ECCCos from the Black Box: Faithful Explanations through Energy-Constrained Conformal Counterfactuals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Counterfactual Explanations offer an intuitive and straightforward way to explain black-box models and offer Algorithmic Recourse to individuals. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn how the input data is distributed. This effectively reallocates the task of learning realistic explanations for the data from the model itself to the surrogate. Consequently, the generated explanations may seem plausible to humans but need not necessarily describe the behaviour of the black-box model faithfully. We formalise this notion of faithfulness through the introduction of a tailored evaluation metric and propose a novel algorithmic framework for generating **E**nergy-**C**onstrained **C**onformal **Co**unterfactuals (ECCCos) that are only as plausible as the model permits. Through extensive empirical studies, we demonstrate that ECCCos reconcile the need for faithfulness and plausibility. In particular, we show that for models with gradient access, it is possible to achieve state-of-the-art performance without the need for surrogate models. To do so, our framework relies solely on properties defining the black-box model itself by leveraging recent advances in Energy-Based Modelling and Conformal Prediction. To our knowledge, this is the first venture in this direction for generating faithful Counterfactual Explanations. Thus, we anticipate that ECCCos can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

1 Introduction

Counterfactual Explanations (CE) provide a powerful, flexible and intuitive way to not only explain black-box models but also help affected individuals through the means of Algorithmic Recourse. Instead of opening the Black Box, CE works under the premise of strategically perturbing model inputs to understand model behaviour [1]. Intuitively speaking, we generate explanations in this context by asking what-if questions of the following nature: 'Our credit risk model currently predicts that this individual is not credit-worthy. What if they reduced their monthly expenditures by 10%?'

This is typically implemented by defining a target outcome $\mathbf{y}^+ \in \mathcal{Y}$ for some individual $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$ described by D attributes, for which the model $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ initially predicts a different outcome: $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$. Counterfactuals are then searched by minimizing a loss function that compares the predicted model output to the target outcome: $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^+)$. Since CE work directly with the black-box model, valid counterfactuals always have full local fidelity by construction where fidelity is defined as the degree to which explanations approximate the predictions of a black-box model [2, 3].

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

In situations where full fidelity is a requirement, CE offer a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [4] and SHAP [5], which involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring that an explanation faithfully describes the behaviour of a model. That is because multiple very distinct explanations can all lead to the same model prediction, especially when dealing with heavily parameterized models like deep neural networks, which are typically underspecified by the data [6].

In the context of CE, the idea that no two explanations are the same arises almost naturally. A key focus in the literature has therefore been to identify those explanations and algorithmic recourses that are most appropriate based on a myriad of desiderata such as sparsity, actionability and plausibility. In this work, we draw closer attention to model faithfulness rather than fidelity as a desideratum for counterfactuals. Our key contributions are as follows:

- We show that fidelity is an insufficient evaluation metric for counterfactuals (Section 3) and propose a definition of faithfulness that gives rise to more suitable metrics (Section 4).
- We introduce a novel algorithmic approach for generating Energy-Constrained Conformal Counterfactuals (ECCCos) in Section 5.
- We provide extensive empirical evidence demonstrating that ECCCos faithfully explain model behaviour and attain plausibility only when appropriate (Section 6).

To our knowledge, this is the first venture in this direction for generating faithful counterfactuals. Thus, we anticipate that ECCCos can serve as a baseline for future research. We believe that our work opens avenues for researchers and practitioners seeking tools to better distinguish trustworthy from unreliable models.

2 Background

While CE can also be generated for arbitrary regression models [7], existing work has primarily focused on classification problems. Let $\mathcal{Y} = (0, 1)^K$ denote the one-hot-encoded output domain with K classes. Then most counterfactual generators rely on gradient descent to optimize different flavours of the following counterfactual search objective:

$$\mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^L} \left\{ \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda \text{cost}(f(\mathbf{Z}')) \right\} \tag{1}$$

Here yloss$(\cdot)$ denotes the primary loss function, $f(\cdot)$ is a function that maps from the counterfactual state space to the feature space and cost$(\cdot)$ is either a single penalty or a collection of penalties that are used to impose constraints through regularization. Equation 1 restates the baseline approach to gradient-based counterfactual search proposed by Wachter et al. [1] in general form as introduced by Altmeyer et al. [8]. To explicitly account for the multiplicity of explanations, $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes an L-dimensional array of counterfactual states.

The baseline approach, which we will simply refer to as *Wachter*, searches a single counterfactual directly in the feature space and penalises its distance to the original factual. In this case, $f(\cdot)$ is simply the identity function and $\mathcal{Z}$ corresponds to the feature space itself. Many derivative works of Wachter et al. [1] have proposed new flavours of Equation 1, each of them designed to address specific *desiderata* that counterfactuals ought to meet in order to properly serve both AI practitioners and individuals affected by algorithmic decision-making systems. The list of desiderata includes but is not limited to the following: sparsity, proximity [1], actionability [9], diversity [2], plausibility [10, 11, 12], robustness [13, 14, 8] and causality [15]. Different counterfactual generators addressing these needs have been extensively surveyed and evaluated in various studies [16, 17, 18, 19, 20].

Perhaps unsurprisingly, the different desiderata are often positively correlated. For example, Artelt et al. [19] find that plausibility typically also leads to improved robustness. Similarly, plausibility has also been connected to causality in the sense that plausible counterfactuals respect causal relationships [21]. Consequently, the plausibility of counterfactuals has been among the primary concerns for researchers. Achieving plausibility is equivalent to ensuring that the generated counterfactuals comply with the true and unobserved data-generating process (DGP). We define plausibility formally in this work as follows:

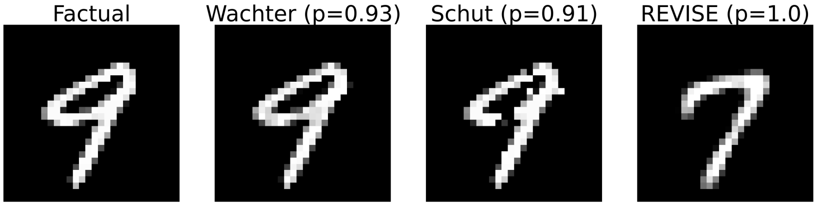

Figure 1: Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left); then from left to right the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

Definition 2.1 (Plausible Counterfactuals)**.** *Let $\mathcal{X}|\mathbf{y}^+ = p(\mathbf{x}|\mathbf{y}^+)$ denote the true conditional distribution of samples in the target class $\mathbf{y}^+$. Then for $\mathbf{x}'$ to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.*

To generate plausible counterfactuals, we need to be able to quantify the DGP: $\mathcal{X}|\mathbf{y}^+$. One straightforward way to do this is to use surrogate models for the task. Joshi et al. [10], for example, suggest that instead of searching counterfactuals in the feature space $\mathcal{X}$, we can instead traverse a latent embedding $\mathcal{Z}$ (Equation 1) that implicitly codifies the DGP. To learn the latent embedding, they propose using a generative model such as a Variational Autoencoder (VAE). Provided the surrogate model is well-specified, their proposed approach called *REVISE* can yield plausible explanations. Others have proposed similar approaches: Dombrowski et al. [22] traverse the base space of a normalizing flow to solve Equation 1; Poyiadzi et al. [11] use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to constrain the counterfactuals to dense regions in the feature space; and, finally, Karimi et al. [15] assume knowledge about the structural causal model that generates the data.

A competing approach towards plausibility that is also closely related to this work instead relies on the black-box model itself. Schut et al. [12] show that to meet the plausibility objective we need not explicitly model the input distribution. Pointing to the undesirable engineering overhead induced by surrogate models, they propose that we rely on the implicit minimisation of predictive uncertainty instead. Their proposed methodology, which we will refer to as *Schut*, solves Equation 1 by greedily applying Jacobian-Based Saliency Map Attacks (JSMA) in the feature space with cross-entropy loss and no penalty at all. The authors demonstrate theoretically and empirically that their approach yields counterfactuals for which the model M_θ predicts the target label $\mathbf{y}^+$ with high confidence. Provided the model is well-specified, these counterfactuals are plausible. This idea hinges on the assumption that the black-box model provides well-calibrated predictive uncertainty estimates.

3 Why Fidelity is not Enough

As discussed in the introduction, any valid counterfactual also has full fidelity by construction: solutions to Equation 1 are considered valid as soon as the label predicted by the model matches the target class. So while fidelity always applies, counterfactuals that address the various desiderata introduced above can look vastly different from each other.

To demonstrate this with an example, we have trained a simple image classifier M_θ on the well-known *MNIST* dataset [23]: a Multi-Layer Perceptron (*MLP*) with above 90 percent test accuracy. No measures have been taken to improve the model's adversarial robustness or its capacity for predictive uncertainty quantification. The far left panel of Figure 1 shows a random sample drawn from the dataset. The underlying classifier correctly predicts the label 'nine' for this image. For the given factual image and model, we have used *Wachter*, *Schut* and *REVISE* to generate one counterfactual each in the target class 'seven'. The perturbed images are shown next to the factual image from left to right in Figure 1. Captions on top of the individual images indicate the generator along with the predicted probability that the image belongs to the target class. In all three cases that probability is above 90 percent and yet the counterfactuals look very different from each other.

Since *Wachter* is only concerned with proximity, the generated counterfactual is almost indistinguishable from the factual. The approach by Schut et al. [12] expects a well-calibrated model that can generate predictive uncertainty estimates. Since this is not the case, the generated counterfactual looks like an adversarial example. Finally, the counterfactual generated by *REVISE* looks much more plausible than the other two. But is it also more faithful to the behaviour of our *MNIST* classifier?

That is much less clear because the surrogate used by *REVISE* introduces friction: the generated explanations no longer depend exclusively on the black-box model itself.

So which of the counterfactuals most faithfully explains the behaviour of our image classifier? Fidelity cannot help us to make that judgement, because all of these counterfactuals have full fidelity. Thus, fidelity is an insufficient evaluation metric to assess the faithfulness of CE.

4 A New Notion of Faithfulness

Considering the limitations of fidelity as demonstrated in the previous section, analogous to Definition 2.1, we introduce a new notion of faithfulness in the context of CE:

Definition 4.1 (Faithful Counterfactuals)**.** *Let $\mathcal{X}_\theta|\mathbf{y}^+ = p_\theta(\mathbf{x}|\mathbf{y}^+)$ denote the conditional distribution of $\mathbf{x}$ in the target class $\mathbf{y}^+$, where θ denotes the parameters of model M_θ. Then for $\mathbf{x}'$ to be considered a faithful counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^+$.*

In doing this, we merge in and nuance the concept of plausibility (Definition 2.1) where the notion of 'consistent with the data' becomes 'consistent with what the model has learned about the data'.

4.1 Quantifying the Model's Generative Property

To assess counterfactuals with respect to Definition 4.1, we need a way to quantify the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}^+)$. To this end, we draw on recent advances in Energy-Based Modelling (EBM), a subdomain of machine learning that is concerned with generative or hybrid modelling [24, 25]. In particular, note that if we fix $\mathbf{y}$ to our target value $\mathbf{y}^+$, we can conditionally draw from $p_\theta(\mathbf{x}|\mathbf{y}^+)$ by randomly initializing $\mathbf{x}_0$ and then using Stochastic Gradient Langevin Dynamics (SGLD) as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon^2}{2}\mathcal{E}(\mathbf{x}_j|\mathbf{y}^+) + \epsilon \mathbf{r}_j, \quad j = 1, ..., J \tag{2}$$

where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ is typically polynomially decayed [26]. The term $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^+)$ denotes the model energy conditioned on the target class label $\mathbf{y}^+$ which we specify as the negative logit corresponding to the target class label $\mathbf{y}^*$. To allow for faster sampling, we follow the common practice of choosing the step-size ϵ and the standard deviation of $\mathbf{r}_j$ separately. While $\mathbf{x}_J$ is only guaranteed to distribute as $p_\theta(\mathbf{x}|\mathbf{y}^*)$ if $\epsilon \rightarrow 0$ and $J \rightarrow \infty$, the bias introduced for a small finite ϵ is negligible in practice [27, 24]. Appendix A provides additional implementation details for any tasks related to energy-based modelling.

Generating multiple samples using SGLD thus yields an empirical distribution $\hat{\mathbf{X}}_{\theta,\mathbf{y}^+}$ that approximates what the model has learned about the input data. While in the context of EBM, this is usually done during training, we propose to repurpose this approach during inference in order to evaluate and generate faithful model explanations.

4.2 Evaluating Plausibility and Faithfulness

The parallels between our definitions of plausibility and faithfulness imply that we can also use similar evaluation metrics in both cases. Since existing work has focused heavily on plausibility, it offers a useful starting point. In particular, Guidotti [20] have proposed an implausibility metric that measures the distance of the counterfactual from its nearest neighbour in the target class. As this distance is reduced, counterfactuals get more plausible under the assumption that the nearest neighbour itself is plausible in the sense of Definition 2.1. In this work, we use the following adapted implausibility metric,

$$\text{impl}(\mathbf{x}', \mathbf{X}_{\mathbf{y}^+}) = \frac{1}{|\mathbf{X}_{\mathbf{y}^+}|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{3}$$

where $\mathbf{x}'$ denotes the counterfactual and $\mathbf{X}_{\mathbf{y}^+}$ is a subsample of the training data in the target class $\mathbf{y}^+$. By averaging over multiple samples in this manner, we avoid the risk that the nearest neighbour of $\mathbf{x}'$ itself is not plausible according to Definition 2.1 (e.g an outlier).

Equation 3 gives rise to a similar evaluation metric for unfaithfulness. We merely swap out the subsample of individuals in the target class for a subset $\hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}$ of the generated conditional samples:

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

Specifically, we form this subset based on the n_E generated samples with the lowest energy.

5 Energy-Constrained Conformal Counterfactuals

In this section, we describe *ECCCo*, our proposed framework for generating Energy-Constrained Conformal Counterfactuals (ECCCos). It is based on the premise that counterfactuals should first and foremost be faithful. Plausibility, as a secondary concern, is then still attainable, but only to the degree that the black-box model itself has learned plausible explanations for the underlying data.

We begin by stating our proposed objective function, which involves tailored loss and penalty functions that we will explain in the following. In particular, we extend Equation 1 as follows:

$$\begin{aligned} \mathbf{Z}' = \arg \min_{\mathbf{Z}' \in \mathcal{Z}^M} \{ & \text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ & + \lambda_2 \text{unfaith}(f(\mathbf{Z}'), \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha)) \} \end{aligned} \tag{5}$$

The first penalty term involving λ_1 induces proximity like in Wachter et al. [1]. Our default choice for dist$(\cdot)$ is the L1 Norm due to its sparsity-inducing properties. The second penalty term involving λ_2 induces faithfulness by constraining the energy of the generated counterfactual where unfaith$(\cdot)$ corresponds to the metric defined in Equation 4. The third and final penalty term involving λ_3 introduces a new concept: it ensures that the generated counterfactual is associated with low predictive uncertainty. As mentioned above, Schut et al. [12] have shown that plausible counterfactuals can be generated implicitly through predictive uncertainty minimization. Unfortunately, this relies on the assumption that the model itself can provide predictive uncertainty estimates, which may be too restrictive in practice.

To relax this assumption, we leverage recent advances in Conformal Prediction (CP), an approach to predictive uncertainty quantification that has recently gained popularity [28, 29]. Crucially for our intended application, CP is model-agnostic and can be applied during inference without placing any restrictions on model training. Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates by repeatedly sifting through the training data or a dedicated calibration dataset. Conformal classifiers produce prediction sets for individual inputs that include all output labels that can be reasonably attributed to the input. These sets tend to be larger for inputs that do not conform with the training data and are characterized by high predictive uncertainty.

In order to generate counterfactuals that are associated with low predictive uncertainty, we use a smooth set size penalty introduced by Stutz et al. [30] in the context of conformal training:

$$\Omega(C_\theta(\mathbf{x}; \alpha)) = \max \left(0, \sum_{\mathbf{y} \in \mathcal{Y}} C_{\theta,\mathbf{y}}(\mathbf{x}_i; \alpha) - \kappa \right) \tag{6}$$

Here, $\kappa \in \{0, 1\}$ is a hyper-parameter and $C_{\theta,\mathbf{y}}(\mathbf{x}_i; \alpha)$ can be interpreted as the probability of label $\mathbf{y}$ being included in the prediction set. In order to compute this penalty for any black-box model we merely need to perform a single calibration pass through a holdout set $\mathcal{D}_{\text{cal}}$. Arguably, data is typically abundant and in most applications, practitioners tend to hold out a test data set anyway. Consequently, CP removes the restriction on the family of predictive models, at the small cost of reserving a subset of the available data for calibration. This particular case of conformal prediction is referred to as Split Conformal Prediction (SCP) as it involves splitting the training data into a proper training dataset and a calibration dataset. In addition to the smooth set size penalty, we have also experimented with the use of a tailored function for yloss$(\cdot)$ that enforces that only the target label $\mathbf{y}^+$ is included in the prediction set Stutz et al. [30]. Further details are provided in Appendix B.

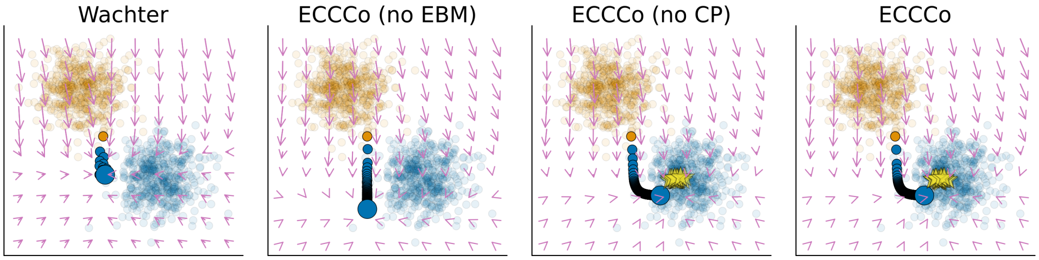

Figure 2: Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfual in the 'blue' class for a sample from the 'orange' class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

Algorithm 1 The *ECCCo* generator

```
Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T, \eta, n_{\mathcal{B}}, n_E$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: $\mathbf{x}'$
1: Initialize $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$    ▷ Map to counterfactual state space.
2: Generate $\left\{\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}\right\}_{n_{\mathcal{B}}} \leftarrow p_\theta(\mathbf{x}_{\mathbf{y}^+})$    ▷ Generate $n_{\mathcal{B}}$ samples using SGLD (Equation 2).
3: Store $\hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+} \leftarrow \left\{\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}\right\}_{n_{\mathcal{B}}}$    ▷ Choose $n_E$ lowest-energy samples.
4: Run SCP for $M_\theta$ using $\mathcal{D}$    ▷ Calibrate model through Split Conformal Prediction.
5: Initialize $t \leftarrow 0$
6: while not converged or $t < T$ do    ▷ For convergence conditions see Appendix C.
7:     $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}; \Lambda, \alpha)$    ▷ Take gradient step of size $\eta$.
8:     $t \leftarrow t + 1$
9: end while
10: $\mathbf{x}' \leftarrow f(\mathbf{z}')$    ▷ Map back to feature space.
```

To provide some further intuition about our objective defined in Equation 5, Figure 2 illustrates how the different components affect the counterfactual search for a synthetic dataset. The underlying classifier is a Joint Energy Model (*JEM*) that was trained to predict the output class ('blue' or 'orange') and generate class-conditional samples [24]. We have used four different generator flavours to produce a counterfactual in the 'blue' class for a sample from the 'orange' class: *Wachter*, which only uses the first penalty ($\lambda_2 = \lambda_3 = 0$); *ECCCo (no EBM)*, which does not constrain energy ($\lambda_2 = 0$); *ECCCo (no CP)*, which involves no set size penalty ($\lambda_3 = 0$); and, finally, *ECCCo*, which involves all penalties defined in Equation 5. Arrows indicate (negative) gradients with respect to the objective function at different points in the feature space.

While *Wachter* generates a valid counterfactual, it ends up close to the original starting point consistent with its objective. *ECCCo (no EBM)* pushes the counterfactual further into the target domain to minimize predictive uncertainty, but the outcome is still not plausible. The counterfactual produced by *ECCCo (no CP)* is attracted by the generated samples shown in bright yellow. Since the *JEM* has learned the conditional input distribution reasonably well in this case, the counterfactuals are both faithful and plausible. Finally, the outcome for *ECCCo* looks similar, but the additional smooth set size penalty leads to somewhat faster convergence.

Algorithm 1 describes how exactly *ECCCo* works. For the sake of simplicity and without loss of generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$. The counterfactual state $\mathbf{z}'$ is initialized by passing the factual $\mathbf{x}$ through a simple feature transformer f^{-1}. Next, we generate $n_{\mathcal{B}}$ conditional samples $\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}$ using SGLD (Equation 2) and store the n_E instances with the lowest energy. We then calibrate the model M_θ through Split Conformal Prediction. Finally, we search counterfactuals through gradient descent where $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 5. The search terminates once the convergence criterium is met or the maximum number of iterations T has been exhausted. Note that the choice of convergence criterium has important implications on the final counterfactual which we explain in Appendix C.

6 Empirical Analysis

Our goal in this section is to shed light on the following research questions:

Research Question 6.1 (Faithfulness)**.** *Are ECCCos more faithful than counterfactuals produced by our benchmark generators?*

Research Question 6.2 (Balancing Objectives)**.** *Compared to our benchmark generators, how do ECCCos balance the two key objectives of faithfulness and plausibility?*

The second question is motivated by the intuition that faithfulness and plausibility should coincide for models that have learned plausible explanations of the data. Next, we first briefly describe our experimental setup before presenting our main results.

6.1 Experimental Setup

To assess and benchmark the performance of our proposed generator against the state of the art, we generate multiple counterfactuals for different models and datasets. In particular, we compare *ECCCo* and its variants to the following counterfactual generators that were introduced above: firstly; *Schut*, which works under the premise of minimizing predictive uncertainty; secondly, *REVISE*, which is state-of-the-art with respect to plausibility; and, finally, *Wachter*, which serves as our baseline.

We use both synthetic and real-world datasets from different domains, all of which are publicly available and commonly used to train and benchmark classification algorithms. We synthetically generate a dataset containing two *Linearly Separable* Gaussian clusters ($n = 1000$), as well as the well-known *Circles* ($n = 1000$) and *Moons* ($n = 2500$) data. Since these data are generated by distributions of varying degrees of complexity, they allow us to assess how the generators and our proposed evaluation metrics handle this.

As for real-world data, we follow Schut et al. [12] and use the *MNIST* [23] dataset containing images of handwritten digits such as the example shown above in Figure 1. From the social sciences domain, we include Give Me Some Credit (*GMSC*) [31]: a tabular dataset that has been studied extensively in the literature on Algorithmic Recourse [18]. It consists of 11 numeric features that can be used to predict the binary outcome variable indicating whether retail borrowers experience financial distress.

For the predictive modelling tasks, we use simple neural networks (*MLP*) and Joint Energy Models (*JEM*). For the more complex real-world datasets we also use ensembling in each case. Both joint-energy modelling and ensembling have been associated with improved generative properties and adversarial robustness [24, 32], so we expect this to be positively correlated with the plausibility of ECCCos. To account for stochasticity, we generate multiple counterfactuals for each target class, generator, model and dataset. Specifically, we randomly sample n^{-} times from the subset of individuals for which the given model predicts the non-target class $\mathbf{y}^{-}$ given the current target. We set $n^{-} = 25$ for all of our synthetic datasets, $n^{-} = 10$ for *GMSC* and $n^{-} = 5$ for *MNIST*. Full details concerning our parameter choices, training procedures and model performance can be found in Appendix D.

6.2 Results for Synthetic Data

Table 1 shows the key results for the synthetic datasets separated by model (first column) and generator (second column). The numerical columns show sample averages and standard deviations of our key evaluation metrics computed across all counterfactuals. We have highlighted the best outcome for each model and metric in bold. To provide some sense of effect sizes, we have added asterisks to indicate that a given value is at least one ($*$) or two ($**$) standard deviations lower than the baseline (*Wachter*).

Starting with the high-level results for our *Linearly Separable* data, we find that *ECCCo* produces the most faithful counterfactuals for both black-box models. This is consistent with our design since *ECCCo* directly enforces faithfulness through regularization. Crucially though, *ECCCo* also produces the most plausible counterfactuals for both models. This dataset is so simple that even the *MLP* has learned plausible explanations of the input data. Zooming in on the granular details for the *Linearly Separable* data, the results for *ECCCo (no CP)* and *ECCCo (no EBM)* indicate that the positive results are dominated by the effect of quantifying and leveraging the model's generative property (EBM). Conformal Prediction alone only leads to marginally improved faithfulness and plausibility.

Table 1: Results for synthetic datasets: sample averages +/- one standard deviation across counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (Wachter).

		Linearly Separable		Moons		Circles	
Model	Generator	Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓
JEM	ECCCo	**0.03 ± 0.06****	**0.20 ± 0.08****	**0.31 ± 0.30***	**1.20 ± 0.15****	0.52 ± 0.36	1.22 ± 0.46
	ECCCo (no CP)	0.03 ± 0.06**	0.20 ± 0.08**	0.37 ± 0.30*	1.21 ± 0.17**	0.54 ± 0.39	1.21 ± 0.46
	ECCCo (no EBM)	0.16 ± 0.11	0.34 ± 0.19	0.91 ± 0.32	1.71 ± 0.25	0.70 ± 0.33	1.30 ± 0.37
	REVISE	0.19 ± 0.03	0.41 ± 0.01**	0.78 ± 0.23	1.57 ± 0.26	**0.48 ± 0.16***	**0.95 ± 0.32***
	Schut	0.39 ± 0.07	0.73 ± 0.17	0.67 ± 0.27	1.50 ± 0.22*	0.54 ± 0.43	1.28 ± 0.53
	Wachter	0.18 ± 0.10	0.44 ± 0.17	0.80 ± 0.27	1.78 ± 0.24	0.68 ± 0.34	1.33 ± 0.32
MLP	ECCCo	**0.29 ± 0.05****	0.23 ± 0.06**	0.80 ± 0.62	1.69 ± 0.40	0.65 ± 0.53	1.17 ± 0.41
	ECCCo (no CP)	0.29 ± 0.05**	**0.23 ± 0.07****	**0.79 ± 0.62**	1.68 ± 0.42	**0.49 ± 0.35**	1.19 ± 0.44
	ECCCo (no EBM)	0.46 ± 0.05	0.28 ± 0.04**	1.34 ± 0.47	1.68 ± 0.47	0.84 ± 0.51	1.23 ± 0.31
	REVISE	0.56 ± 0.05	0.41 ± 0.01	1.45 ± 0.44	**1.64 ± 0.31**	0.58 ± 0.52	**0.95 ± 0.32**
	Schut	0.43 ± 0.06*	0.47 ± 0.36	1.45 ± 0.55	1.73 ± 0.48	0.58 ± 0.37	1.23 ± 0.43
	Wachter	0.51 ± 0.04	0.40 ± 0.08	1.32 ± 0.41	1.69 ± 0.32	0.83 ± 0.50	1.24 ± 0.29

The findings for the *Moons* dataset are broadly in line with the findings so far: for the *JEM*, *ECCCo* yields substantially more faithful and plausible counterfactuals than all other generators. For the *MLP*, faithfulness is maintained but counterfactuals are not plausible. This high-level pattern is broadly consistent with other more complex datasets and supportive of our narrative, so it is worth highlighting: ECCCos consistently achieve high faithfulness, which—subject to the quality of the model itself—coincides with high plausibility. By comparison, *REVISE* yields the most plausible counterfactuals for the *MLP*, but it does so at the cost of faithfulness. We also observe that the best results for *ECCCo* are achieved when using both penalties. Once again though, the generative component (EBM) has a stronger impact on the positive results for the *JEM*.

For the *Circles* data, it appears that *REVISE* performs well, but we note that it generates valid counterfactuals only half of the time (see Appendix E for a complete overview including additional common evaluation metrics). The underlying VAE with default parameters has not adequately learned the data-generating process. Of course, it is possible to improve generative performance through hyperparameter tuning but this example serves to illustrate that *REVISE* depends on the quality of its surrogate. Independent of the outcome for *REVISE*, however, the results do not seem to indicate that *ECCCo* substantially improves faithfulness and plausibility for the *Circles* data. We think this points to a limitation of our evaluation metrics rather than *ECCCo* itself: computing average distances fails to account for the 'wraparound' effect associated with circular data [33].

6.3 Results for Real-World Data

The results for our real-world datasets are shown in Table 2. Once again the findings indicate that the plausibility of ECCCos is positively correlated with the capacity of the black-box model to distinguish plausible from implausible inputs. The case is very clear for *MNIST*: ECCCos are consistently more faithful than the counterfactuals produced by our benchmark generators and their plausibility gradually improves through ensembling and joint-energy modelling. Interestingly, faithfulness also gradually improves for *REVISE*. This indicates that as our models improve, their generative capacity approaches that of the surrogate VAE used by *REVISE*. The VAE still outperforms our classifiers in this regard, as evident from the fact that *ECCCo* never quite reaches the same level of plausibility as *REVISE*. With reference to Appendix E we note that the results for *Schut* need to be discounted as it rarely produces valid counterfactuals for *MNIST*. Relatedly, we find that *ECCCo* is the only generator that consistently achieves full validity. Finally, it is worth noting that *ECCCo* produces counterfactual images with the lowest average predictive uncertainty for all models.

For the tabular credit dataset (*GMSC*) it is inherently challenging to use deep neural networks in order to achieve good discriminative performance [34, 35] and generative performance [36], respectively. In order to achieve high plausibility, *ECCCo* effectively requires classifiers to achieve good performance for both tasks. Since this is a challenging task even for Joint Energy Models, it is not surprising to find that even though *ECCCo* once again achieves state-of-the-art faithfulness, it is outperformed by *REVISE* and *Schut* with respect to plausibility.

Table 2: Results for real-world datasets: sample averages +/- one standard deviation across counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (Wachter).

		MNIST		GMSC	
Model	Generator	Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓
JEM	ECCCo	**19.28 ± 5.01****	314.76 ± 32.36*	**79.16 ± 11.67****	18.26 ± 4.92**
	REVISE	188.70 ± 26.18*	**255.26 ± 41.50****	186.40 ± 28.06	**5.34 ± 2.38****
	Schut	211.00 ± 27.21	286.61 ± 39.85*	200.98 ± 28.49	6.50 ± 2.01**
	Wachter	222.90 ± 26.56	361.88 ± 39.74	214.08 ± 45.35	61.04 ± 2.58
JEM Ensemble	ECCCo	**15.99 ± 3.06****	294.72 ± 30.75**	**83.28 ± 13.26****	17.21 ± 4.46**
	REVISE	173.59 ± 20.65**	**246.32 ± 37.46****	194.24 ± 35.41	**4.95 ± 1.26****
	Schut	205.33 ± 24.07	287.39 ± 39.33*	208.45 ± 34.60	6.12 ± 1.91**
	Wachter	217.67 ± 23.78	363.23 ± 39.24	186.19 ± 33.88	60.70 ± 44.32
MLP	ECCCo	**41.95 ± 6.50****	591.58 ± 36.24	**75.93 ± 14.27****	17.20 ± 3.15**
	REVISE	365.82 ± 15.35*	**249.49 ± 41.55****	196.75 ± 41.25	**4.84 ± 0.60****
	Schut	382.44 ± 17.81	285.98 ± 42.48*	212.00 ± 41.15	6.44 ± 1.34**
	Wachter	386.05 ± 16.60	361.83 ± 42.18	218.34 ± 53.26	45.84 ± 39.39
MLP Ensemble	ECCCo	**31.43 ± 3.91****	490.88 ± 27.19	**73.86 ± 14.63****	17.92 ± 4.17**
	REVISE	337.74 ± 11.89*	**247.67 ± 38.36****	207.21 ± 43.20	**5.78 ± 2.10****
	Schut	359.54 ± 14.52	283.99 ± 41.08*	205.36 ± 32.11	7.00 ± 2.15**
	Wachter	360.79 ± 14.39	357.73 ± 42.55	213.71 ± 54.17	73.09 ± 64.50

6.4 Key Takeways

To conclude this section, we summarize our findings with reference to the opening questions. The results clearly demonstrate that *ECCCo* consistently achieves state-of-the-art faithfulness, as it was designed to do (Research Question 6.1). A related important finding is that *ECCCo* yields highly plausible explanations provided that they faithfully describe model behaviour (Research Question 6.2). *ECCCo* achieves this result primarily by leveraging the model's generative property.

7 Limitations

Even though we have taken considerable measures to study our proposed methodology carefully, limitations can still be identified. In particular, we have found that the performance of *ECCCo* is sensitive to hyperparameter choices. In order to achieve faithfulness, we generally had to penalise the distance from generated samples slightly more than the distance from factual values.

Conversely, we have not found that strongly penalising prediction set sizes had any discernable effect. Our results indicate that CP alone is often not sufficient to achieve faithfulness and plausibility, although we acknowledge that this needs to be investigated more thoroughly through future work.

While our approach is readily applicable to models with gradient access like deep neural networks, more work is needed to generalise it to other machine learning models such as decision trees. Relatedly, common challenges associated with Energy-Based Modelling including sensitivity to scale, training instabilities and sensitivity to hyperparameters also apply to *ECCCo*.

8 Conclusion

This work leverages recent advances in Energy-Based Modelling and Conformal Prediction in the context of Explainable Artificial Intelligence. We have proposed a new way to generate counterfactuals that are maximally faithful to the black-box model they aim to explain. Our proposed generator, *ECCCo*, produces plausible counterfactuals if and only if the black-box model itself has learned realistic explanations for the data, which we have demonstrated through rigorous empirical analysis. This should enable researchers and practitioners to use counterfactuals in order to discern trustworthy models from unreliable ones. While the scope of this work limits its generalizability, we believe that *ECCCo* offers a solid baseline for future work on faithful Counterfactual Explanations.

References

[1] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[2] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[3] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[6] Andrew Gordon Wilson. The case for Bayesian deep learning. 2020.

[7] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. Counterfactual Explanations for Arbitrary Regression Models. 2021.

[8] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[9] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[10] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. 2019.

[11] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[12] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.

[13] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. 2021.

[14] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*, 2022.

[15] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[16] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. 2020.

[17] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.

[18] Martin Pawelczyk, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms. 2021.

[19] André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating Robustness of Counterfactual Explanations. Technical report, arXiv. URL http://arxiv.org/abs/2103.02354. arXiv:2103.02354 [cs] type: article.

[20] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6. URL https://doi.org/10.1007/s10618-022-00831-6.

[21] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. Technical report, arXiv. URL http://arxiv.org/abs/1912.03277. arXiv:1912.03277 [cs, stat] type: article.

[22] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[23] Yann LeCun. The MNIST database of handwritten digits. 1998.

[24] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. March 2020. URL https://openreview.net/forum?id=Hkxzx0NtDB.

[25] Yilun Du and Igor Mordatch. Implicit Generation and Generalization in Energy-Based Models. Technical report, arXiv. URL http://arxiv.org/abs/1903.08689. arXiv:1903.08689 [cs, stat] type: article.

[26] M. Welling and Y. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. URL https://www.semanticscholar.org/paper/Bayesian-Learning-via-Stochastic-Gradient-Langevin-Welling-Teh/aeed631d6a84100b5e9a021ec1914095c66de415.

[27] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.

[28] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. 2021.

[29] Valery Manokhin. Awesome conformal prediction.

[30] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning Optimal Conformal Classifiers. May 2022. URL https://openreview.net/forum?id=t8O-4LKFVx.

[31] 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., 2011. URL https://www.kaggle.com/c/GiveMeSomeCredit.

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 2016.

[33] Jeff Gill and Dominik Hangartner. Circular Data in Political Science and How to Handle It. 18(3):316–336. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpq009. URL https://www.cambridge.org/core/journals/political-analysis/article/circular-data-in-political-science-and-how-to-handle-it/6DF2D9DA60C455E6A48FFB0FF011F747.

[34] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. 2021.

[35] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? 2022.

[36] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure. URL https://openreview.net/forum?id=fPVRcJqspu.

[37] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. Technical report, arXiv. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs] type: article.