effective for surrogate-based CE methods, why intervention in the target model is preferable to intervention in the surrogate model. The paper also claims that the proposed method improves actionability by reducing action costs and increasing robustness. However, the desirability of cost reduction is debatable; for instance, in the context of loan approval, such a change could make approvals overly lenient, which might be detrimental to the lender. This could be seen as "moving the goalposts" rather than providing meaningful feedback to loan applicants. The robustness aspect is more compelling; however, if robustness is a central contribution, the paper should include direct comparisons with established adversarial training methods.

d) Weakness 2: Unclear and Insufficient Evaluations: The experimental section lacks clarity in several respects. While three CE methods (Generic, REVISE, and ECCCo) are mentioned, it is not always specified which method was used in each result. For example, Table 1 reports improvements in CE quality, but the caption suggests that ECCCo was used; if so, it is unclear where results for REVISE are reported. If the proposed method is intended to be method-agnostic, results for surrogate-based CE methods such as REVISE should be presented explicitly.

In addition, the claim that adversarial robustness is "greatly improved" is not supported by comparisons to relevant baselines, such as adversarial training or other robustness-oriented methods described in Section 2.2.

e) Suggestion: One possible way to address the issues outlined above is to narrow the focus to improving non-surrogate-based methods such as ECCCo, and to restructure the introduction accordingly:

Introduce non-surrogate-based methods such as ECCCo. Highlight their limitation—specifically, that their plausibility is determined by the classification model and therefore cannot be improved once training is complete. Propose counterfactual training as a solution, emphasizing that non-surrogate-based methods' plausibility can only be enhanced during the training phase. While this restructuring narrows the overall research scope, it would make the motivation for the training-time intervention clearer and more compelling.

f) Conclusion: The paper presents an interesting and potentially valuable approach to training models for improved counterfactual explanations. However, the contribution would be significantly strengthened by (i) more clearly identifying the limitations of existing methods, (ii) providing a well-reasoned justification for training-time intervention, and (iii) clarifying and expanding the experimental evaluation.

Rating: 4: Ok but not good enough - rejection

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

4) AI review:

a) Synopsis of the paper: The paper proposes counterfactual training, a learning regime that integrates counterfactual explanations into the training objective to align models with explanations that are both plausible and actionable. It constructs a contrastive divergence between mature counterfactuals and target-class data, reuses nascent counterfactuals

as adversarial examples to improve robustness, and enforces domain and mutability constraints during training. A theoretical result shows that, under linear-Gaussian assumptions, masking immutable features in the divergence reduces model sensitivity to them. Experiments across synthetic, tabular, and image datasets indicate improved counterfactual plausibility, reduced recourse costs under actionability constraints in many cases, and stronger robustness to adversarial perturbations.

b) Summary of Review: This is a timely and well-motivated contribution that moves beyond post-hoc explanation by "holding the model accountable" to produce plausible and actionable counterfactuals through an integrated training objective. The design—combining an energy-style contrastive term over mature counterfactuals with an adversarial loss over nascent counterfactuals and explicit actionability handling—is conceptually appealing and largely clearly presented, with a helpful suite of experiments. There are correctable technical issues in definitions and notation, and several implementation details that affect reproducibility need to be clarified. The empirical evaluation would be significantly stronger with stronger robustness baselines and attacks, ablations, and measurements of faithfulness and recourse robustness. Overall, the idea is promising and relevant; addressing the noted issues would materially strengthen both the technical soundness and the validity of the empirical claims.

c) Strengths:

- Problem framing and significance
  - Clearly articulates the limitation of post-hoc counterfactual generation that optimizes a single desideratum and may inadvertently harm others, and motivates training models to inherently support plausible and actionable explanations.
  - Situates the work within energy-based modeling and adversarial robustness, highlighting methodological links between counterfactuals and adversarial examples.
- Novel training objective with practical flexibility
  - Defines a unified objective that (i) minimizes a contrastive energy divergence between mature counterfactuals and data from the target class, (ii) applies an adversarial loss on nascent counterfactuals within an -ball, and (iii) encodes domain and mutability constraints during counterfactual generation. The approach is broadly applicable to differentiable classifiers and compatible with multiple counterfactual generators.
- Actionability encoding and theoretical insight
  - Proposes a principled procedure to protect immutable features by masking their contribution in the divergence term, so the model seeks plausibility through mutable features. Proposition 3.1 shows under linear-Gaussian assumptions that masking an immutable feature reduces classifier sensitivity to it relative to mutable features, aligning with theory on robust recourse that favors actionable features (Dominguez-Olmedo, Karimi, & Schölkopf, 2022).
- Empirical evidence of benefits
  - Across nine datasets, reports substantial and often statistically significant reductions in implausibility, and consistent robustness gains to gradient-based perturbations. On many datasets, valid counterfactuals become less costly under mutability constraints when training with the proposed masking.
- Insightful practical observations
  - Finds that faithful generators (e.g., energy-constrained approaches) benefit training most, while generators that compromise faithfulness yield weaker results. Shows that more counterfactual steps and larger numbers of generated counterfactuals per epoch tend to improve outcomes, and that the method can be used as a fine-tuning stage.

d) Weaknesses:

- Technical issues and ambiguities (correctable without changing the overall story)
  - Faithfulness and plausibility definitions: The conditions "$\int^A p_\theta(\mathbf{x}'|\mathbf{y}^+)d\mathbf{x} \to 1$" and "$\int^A p(\mathbf{x}'|\mathbf{y}^+)d\mathbf{x} \to 1$" over an "arbitrarily small" region A cannot approach 1 for continuous densities, and the integration variable should be x (not x' as a fixed point). A high-density or fixed-radius neighborhood criterion, or a threshold on $p_\theta(\mathbf{x}'|\mathbf{y}^+)$ and $p(\mathbf{x}'|\mathbf{y}^+)$, would be well-defined.
  - Energy divergence label mismatch: $\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta)$ uses y while $\mathcal{E}_\theta(\mathbf{x}, y) = -\mathbf{M}_\theta(\mathbf{x})[y^+]$ uses y^+. The divergence should consistently use the sampled target class index y^+.
  - Adversarial loss notation and linkage: The text defines $t_\varepsilon = \max_t\{t : ||\Delta_t||_\infty < \varepsilon\}$ but does not clearly link t_ε to x'_{AE} in the loss, and Δ_t (presumably $x'_t - x$) is not defined. It remains ambiguous whether $x'_{\text{AE}} = x_{t_\varepsilon}$ and what happens if no iterate satisfies the ε-constraint; ε is also omitted from the main objective's call to advloss$(\cdot)$.
  - Ridge regularization: The objective includes λ_{reg},ridge$(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta)$ but the exact form of ridge$(\cdot)$ is unspecified (e.g., penalty on energy magnitudes versus parameters). This is necessary for reproducibility and to understand its interaction with the contrastive term.
  - Algorithm 1 sampling details: The initialization of x'_0, the sampling policy for y^+ (uniform vs. empirical, ensuring $y^+ \neq y$), and how x^+ is selected (e.g., per-sample, in-batch, memory bank) are not described. Handling of non-maturing counterfactuals within T steps should also be specified.
  - Protected-feature masking terminology: The text describes setting $\mathbf{x}^+[d] - \mathbf{x}'[d] := 0$ for protected coordinates as a "point mass prior", but functionally this is a masking/projection in the divergence term rather than a prior over x^+. The Bayesian phrasing is misleading.
  - Typographical and notation issues: Minor typesetting artifacts (e.g., $\min_{x'} \in X^D$ with a stray D, $\lambda_{regridge}$),

an ambiguous phrase $\arg\max y^+ = y^+$ and inconsistent acronym usage ("ECCCo" vs. "ECCo") can confuse readers.
  - Plausibility metrics under-specified: The distance function in $IP(x', X^+)$ is not specified, and the kernel and bandwidth for $IP^*(X', X^*)$ (MMD) are not given. The statement that $IP^* = 0$ iff distributions match holds for characteristic kernels at the population level; these conditions should be stated.
- Evaluation limitations relative to the claims
  - Baselines: Comparisons are primarily against a conventional multilayer perceptron without counterfactual training. There are no head-to-head comparisons with strong adversarial training baselines (e.g., projected gradient descent training or TRADES), joint energy-based classifiers, deep ensembles, or recourse-aware learning methods (Ross, Lakkaraju, & Bastani, 2021; Guo, Nguyen, & Yadav, 2023), making it difficult to attribute gains specifically to the proposed objective.
  - Robustness assessment: Robust accuracy is reported under single-step FGSM; stronger iterative attacks (e.g., projected gradient descent, Carlini–Wagner) and transfer attacks are absent, limiting conclusions about robustness.
  - Faithfulness measurement: The approach relies on faithful counterfactuals for training, but empirical faithfulness is not quantified (e.g., conformity to learned energy surfaces or consistency across model perturbations/retraining).
  - Datasets and preprocessing: California Housing is originally a regression dataset; the binarization protocol is not described. The evaluation omits standard recourse datasets that test credit-risk settings and fairness aspects (e.g., HELOC, COMPAS) and omits more challenging vision benchmarks (e.g., CIFAR-10).
  - Ablations and diagnostics: There is no ablation isolating the contributions of the divergence term, the adversarial loss, and protected-feature masking, nor a comparison to simple counterfactual data augmentation. Convergence/validity rates of counterfactual generation, per-class outcomes, and sensitivity to the choice of protected features are not analyzed. Compute overheads are not quantified.
- Scope of theoretical guarantees and fairness analysis
  - Proposition 3.1 applies to linear classifiers with Gaussian class-conditional densities; this is a useful first step but does not establish behavior for nonlinear networks. While empirical findings are suggestive, additional systematic analyses of feature sensitivity would strengthen the claim.
  - The paper discusses fairness implications qualitatively (burden of recourse, proxy features) but does not include quantitative fairness or recourse-burden measurements on real datasets.

e) Suggestions for Improvement:

- Clarify and correct the mathematical formulation
  - Redefine faithfulness and plausibility in a way that is meaningful for continuous densities, e.g., via high-density thresholds, fixed-radius balls with nonzero measure, or explicit thresholds on $p_\theta(\mathbf{x}'|\mathbf{y}^+)$ and $p(\mathbf{x}'|\mathbf{y}^+)$. Use $p_\theta(\mathbf{x}'|\mathbf{y}^+)$ and $p(\mathbf{x}'|\mathbf{y}^+)$ with as the integration variable.
  - Make the divergence term self-consistent: use y^+ throughout in $\text{div}(\mathbf{x}^+, \mathbf{x}'_{\text{CE}}, y; \theta)$ and in $\mathcal{E}_\theta(\mathbf{x}, y^+)$.
  - Define $\Delta_t := x'_t - x$, explicitly connect x'_{AE} to t_ε (e.g., $x'_{AE} := x_{t_\varepsilon}$), include ε in the advloss$(\cdot)$ call in the main objective, and specify the fallback if no iterate satisfies the ε-constraint (e.g., use the last iterate, reduce ε, or skip the adversarial term for that sample).
  - Specify the exact ridge$(\cdot)$ term (e.g., ℓ_2 penalty on energy magnitudes or parameters), and discuss its role in preventing gradient explosion and stabilizing the contrastive term.
  - In Algorithm 1, add concrete sampling details: how x_0 is initialized (e.g., x plus small noise), how y^+ is sampled (uniform or class-balanced, ensuring $y^+ \neq y$), how x^+ is selected (e.g., in-batch same-class sample, memory bank with nearest neighbors), and how non-maturing counterfactuals are handled; report the fraction of mature counterfactuals per dataset.
  - Present the protected-feature operation as a masking/projection in the divergence (not a prior), and explain its effect on the optimization landscape.
  - Standardize notation (e.g. X vs. X^+,x'_{CE},x'_{AE},y^+,E_θ,S,τ,T,Δ_t,ε,λ_{div},λ_{adv},λ_{reg}), fix typographical artifacts, and clarify the ambiguous phrase "$\arg\max y^+ = y^+$".
- Strengthen the empirical evaluation
  - Add strong baselines for robustness and explainability: Adversarial training baselines such as projected gradient descent training (Madry et al., 2018) and TRADES (Zhang et al., 2019). Joint energy-based classifiers (Grathwohl et al., 2020) or deep ensembles (Lakshminarayanan, Pritzel, & Blundell, 2017) to test the incremental value of the contrastive divergence term. Recourse-aware learning baselines trained for actionable recourse (Ross, Lakkaraju, & Bastani, 2021) and end-to-end predictor–counterfactual architectures (Guo, Nguyen, & Yadav, 2023).
  - Evaluate robustness under stronger attacks: Iterative projected gradient descent with multiple steps and restarts; Carlini–Wagner attacks; and transfer attacks (e.g., ensemble-based) (Madry et al., 2018; Carlini & Wagner, 2017; Tramèr et al., 2018). Report clean and robust accuracy across , including attack step size, number of steps, and restarts.
  - Quantify faithfulness, stability, and robust recourse: Measure counterfactual consistency across retrained models or bootstrapped datasets (Black, Wang, & Fredrikson, 2022). Evaluate robustness of counterfac-

tuals with probabilistic guarantees or stability metrics (Hamman et al., 2023). Probe the cost–robustness frontier and recourse invalidation using PROBE (Pawelczyk et al., 2023), and consider verification-style tests to distinguish actionable perturbations from adversarial artifacts (Friedbaum, Adiga, & Tandon, 2024).

  - Expand datasets and clarify preprocessing: Describe the binarization protocol for California Housing. Add standard recourse datasets with credit/fairness relevance (e.g., HELOC; COMPAS) and a more challenging vision dataset (e.g., CIFAR-10) to assess generality (FICO, 2018; Larson et al., 2016; Krizhevsky, 2009).
  - Add ablations and diagnostics: Isolate contributions of each component: classification-only; +divergence; +adversarial; with/without protected-feature masking; versus simple counterfactual data augmentation (Abbasnejad et al., 2020; Balashankar et al., 2023). Report convergence and validity rates for counterfactual generation (mature vs. nascent), per-class outcomes, and sensitivity to which features are protected.
  - Quantify compute overhead: wall-clock time, number of counterfactuals per epoch, and resource usage relative to baseline, adversarial training, and joint energy-based methods.

- Improve presentation and usability
  - Include a schematic of the training pipeline showing how mature and nascent counterfactuals are produced per batch, where mutability/domain constraints and masking are applied, and how the different loss terms combine.
  - Clearly define the distance used in (e.g., or Mahalanobis) and the kernel and bandwidth used for (e.g., Gaussian kernel with median heuristic). State that implies distributional equality under a characteristic kernel in the population limit.
  - Where actionability and fairness are central (e.g., protecting age), complement qualitative observations with quantitative fairness analyses of burden of recourse across groups and proxy effects (Sharma, Henderson, & Ghosh, 2020; Dominguez-Olmedo, Karimi, & Schölkopf, 2022).

f) References: Abbasnejad, E., Teney, D., Parvaneh, A., Shi, J., & van den Hengel, A. (2020). Counterfactual vision and language learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10041–10051). Altmeyer, P., Farmanbar, M., van Deursen, A., & Liem, C. C. S. (2024). Faithful model explanations through Energy-Constrained Conformal Counterfactuals. Proceedings of the AAAI Conference on Artificial Intelligence, 38(10), 10829–10837. https://doi.org/10.1609/aaai.v38i10.28956 Balashankar, A., Wang, X., Qin, Y., Packer, B., Thain, N., Chi, E., Chen, J., & Beutel, A. (2023). Improving classifier robustness through active generative counterfactual data augmentation. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 127–139). Association for Computational Linguistics. Black, E., Wang, Z., & Fredrikson, M. (2022). Consistent counterfactuals for deep models. In International Conference on Learning Representations (ICLR 2022). https://openreview.net/forum?id=St6eyiTEHnG Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39–57). IEEE. Dominguez-Olmedo, R., Karimi, A. H., & Schölkopf, B. (2022). On the adversarial robustness of causal algorithmic recourse. In Proceedings of the 39th International Conference on Machine Learning (pp. 5324–5342). PMLR. FICO. (2018). Explainable Machine Learning Challenge: Home Equity Line of Credit (HELOC) dataset. FICO Community. https://community.fico.com/s/explainable-machine-learning-challenge Friedbaum, J., Adiga, S., & Tandon, R. (2024). Trustworthy actionable perturbations. In Proceedings of the 41st International Conference on Machine Learning (pp. 14006–14034). PMLR. Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., & Swersky, K. (2020). Your classifier is secretly an energy based model and you should treat it like one. In International Conference on Learning Representations (ICLR 2020). Guo, H., Nguyen, T. H., & Yadav, A. (2023). CounterNet: End-to-end training of prediction aware counterfactual explanations. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 577–589). Association for Computing Machinery. Hamman, F., Noorani, E., Mishra, S., Magazzeni, D., & Dutta, S. (2023). Robust counterfactual explanations for neural networks with probabilistic guarantees. In Proceedings of the 40th International Conference on Machine Learning (pp. 12351–12367). PMLR. Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report. University of Toronto. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30 (pp. 6405–6416). Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR 2018). Pawelczyk, M., Datta, T., van den Heuvel, J., Kasneci, G., & Lakkaraju, H. (2023). Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. In International Conference on Learning Representations (ICLR 2023). Ross, A., Lakkaraju, H., & Bastani, O. (2021). Learning models for actionable recourse. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: A common framework to provide explanations and analyze the fairness and robustness of black-box models. In Proceedings of the AAAI/ACM Conference

on AI, Ethics, and Society (pp. 166–172). Association for Computing Machinery. Tramèr, N., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations (ICLR 2018). Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., & Jordan, M. I. (2019). Theoretically principled trade-off between robustness and accuracy. In Proceedings of the 36th International Conference on Machine Learning (pp. 7472–7482). PMLR.