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 involving multiple synthetic and real-world datasets, we demonstrate that ECCCos reconcile the need for plausibility and faithfulness. In particular, we show that it is possible to achieve state-of-the-art plausibility for models with gradient access 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 [31]. 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 Counterfactual Explanations 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 [19, 18].

In situations where full fidelity is a requirement, CE offers a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [24] and SHAP [14], 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 [32].

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 as a desideratum for counterfactuals. Our key contributions are as follows: firstly, we propose a definition of faithfulness that is suitable for counterfactuals; secondly, we introduce a novel algorithmic approach for generating Energy-Constrained Conformal Counterfactuals (ECCCos) that explicitly addresses the need for faithfulness; finally, we provide extensive empirical evidence demonstrating that ECCCos faithfully explain model behaviour without sacrificing plausibility.

2 Background and Related Work

In this section, we provide some background on CE and our motivation for this work. To start, we briefly introduce the methodology underlying most state-of-the-art (SOTA) counterfactual generators.

2.1 Gradient-Based Counterfactual Search

While Counterfactual Explanations can be generated for arbitrary regression models [26], 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 denotes the primary loss function already introduced above and cost 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. [31] in general form where $\mathbf{Z}' = \{\mathbf{z}_l\}_L$ denotes an L-dimensional array of counterfactual states [2]. This is to explicitly account for the multiplicity of explanations and the fact that we may choose to generate multiple counterfactuals and traverse a latent encoding $\mathcal{Z}$ of the feature space $\mathcal{X}$ where we denote $f^{-1} : \mathcal{X} \mapsto \mathcal{Z}$. Encodings may involve simple feature transformations or more advanced techniques involving generative models, as we will discuss further below. The baseline approach, which we will simply refer to as **Wachter** [31], searches a single counterfactual directly in the feature space and penalises its distance to the original factual.

Solutions to Equation 1 are considered valid as soon as the predicted label matches the target label. A stripped-down counterfactual explanation is therefore little different from an adversarial example. In Figure 1, for example, we have applied Wachter to MNIST data (centre panel) where the underlying classifier M_θ is a simple Multi-Layer Perceptron (MLP) with above 90 percent test accuracy. For the generated counterfactual $\mathbf{x}'$ the model predicts the target label with high confidence (centre panel in Figure 1). The explanation is valid by definition, even though it looks a lot like an Adversarial Example [6]. Schut et al. [25] make the connection between Adversarial Examples and Counterfactual Explanations explicit and propose using a Jacobian-Based Saliency Map Attack (JSMA) to solve Equation 1. They demonstrate that this approach yields realistic and sparse counterfactuals for Bayesian, adversarially robust classifiers. Applying their approach to our simple MNIST classifier does not yield a realistic counterfactual but this one, too, is valid (right panel in Figure 1).

2.2 From Adversial Examples to Plausible Explanations

The crucial difference between Adversarial Examples (AE) and Counterfactual Explanations is one of intent. While an AE is intended to go unnoticed, a CE should have certain desirable properties. The literature has made this explicit by introducing various so-called *desiderata* that counterfactuals should

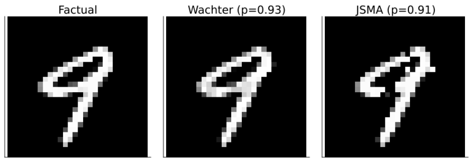

Figure 1: Explanations or Adversarial Examples? Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left); counterfactual produced using Wachter et al. [31] (centre); and a counterfactual produced using the approach introduced by [25] that uses Jacobian-Based Saliency Map Attacks to solve Equation 1.

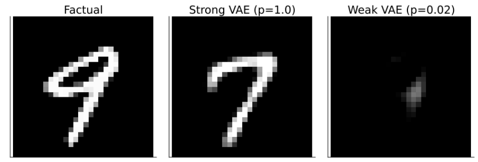

Figure 2: Using surrogates can improve plausibility, but also increases vulnerability. Counterfactuals for turning an 9 (nine) into a 7 (seven): original image (left); counterfactual produced using REVISE [9] with a well-specified surrogate (centre); and a counterfactual produced using REVISE [9] with a poorly specified surrogate (right).

meet in order to properly serve both AI practitioners and individuals affected by AI decision-making systems. The list of desiderate includes but is not limited to the following: sparsity, proximity [31], actionability [29], diversity [19], plausibility [9, 23, 25], robustness [28, 22, 2] and causality [12].

Researchers have come up with various ways to meet these desiderata, which have been extensively surveyed and evaluated in various studies [30, 11, 21, 4, 8]. Perhaps unsurprisingly, the different desiderata are often positively correlated. For example, Artelt et al. [4] 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 [15].

2.2.1 Plausibility through Surrogates

Arguably, the plausibility of counterfactuals has been among the primary concerns and some have focused explicitly on this goal. Joshi et al. [9], for example, were among the first to 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 data generating process (DGP) of $\mathbf{x} \sim \mathcal{X}$. To learn the latent embedding, they introduce a surrogate model. In particular, they propose to use the latent embedding of a Variational Autoencoder (VAE) trained to generate samples $\mathbf{x}^* \leftarrow \mathcal{G}(\mathbf{z})$ where $\mathcal{G}$ denotes the decoder part of the VAE. Provided the surrogate model is well-trained, their proposed approach —**REVISE**— can yield plausible explanations like the one in the centre panel of Figure 2.

Others have proposed similar approaches. Dombrowski et al. [5] traverse the base space of a normalizing flow to solve Equation 1, essentially relying on a different surrogate model for the generative task. Poyiadzi et al. [23] use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to constrain the counterfactuals to dense regions in the feature space. Karimi et al. [12] argue that counterfactuals should comply with the causal model that generates the data. All of these different approaches share a common goal: ensuring that the generated counterfactuals comply with the true and unobserved DGP. To summarize this broad objective, we propose the following definition:

Definition 2.1 (Plausible Counterfactuals)**.** *Let $\mathcal{X}|\mathbf{y}^+$ 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:* $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.

Surrogate models offer an obvious solution to achieve this objective. Unfortunately, surrogates also introduce a dependency: the generated explanations no longer depend exclusively on the black-box model itself, but also on the surrogate model. This is not necessarily problematic if the primary objective is not to explain the behaviour of the model but to offer recourse to individuals affected by it. It may become problematic even in this context if the dependency turns into a vulnerability. To illustrate this point, we have used REVISE [9] with an underfitted VAE to generate the counterfactual in the right panel of Figure 2: in this case, the decoder step of the VAE fails to yield plausible values ($\{\mathbf{x}' \leftarrow \mathcal{G}(\mathbf{z})\} \not\sim \mathcal{X}|\mathbf{y}^+$) and hence the counterfactual search in the learned latent space is doomed.

2.2.2 Plausibility through Minimal Predictive Uncertainty

Schut et al. [25] 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 solves Equation 1 by greedily applying JSMA in the feature space with standard cross-entropy loss and no penalty at all. They 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.

2.3 From Fidelity to Model Faithfulness

Above we explained that since Counterfactual Explanations work directly with the Black Box model, the fidelity of explanations as we defined it earlier is not a concern. This may explain why research has primarily focused on other desiderata, most notably plausibility (Definition 2.1). Enquiring about the plausibility of a counterfactual essentially boils down to the following question: 'Is this counterfactual consistent with the data'? We ask a related question: 'Is this counterfactual consistent with what the model has learned about the data'? To answer this question and propose a novel way to assess if explanations conform with model behaviour.

The word *fidelity* stems from the Latin word 'fidelis', which means 'faithful, loyal, trustworthy' [17]. As we explained in Section 2, model explanations are generally considered faithful if their corresponding predictions coincide with the predictions made by the model itself. Since this definition of faithfulness is not useful in the context of CE, we propose the following definition:

Definition 2.2 (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 conformal counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^+$.*

To assess counterfactuals with respect to Definition 2.2, we need to be able to quantify the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y}^+)$. This is very much at the core of our proposed methodological framework, which we will introduce next.

3 Methodological Framework

The primary objective of this work has been to develop a methodology for generating maximally plausible counterfactuals under minimal intervention. Our proposed framework is based on the premise that explanations should be plausible but not plausible at all costs. Energy-Constrained Conformal Counterfactuals (ECCCo) achieve this goal in two ways: firstly, they rely on the Black Box itself for the generative task; and, secondly, they involve an approach to predictive uncertainty quantification that is model-agnostic.

3.1 Quantifying the Model's Generative Property

Recent work by Grathwohl et al. [7] on Energy-Based Models (EBM) has shown that there is a 'generative model hidden within every standard discriminative model'. They use Stochastic Gradient Langevin Dynamics (SGLD) to estimate the posterior conditional distribution $p_\theta(\mathbf{x}|\mathbf{y})$ using and leverage this to train classifiers jointly for the discriminative and generative task.

Crucially for our purpose, SGLD can be applied during inference to essentially any standard discriminative model. Even models that are not explicitly trained for the joint objective learn about the distribution of inputs X by learning to make conditional predictions about the output y. We can leverage this observation to quantify the generative property of models. In particular, note that if we fix $\mathbf{y}$ to our target value $\mathbf{y}^+$, we can sample from $p_\theta(\mathbf{x}|\mathbf{y}^+)$ using 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. The term $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^+)$ denotes the model energy conditioned on the target class label $\mathbf{y}^+$. Generating multiple samples in this manner yields an empirical distribution $\hat{\mathbf{X}}_{\theta,\mathbf{y}^+}$ that we use in our search for plausible counterfactuals, as discussed in more detail below. Appendix A provides additional implementation details for any tasks related to energy-based modelling.

3.2 Quantifying the Model's Predictive Uncertainty

To quantify the model's predictive uncertainty we use Conformal Prediction (CP), an approach that has recently gained popularity in the Machine Learning community [3, 16]. 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 therefore 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. [27] in the context of conformal training:

$$\Omega(C_\theta(\mathbf{x};\alpha)) = \max\left(0, \sum_{\mathbf{y}\in\mathcal{Y}} C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) - \kappa\right) \tag{3}$$

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. Details concerning our implementation of Conformal Prediction can be found in Appendix B.

3.3 Energy-Constrained Conformal Counterfactuals (ECCCo)

Our framework for generating ECCCos combines the ideas introduced in the previous two subsections. Formally, we extend Equation 1 as follows,

$$\begin{aligned}\mathbf{Z}' = \arg\min_{\mathbf{Z}'\in\mathcal{Z}^M} \{&\text{yloss}(M_\theta(f(\mathbf{Z}')),\mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'),\mathbf{x}) \\ &+ \lambda_2 \text{dist}(f(\mathbf{Z}'),\hat{\mathbf{x}}_\theta) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}');\alpha))\}\end{aligned} \tag{4}$$

where $\hat{\mathbf{x}}_\theta$ denotes samples generated using SGLD (Equation 2) and $\text{dist}(\cdot)$ is a generic term for a distance metric. Our default choice for $\text{dist}(\cdot)$ is the L1 Norm since it induces sparsity.

The first two terms in Equation 4 correspond to the counterfactual search objective defined in Wachter et al. [31] which merely penalises the distance of counterfactuals from their factual values. The additional two penalties in ECCCo ensure that counterfactuals conform with the model's generative property and lead to minimally uncertain predictions, respectively. The hyperparameters $\lambda_1, ..., \lambda_3$ can be used to balance the different objectives: for example, we may choose to incur larger deviations from the factual in favour of faithfulness with the model's generative property by choosing lower values of λ_1 and relatively higher values of λ_2. Figure 3 illustrates this balancing act for an example involving synthetic data: vector fields indicate the direction of gradients with respect to the different components our proposed objective function (Equation 4).

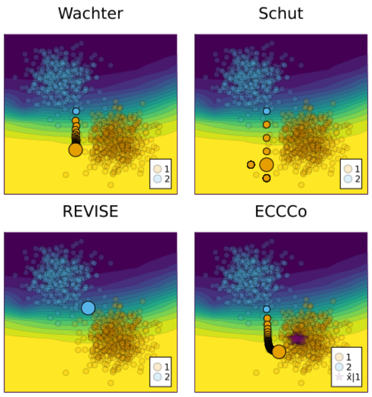

Figure 3: [PLACEHOLDER; may swap for:] Vector fields indicating the direction of gradients with respect to the different components of the ECCCo objective (Equation 4).

Algorithm 1: Generating ECCCos (For more details, see Appendix C)

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda, \alpha, \mathcal{D}, T, \eta, n_\mathcal{B}, N_\mathcal{B}$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: $\mathbf{x}'$

Initialize $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$
Generate buffer $\mathcal{B}$ of $N_\mathcal{B}$ conditional samples $\hat{\mathbf{x}}_\theta|\mathbf{y}^+$ using SGLD (Equation 2)
Run *SCP* for M_θ using $\mathcal{D}$
Initialize $t \leftarrow 0$
while *not converged* or $t < T$ **do**
  $\hat{\mathbf{x}}_{\theta,t} \leftarrow \text{rand}(\mathcal{B}, n_\mathcal{B})$
  $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{x}}_{\theta,t}; \Lambda, \alpha)$
  $t \leftarrow t + 1$
end while
$\mathbf{x}' \leftarrow f(\mathbf{z}')$

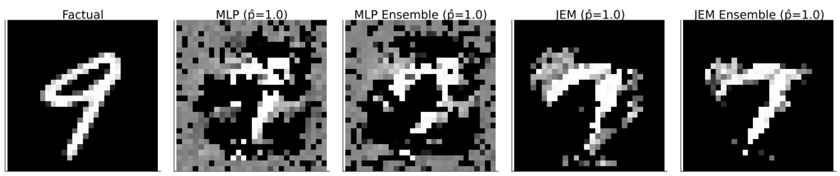

Figure 4: Original image (left) and ECCCos for turning a 9 (nine) into a 7 (seven) for different black-box models from left to right: Multi-Layer Perceptron (MLP), Ensemble of MLPs, Joint Energy Model (JEM), Ensemble of JEMs.

The entire procedure for Generating ECCCos is described in Algorithm 1. For the sake of simplicity and without loss of generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$ where in contrast to Equation 4 $\mathbf{z}'$ denotes a 1-dimensional array containing a single counterfactual state. That state is initialized by passing the factual $\mathbf{x}$ through the encoder f^{-1} which in our case corresponds to a simple feature transformer, rather than the encoder part of VAE as in REVISE [9]. Next, we generate a buffer of $N_\mathcal{B}$ conditional samples $\hat{\mathbf{x}}_\theta|\mathbf{y}^+$ using SGLD (Equation 2) and conformalise the model M_θ through Split Conformal Prediction on training data $\mathcal{D}$.

Finally, we search counterfactuals through gradient descent. Let $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{x}}_{\theta,t}; \Lambda, \alpha)$ denote our loss function defined in Equation 4. Then in each iteration, we first randomly draw $n_\mathcal{B}$ samples from the buffer $\mathcal{B}$ before updating the counterfactual state $\mathbf{z}'$ by moving in the negative direction of that loss function. 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 (for more detail on this see Appendix C).

Figure 4 presents ECCCos for the MNIST example from Section 2 for various black-box models of increasing complexity from left to right: a simple Multi-Layer Perceptron (MLP); an Ensemble of MLPs, each of the same architecture as the single MLP; a Joint Energy Model (JEM) based on the same MLP architecture; and finally, an Ensemble of these JEMs. Since Deep Ensembles have an improved capacity for predictive uncertainty quantification and JEMs are explicitly trained to learn plausible representations of the input data, it is intuitive to see that the plausibility of counterfactuals visibly improves from left to right. This provides some first anecdotal evidence that ECCCos achieve plausibility while maintaining faithfulness to the Black Box.

4 Empirical Analysis

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

Research Question 4.1 (Feasibility)**.** *Is it feasible to generate plausible Counterfactual Explanations through ECCCo without relying on surrogate models?*

Research Question 4.2 (Drivers). *Subject to feasibility, what drives the performance of ECCCo? Is it sufficient to rely on energy-based modelling to quantify the model's generative property? Is it sufficient to rely on conformal prediction to quantify the model's uncertainty?*

We first briefly describe our experimental setup, before presenting our main results.

4.1 Key Evaluation Metrics

Above we have defined plausibility (Definition 2.1) and faithfulness (Definition 2.2) for Counterfactual Explanations. These are the main criteria we use to evaluate counterfactuals in this study. In order to quantify the plausibility of counterfactuals we use a slightly adapted version of the implausibility metric proposed in Guidotti [8]. Formally, we define implausibility as follows,

$$\text{impl} = \frac{1}{|\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}|} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{5}$$

where $\mathbf{X}_{\mathbf{y}^+}$ is a subsample of the training data in the target class $\mathbf{y}^+$. This gives rise to a very 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:

$$\text{unfaith} = \frac{1}{|\mathbf{x} \in \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}|} \sum_{\mathbf{x} \in \hat{\mathbf{X}}^{n_E}_{\theta,\mathbf{y}^+}} \text{dist}(\mathbf{x}', \mathbf{x}) \tag{6}$$

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

While we focus on these key evaluation metrics in the body of this paper, we also sporadically discuss outcomes with respect to other common measures used to evaluate the validity, proximity and sparsity of counterfactuals. Details can be found in Appendix E.

4.2 Experimental Setup

To assess and benchmark the performance of ECCCo against the state of the art, we generate multiple counterfactuals for different black-box models and datasets. In particular, we compare ECCCo to the following counterfactual generators that were introduced above: firstly; **Schut** [25], which minimizes predictive uncertainty; secondly, **REVISE** [9], which uses a VAE as its surrogate model; and, finally, **Wachter** [31], which serves as our baseline.

We use both synthetic and real-world datasets from different domains, all of which are publically available and commonly used to train and benchmark classification algorithms. The synthetic datasets include: a dataset containing two **Linearly Separable** Gaussian clusters ($n = 1000$), as well as the well-known **Circles** ($n = 1000$) and **Moons** ($n = 2500$) data. As for real-world data, we follow Schut et al. [25] and use the **MNIST** [13] dataset containing images of handwritten digits such as the examples shown above. From the social sciences domain, we include Give Me Some Credit (**GMSC**) [10]: a tabular dataset that has been studied extensively in the literature on Algorithmic Recourse [21]. It consists of 11 numeric features that can be used to predict the binary outcome variable indicating whether or not retail borrowers experience financial distress.

As with the example in Section 3, 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. To account for stochasticity, we generate multiple counterfactuals for each possible 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.

4.3 Results

Table 1 shows the key results for the synthetic datasets separated by model (first columns) and generator (second column). The numerical columns show the average values of our key evaluation

Table 1: Results for synthetic datasets. Standard deviations across samples are shown in parentheses. 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	0.10 (0.06)**	0.19 (0.03)**	**0.57 (0.58)****	**1.29 (0.21)***	**0.63 (1.58)**	1.44 (1.37)
	ECCCo (no CP)	**0.10 (0.07)****	**0.19 (0.03)****	0.63 (0.64)*	1.30 (0.21)*	0.64 (1.61)	1.45 (1.38)
	ECCCo (no EBM)	0.37 (0.28)	0.38 (0.26)	1.73 (1.34)	1.73 (1.42)	1.41 (1.51)	1.50 (1.38)
	REVISE	0.41 (0.02)**	0.41 (0.01)**	1.59 (0.55)	1.55 (0.20)	0.96 (0.32)*	**0.95 (0.32)***
	Schut	0.66 (0.23)	0.66 (0.22)	1.55 (0.61)	1.42 (0.16)*	0.99 (0.80)	1.28 (0.53)
	Wachter	0.44 (0.16)	0.44 (0.15)	1.77 (0.48)	1.67 (0.15)	1.41 (1.50)	1.51 (1.35)
MLP	ECCCo	**0.03 (0.02)****	0.69 (0.10)	1.68 (1.74)	2.02 (0.86)	**0.37 (0.65)****	1.30 (0.68)
	ECCCo (no CP)	**0.03 (0.02)****	0.68 (0.10)	**1.34 (1.66)**	2.11 (0.88)	0.50 (0.85)*	1.28 (0.66)
	ECCCo (no EBM)	1.25 (0.87)	1.84 (1.10)	2.98 (1.89)	2.29 (1.75)	2.00 (1.46)	1.83 (1.00)
	REVISE	1.10 (0.10)	**0.40 (0.01)****	2.46 (1.05)	**1.54 (0.27)***	1.16 (1.05)	**0.95 (0.32)***
	Schut	0.81 (0.10)*	0.47 (0.24)	2.71 (1.15)	1.62 (0.42)	1.60 (1.15)	1.24 (0.44)
	Wachter	0.94 (0.11)	0.44 (0.15)	2.95 (1.42)	1.84 (1.33)	1.67 (1.05)	1.31 (0.43)

metrics computed across all counterfactuals. Standard deviations are shown in parentheses. In bold we have highlighted the best outcome for each model and metric. To provide some sense of the statistical significance of our findings, 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 not surprising, since ECCCo directly enforces faithfulness through regularization. Crucially though, ECCCo also produces the most plausible counterfactuals for the Joint Energy Model, which was explicitly trained to learn plausible representations of the input data. This high-level pattern is broadly consistent across all 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.

Zooming in on the granular details for the Linearly Separable data, note that the list of generators in Table 1 includes 'ECCCo (no CP)' and 'ECCCo (no EBM)' in addition to 'ECCCo' and our benchmark generators. These have been added to gain some sense of the degree to which the two components underlying ECCCo—namely energy-based modelling (EBM) and conformal prediction (CP)—drive the results. Specifically, 'ECCCo (no CP)' involves no set size penalty ($\lambda_3 = 0$ in Equation 4), while 'ECCCo (no EBM)' does not penalise the distance to samples generated through SGLD ($\lambda_2 = 0$ in Equation 4). The corresponding results indicate that the positive results are dominated by the effect of quantifying and leveraging the model's generative property (EBM) in our search for counterfactuals. Conformal Prediction alone only leads to marginally improved faithfulness and plausibility relative to the benchmark generators for our JEM. As a final observation for the Linearly Separable data we note that for the MLP, increased faithfulness comes at the cost of reduced plausibility. Specifically, this means that counterfactuals generated through ECCCo end up further away from individuals in the target class than those produced by our benchmark generators.

The findings for the Moons dataset are broadly in line with the findings so far: for the JEM, ECCCo yields significantly more faithful and plausible counterfactuals than all other generators. For the MLP, faithfulness is maintained but counterfactuals are not plausible. By comparison, REVISE yields fairly plausible counterfactuals in both cases, 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, the most faithful counterfactuals are generated by ECCCo. While it appears that REVISE generates the most plausible counterfactuals in this case, we note that they are valid only half of the time (see Appendix E for a complete overview of all evaluation metrics). It turns out that in this case, the underlying VAE with default parameters has not adequately learned the data-generating process. Of course, it is possible to achieve better generative performance through hyperparameter tuning. But this example serves to illustrate that REVISE depends strictly on the quality of the surrogate model. Independent of the outcome for REVISE, however, the results do not seem to indicate that ECCCo significantly improves our plausibility metric for the Circles data.

Table 2: Results for real-world datasets. Standard deviations across samples are shown in parentheses. 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	**116.09 (30.70)****	281.33 (41.51)**	**41.65 (17.24)****	40.57 (8.74)**
	REVISE	348.74 (65.65)**	**246.69 (36.69)****	74.89 (15.82)**	**6.01 (5.75)****
	Schut	355.58 (64.84)**	270.06 (40.41)**	76.23 (15.54)**	6.02 (0.72)**
	Wachter	694.08 (50.86)	630.99 (33.01)	146.02 (64.48)	128.93 (74.00)
JEM Ensemble	ECCCo	**89.89 (27.26)****	240.59 (37.41)**	**26.55 (12.94)****	33.65 (8.33)**
	REVISE	292.52 (53.13)**	**240.50 (35.73)****	52.47 (14.12)**	6.69 (3.37)**
	Schut	319.45 (59.02)**	266.80 (40.46)**	56.34 (15.00)**	**6.27 (1.06)****
	Wachter	582.52 (58.46)	543.90 (44.24)	125.72 (70.80)	126.55 (93.75)
MLP	ECCCo	**212.45 (36.70)****	649.63 (58.80)	**46.90 (15.80)****	37.78 (8.40)**
	REVISE	839.79 (77.14)*	**244.33 (38.69)****	81.08 (19.53)**	**4.60 (0.72)****
	Schut	842.80 (82.01)*	264.94 (42.18)**	90.67 (20.80)**	5.56 (0.81)**
	Wachter	982.32 (61.81)	561.23 (45.08)	191.68 (30.86)	200.23 (15.05)
MLP Ensemble	ECCCo	**162.21 (36.21)****	587.65 (95.01)	**74.65 (144.69)***	71.87 (145.19)
	REVISE	741.30 (125.98)*	**242.76 (41.16)****	80.90 (14.59)**	**5.20 (1.52)****
	Schut	754.35 (132.26)	266.94 (42.55)**	85.63 (19.15)**	6.00 (0.99)**
	Wachter	871.09 (92.36)	536.24 (48.73)	220.05 (17.41)	203.65 (14.77)

Moving on to our real-world datasets, the results 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 corresponding counterfactuals produced by any of the benchmark generators and their plausibility gradually improves through ensembling and joint-energy modelling. For the JEM Ensemble, ECCCo is essentially on par with REVISE and does significantly better than the baseline generator. We also note that ECCCo is the only generator that consistently achieves full validity for all models (Appendix E). Interestingly, ECCCo also yields lower-cost outcomes than the baseline generator for the JEMs.

For the tabular credit dataset (GMSC) we have struggled to get good generative and discriminative performance for our JEMs. Consequently, it is not surprising to find that ECCCo never achieves state-of-the-art plausibility, although it does improve outcomes compared to the baseline (Wachter). Concerning faithfulness, ECCCo once again consistently outperforms all other generators.

To conclude this section, we summarize our findings with reference to the opening questions. Concerning the feasibility of our proposed methodology (Research Question 4.1), our findings demonstrate that it is indeed possible to generate plausible counterfactuals without the need for surrogate models. A related important finding is that ECCCo never sacrifices faithfulness for plausibility: any plausible ECCCo also faithfully describes model behaviour. This mitigates the risk of generating plausible explanations for models that are, in fact, highly susceptible to implausible counterfactuals as well. Our findings here indicate that ECCCo achieves this result primarily by leveraging the model's generative property. We think that further work is needed, however, to definitively answer Research Question 4.2, on which we elaborate in the following section.

5 Limitations

Even though we have taken considerable measures to study our proposed methodology carefully, this work is limited in scope, which caveats our findings. 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. This choice is associated with relatively higher costs to individuals since the proposed recourses typically involve more substantial feature changes than for our benchmark generators.

Conversely, we have not found that penalising prediction set sizes disproportionately strongly had any discernable effect on our results. As discussed above, we also struggled to achieve good results by relying on conformal prediction alone. We want to caveat this finding by acknowledging that the role of CP in this context needs to be investigated more thoroughly through future work. Our suggested approach involving a smooth set size penalty may be insufficient in this context.

The fact that our findings are primarily driven by applying ideas from energy-based modelling presents a challenge in itself: while our approach is readily applicable to models with gradient access like deep neural networks, more work is needed to generalise our methodology to other popular machine learning models such as gradient-boosted trees. Relatedly, we have encountered common challenges associated with energy-based modelling during our experiments including sensitivity to scale, training instabilities and sensitivity to hyperparameters. We have also struggled to apply our proposed approach to low-dimensional tabular data.

6 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 Counterfactual Explanations that are maximally faithful to the black-model they aim to explain. Our proposed counterfactual generator, ECCCo, produces plausible counterfactual if and only if the black-model itself has learned realistic representations of the data. 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] Patrick Altmeyer. Conformal Prediction in Julia. URL `https://www.paltmeyer.com/blog/posts/conformal-prediction/`.

[2] 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.

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

[4] 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.

[5] 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.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

[7] 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`.

[8] 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`.

[9] 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.

[10] 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`.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] Valery Manokhin. Awesome conformal prediction.

[17] Merriam-Webster. "fidelity". URL `https://www.merriam-webster.com/dictionary/fidelity`.

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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`.

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

[29] 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.

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

[31] 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.

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

Appendices

A JEM

While $\mathbf{x}_J$ is only guaranteed to distribute as $p_\theta(\mathbf{x}|\mathbf{y}^+)$ if $\epsilon \to 0$ and $J \to \infty$, the bias introduced for a small finite ϵ is negligible in practice [20, 7]. While Grathwohl et al. [7] use Equation 2 during training, we are interested in applying the conditional sampling procedure in a post-hoc fashion to any standard discriminative model.

B Conformal Prediction

The fact that conformal classifiers produce set-valued predictions introduces a challenge: it is not immediately obvious how to use such classifiers in the context of gradient-based counterfactual search. Put differently, it is not clear how to use prediction sets in Equation 1. Fortunately, Stutz et al. [27] have recently proposed a framework for Conformal Training that also hinges on differentiability. Specifically, they show how Stochastic Gradient Descent can be used to train classifiers not only for the discriminative task but also for additional objectives related to Conformal Prediction. One such objective is *efficiency*: for a given target error rate α, the efficiency of a conformal classifier improves as its average prediction set size decreases. To this end, the authors introduce a smooth set size penalty defined in Equation 3 in the body of this paper

Formally, it is defined as $C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) := \sigma\left((s(\mathbf{x}_i,\mathbf{y}) - \alpha)T^{-1}\right)$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling [27].

Intuitively, CP works under the premise of turning heuristic notions of uncertainty into rigorous uncertainty estimates by repeatedly sifting through the data. It can be used to generate prediction intervals for regression models and prediction sets for classification models [1]. Since the literature on CE and AR is typically concerned with classification problems, we focus on the latter. A particular variant of CP called Split Conformal Prediction (SCP) is well-suited for our purposes, because it imposes only minimal restrictions on model training.

Specifically, SCP involves splitting the data $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n}$ into a proper training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. The former is used to train the classifier in any conventional fashion. The latter is then used to compute so-called nonconformity scores: $\mathcal{S} = \{s(\mathbf{x}_i, \mathbf{y}_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ where $s : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}$ is referred to as *score function*. In the context of classification, a common choice for the score function is just $s_i = 1 - M_\theta(\mathbf{x}_i)[\mathbf{y}_i]$, that is one minus the softmax output corresponding to the observed label $\mathbf{y}_i$ [3].

Finally, classification sets are formed as follows,

$$C_\theta(\mathbf{x}_i;\alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \tag{7}$$

where $\hat{q}$ denotes the $(1-\alpha)$-quantile of $\mathcal{S}$ and α is a predetermined error rate. As the size of the calibration set increases, the probability that the classification set $C(\mathbf{x}_{\text{test}})$ for a newly arrived sample $\mathbf{x}_{\text{test}}$ does not cover the true test label $\mathbf{y}_{\text{test}}$ approaches α [3].

Observe from Equation 7 that Conformal Prediction works on an instance-level basis, much like Counterfactual Explanations are local. The prediction set for an individual instance $\mathbf{x}_i$ depends only on the characteristics of that sample and the specified error rate. Intuitively, the set is more likely to include multiple labels for samples that are difficult to classify, so the set size is indicative of predictive uncertainty. To see why this effect is exacerbated by small choices for α consider the case of $\alpha = 0$, which requires that the true label is covered by the prediction set with probability equal to 1.

C Conformal Prediction

D Experimental Setup

E Results

Table 3: All results for all datasets. Standard deviations across samples are shown in parentheses. 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).

Model	Data	Generator	Cost ↓	Unfaithfulness ↓	Implausibility ↓	Redundancy ↑	Uncertainty ↓	Validity ↑
California Housing	JEM	ECCCo	39.14 (3.71)	**236.79 (51.16)**	39.78 (3.18)	0.00 (0.00)	2.00 (0.00)	1.00 (0.00)
		REVISE	4.39 (2.08)	284.51 (52.74)	**5.58 (0.81)****	0.01 (0.03)	**1.85 (0.32)**	1.00 (0.00)
		Schut	4.17 (1.84)	263.55 (60.56)	8.00 (2.03)	**0.25 (0.24)***	1.88 (0.31)	1.00 (0.00)
		Wachter	**2.03 (1.01)**	274.55 (51.17)	7.32 (1.80)	0.00 (0.00)	1.90 (0.31)	1.00 (0.00)
	JEM Ensemble	ECCCo	34.85 (4.67)	**249.44 (58.53)**	35.09 (5.56)	0.00 (0.00)	2.00 (0.00)	1.00 (0.00)
		REVISE	4.53 (1.97)	268.45 (66.87)	**5.44 (0.74)****	0.00 (0.00)	1.95 (0.21)	1.00 (0.00)
		Schut	**0.98 (0.38)****	279.38 (63.23)	7.64 (1.47)	**0.84 (0.06)****	2.00 (0.00)	1.00 (0.00)
		Wachter	2.00 (0.59)	268.59 (68.66)	7.16 (1.46)	0.00 (0.00)	**1.90 (0.31)**	1.00 (0.00)
	MLP	ECCCo	37.47 (4.59)	**230.92 (48.86)**	37.53 (5.40)	0.00 (0.00)	1.00 (0.00)**	1.00 (0.00)
		REVISE	3.38 (2.06)	281.10 (53.01)	**5.34 (0.67)****	0.00 (0.00)	1.10 (0.31)	1.00 (0.00)
		Schut	**0.88 (0.51)****	285.12 (56.00)	6.48 (1.18)**	**0.72 (0.22)****	**1.00 (0.00)****	1.00 (0.00)
		Wachter	5.35 (10.88)	262.50 (56.87)	9.21 (10.41)	0.00 (0.00)	1.05 (0.22)	1.00 (0.00)
	MLP Ensemble	ECCCo	38.33 (4.99)	**212.47 (59.27)***	38.17 (6.18)	0.00 (0.00)	1.00 (0.00)**	1.00 (0.00)
		REVISE	3.41 (1.79)	284.65 (49.52)	**5.64 (1.13)***	0.00 (0.00)	1.05 (0.22)	1.00 (0.00)
		Schut	**0.84 (0.56)****	269.19 (46.08)	7.30 (1.94)	**0.81 (0.11)****	**1.00 (0.00)****	1.00 (0.00)
		Wachter	2.00 (1.39)	278.09 (73.65)	7.32 (1.75)	0.00 (0.00)	1.07 (0.23)	1.00 (0.00)
Circles	JEM	ECCCo	1.34 (1.48)	**0.63 (1.58)**	1.44 (1.37)	0.00 (0.00)	0.98 (0.14)	0.98 (0.14)
		ECCCo (no CP)	1.33 (1.49)	0.64 (1.61)	1.45 (1.38)	0.00 (0.00)	0.98 (0.14)	0.98 (0.14)
		ECCCo (no EBM)	0.85 (1.49)	1.41 (1.51)	1.50 (1.38)	0.00 (0.00)	1.04 (0.28)	0.98 (0.14)
		REVISE	0.99 (0.35)	0.96 (0.32)*	**0.95 (0.32)***	0.00 (0.00)	**0.50 (0.51)**	0.50 (0.51)
		Schut	1.00 (0.43)	0.99 (0.80)	1.28 (0.53)	**0.25 (0.25)**	1.11 (0.38)	**1.00 (0.00)****
		Wachter	**0.74 (1.50)**	1.41 (1.50)	1.51 (1.35)	0.00 (0.00)	0.98 (0.14)	0.98 (0.14)
	MLP	ECCCo	1.39 (0.23)	**0.37 (0.65)****	1.30 (0.68)	0.00 (0.00)	1.00 (0.00)**	**1.00 (0.00)**
		ECCCo (no CP)	1.33 (0.28)	0.50 (0.85)*	1.28 (0.66)	0.00 (0.00)	1.04 (0.20)*	**1.00 (0.00)**
		ECCCo (no EBM)	1.15 (0.69)	2.00 (1.46)	1.83 (1.00)	0.00 (0.00)	0.97 (0.10)**	**1.00 (0.00)**
		REVISE	0.98 (0.36)	1.16 (1.05)	**0.95 (0.32)***	0.00 (0.00)	**0.50 (0.51)***	0.50 (0.51)
		Schut	0.61 (0.11)	1.60 (1.15)	1.24 (0.44)	**0.34 (0.24)***	1.00 (0.00)**	**1.00 (0.00)**
		Wachter	**0.53 (0.15)**	1.67 (1.05)	1.31 (0.43)	0.00 (0.00)	1.28 (0.46)	**1.00 (0.00)**
FashionMNIST	JEM	ECCCo	859.68 (91.05)	**40.65 (5.67)****	605.67 (19.56)	0.00 (0.00)	3.00 (0.00)**	**1.00 (0.00)**
		REVISE	500.28 (86.07)	693.81 (118.47)*	**467.88 (132.24)**	0.00 (0.00)	3.20 (2.28)**	0.80 (0.45)
		Schut	**10.00 (0.00)****	871.82 (64.75)	561.81 (94.76)	**0.99 (0.00)****	**0.00 (0.00)****	0.00 (0.00)
		Wachter	100.86 (13.85)	902.84 (88.79)	586.49 (97.17)	0.00 (0.00)	10.00 (0.00)	**1.00 (0.00)**
	JEM Ensemble	ECCCo	679.19 (66.95)	**59.61 (32.93)****	500.50 (27.51)	0.00 (0.00)	4.00 (0.00)**	**1.00 (0.00)**
		REVISE	476.47 (147.09)	533.64 (102.81)*	**356.60 (79.57)***	0.00 (0.00)	4.80 (1.30)**	**1.00 (0.00)**
		Schut	**10.00 (0.00)****	688.61 (86.83)	445.55 (99.03)	**0.99 (0.00)****	**0.00 (0.00)****	0.00 (0.00)
		Wachter	92.50 (9.31)	714.63 (54.58)	470.54 (96.18)	0.00 (0.00)	10.00 (0.00)	**1.00 (0.00)**
	MLP	ECCCo	885.97 (29.70)	**65.36 (20.64)****	791.07 (14.51)	0.00 (0.00)	2.00 (0.00)**	**1.00 (0.00)****
		REVISE	323.10 (102.63)	856.08 (73.66)	**394.73 (252.67)**	0.00 (0.00)	1.00 (1.00)**	0.60 (0.55)
		Schut	**10.00 (0.00)****	928.77 (42.27)	518.98 (143.30)	**0.99 (0.00)****	**0.00 (0.00)****	0.00 (0.00)
		Wachter	94.57 (10.26)	916.45 (50.09)	546.35 (145.24)	0.00 (0.00)	3.61 (4.01)	0.80 (0.45)
	MLP Ensemble	ECCCo	869.65 (67.92)	**47.37 (7.72)****	751.83 (11.87)	0.00 (0.00)	1.00 (0.00)**	**1.00 (0.00)**
		REVISE	267.88 (69.67)	822.34 (57.55)	**307.50 (105.09)***	0.00 (0.00)	3.00 (4.00)	0.80 (0.45)
		Schut	**10.00 (0.00)****	891.57 (70.10)	449.79 (149.32)	**0.99 (0.00)****	**0.00 (0.00)****	0.00 (0.00)
		Wachter	91.50 (16.35)	874.21 (59.36)	476.59 (150.76)	0.00 (0.00)	4.60 (4.93)	**1.00 (0.00)**
GMSC	JEM	ECCCo	40.78 (8.79)**	**41.65 (17.24)****	40.57 (8.74)**	0.00 (0.00)	1.50 (0.51)	**1.00 (0.00)****
		REVISE	5.10 (6.48)**	74.89 (15.82)**	**6.01 (5.75)****	0.00 (0.00)	1.81 (0.40)	**1.00 (0.00)****
		Schut	**1.10 (0.39)****	76.23 (15.54)**	6.02 (0.72)**	**0.77 (0.09)****	1.55 (0.51)	**1.00 (0.00)****
		Wachter	127.26 (75.11)	146.02 (64.48)	128.93 (74.00)	0.00 (0.00)	**1.00 (1.03)**	0.50 (0.51)
	JEM Ensemble	ECCCo	33.87 (8.25)**	**26.55 (12.94)****	33.65 (8.33)**	0.00 (0.00)	2.00 (0.00)	**1.00 (0.00)****
		REVISE	6.00 (4.92)**	52.47 (14.12)**	6.69 (3.37)**	0.00 (0.00)	1.80 (0.52)	0.95 (0.22)**
		Schut	**1.29 (0.92)****	56.34 (15.00)**	**6.27 (1.06)****	**0.74 (0.16)****	1.62 (0.52)	**1.00 (0.00)****
		Wachter	124.35 (95.08)	125.72 (70.80)	126.55 (93.75)	0.00 (0.00)	**1.00 (1.03)**	0.50 (0.51)
	MLP	ECCCo	38.91 (7.68)**	**46.90 (15.80)****	37.78 (8.40)**	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		REVISE	4.16 (2.35)**	81.08 (19.53)**	**4.60 (0.72)****	0.00 (0.00)	1.23 (0.40)	1.00 (0.00)
		Schut	**0.72 (0.32)****	90.67 (20.80)**	5.56 (0.81)**	**0.87 (0.06)****	**1.00 (0.00)**	1.00 (0.00)
		Wachter	199.28 (14.78)	191.68 (30.86)	200.23 (15.05)	0.00 (0.00)	**1.00 (0.00)**	1.00 (0.00)
	MLP Ensemble	ECCCo	72.42 (145.72)	**74.65 (144.69)***	71.87 (145.19)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
		REVISE	4.75 (2.94)**	80.90 (14.59)**	**5.20 (1.52)****	0.00 (0.00)	1.07 (0.12)	1.00 (0.00)
		Schut	**0.65 (0.24)****	85.63 (19.15)**	6.00 (0.99)**	**0.88 (0.04)****	**1.00 (0.00)****	1.00 (0.00)
		Wachter	202.64 (14.71)	220.05 (17.41)	203.65 (14.77)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)
Linearly Separable	JEM	ECCCo	0.91 (0.14)	0.10 (0.06)**	0.19 (0.03)**	0.00 (0.00)	0.97 (0.03)**	**1.00 (0.00)**
		ECCCo (no CP)	0.91 (0.14)	**0.10 (0.07)****	**0.19 (0.03)****	0.00 (0.00)	0.98 (0.03)**	**1.00 (0.00)**
		ECCCo (no EBM)	0.90 (0.17)	0.37 (0.28)	0.38 (0.26)	0.00 (0.00)	1.23 (0.49)	**1.00 (0.00)**
		REVISE	**0.42 (0.14)***	0.41 (0.02)**	0.41 (0.01)**	0.00 (0.00)	**0.81 (0.82)**	0.50 (0.51)
		Schut	1.14 (0.27)	0.66 (0.23)	0.66 (0.22)	**0.21 (0.25)**	1.74 (0.43)	**1.00 (0.00)**
		Wachter	0.61 (0.12)	0.44 (0.16)	0.44 (0.15)	0.00 (0.00)	1.50 (0.50)	**1.00 (0.00)**
	MLP	ECCCo	1.52 (0.16)	**0.03 (0.02)****	0.69 (0.10)	0.00 (0.00)	1.00 (0.00)**	**1.00 (0.00)**
		ECCCo (no CP)	1.52 (0.16)	**0.03 (0.02)****	0.68 (0.10)	0.00 (0.00)	1.00 (0.00)**	**1.00 (0.00)**
		ECCCo (no EBM)	2.66 (1.10)	1.25 (0.87)	1.84 (1.10)	0.00 (0.00)	1.00 (0.00)**	**1.00 (0.00)**
		REVISE	**0.44 (0.13)***	1.10 (0.10)	**0.40 (0.01)****	0.00 (0.00)	1.64 (0.78)	0.82 (0.39)
		Schut	0.76 (0.14)	0.81 (0.10)*	0.47 (0.24)	**0.26 (0.25)***	**1.00 (0.00)****	**1.00 (0.00)**
		Wachter	0.60 (0.14)	0.94 (0.11)	0.44 (0.15)	0.00 (0.00)	1.54 (0.50)	**1.00 (0.00)**
MNIST	JEM	ECCCo	269.99 (57.02)**	**116.09 (30.70)****	281.33 (41.51)**	0.00 (0.00)	NA	**1.00 (0.00)****
		REVISE	143.79 (43.43)**	348.74 (65.65)**	**246.69 (36.69)****	0.00 (0.01)	NA	0.80 (0.40)
		Schut	**9.90 (0.55)****	355.58 (64.84)**	270.06 (40.41)**	**0.99 (0.00)****	NA	0.15 (0.36)
		Wachter	453.86 (16.96)	694.08 (50.86)	630.99 (33.01)	0.00 (0.00)	NA	0.90 (0.30)
	JEM Ensemble	ECCCo	260.94 (52.14)**	**89.89 (27.26)****	240.59 (37.41)**	0.00 (0.00)	NA	**1.00 (0.00)****
		REVISE	138.82 (33.99)**	292.52 (53.13)**	**240.50 (35.73)****	0.00 (0.01)	NA	0.81 (0.39)
		Schut	**9.97 (0.28)****	319.45 (59.02)**	266.80 (40.46)**	**0.99 (0.00)****	NA	0.05 (0.22)
		Wachter	365.46 (35.14)	582.52 (58.46)	543.90 (44.24)	0.00 (0.00)	NA	0.96 (0.20)
	MLP	ECCCo	658.48 (65.03)	**212.45 (36.70)****	649.63 (58.80)	0.00 (0.00)	NA	**1.00 (0.00)**
		REVISE	150.41 (51.81)**	839.79 (77.14)*	**244.33 (38.69)****	0.00 (0.00)	NA	0.95 (0.22)
		Schut	**9.95 (0.41)****	842.80 (82.01)*	264.94 (42.18)**	**0.99 (0.00)****	NA	0.06 (0.25)
		Wachter	400.08 (34.33)	982.32 (61.81)	561.23 (45.08)	0.00 (0.00)	NA	**1.00 (0.00)**
	MLP Ensemble	ECCCo	616.12 (102.01)	**162.21 (36.21)****	587.65 (95.01)	0.00 (0.00)	NA	**1.00 (0.00)****
		REVISE	149.48 (47.90)**	741.30 (125.98)*	**242.76 (41.16)****	0.00 (0.01)	NA	0.92 (0.27)
		Schut	**9.98 (0.23)****	754.35 (132.26)	266.94 (42.55)**	**0.99 (0.00)****	NA	0.03 (0.18)
		Wachter	374.37 (41.37)	871.09 (92.36)	536.24 (48.73)	0.00 (0.00)	NA	1.00 (0.05)
Moons	JEM	ECCCo	1.87 (0.79)	**0.57 (0.58)****	**1.29 (0.21)***	0.00 (0.00)	0.99 (0.18)**	1.00 (0.00)
		ECCCo (no CP)	1.83 (0.80)	0.63 (0.64)*	1.30 (0.21)*	0.00 (0.00)	1.13 (0.35)	1.00 (0.00)
		ECCCo (no EBM)	1.30 (1.72)	1.73 (1.34)	1.73 (1.42)	0.00 (0.00)	**0.94 (0.27)***	1.00 (0.00)
		REVISE	1.07 (0.26)	1.59 (0.55)	1.55 (0.20)	0.00 (0.00)	1.30 (0.40)	1.00 (0.00)
		Schut	1.36 (0.35)	1.55 (0.61)	1.42 (0.16)*	**0.03 (0.12)**	1.11 (0.30)*	1.00 (0.00)
		Wachter	**0.89 (0.21)**	1.77 (0.48)	1.67 (0.15)	0.00 (0.00)	1.45 (0.47)	1.00 (0.00)
	MLP	ECCCo	2.53 (1.24)	1.68 (1.74)	2.02 (0.86)	0.00 (0.00)	1.11 (0.31)	**1.00 (0.00)**
		ECCCo (no CP)	2.45 (1.36)	**1.34 (1.66)**	2.11 (0.88)	0.00 (0.00)	1.24 (0.41)	**1.00 (0.00)**
		ECCCo (no EBM)	2.53 (2.03)	2.98 (1.89)	2.29 (1.75)	0.00 (0.00)	0.99 (0.07)**	**1.00 (0.00)**
		REVISE	0.98 (0.33)*	2.46 (1.05)	**1.54 (0.27)***	0.00 (0.00)	1.40 (0.49)	**1.00 (0.00)**
		Schut	**0.75 (0.23)****	2.71 (1.15)	1.62 (0.42)	**0.31 (0.27)***	**0.94 (0.24)***	0.94 (0.24)
		Wachter	1.49 (1.76)	2.95 (1.42)	1.84 (1.33)	0.00 (0.00)	1.33 (0.48)	**1.00 (0.00)**