
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 Energy-Constrained Conformal Counterfactuals (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 [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].

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 without sacrificing plausibility (Section 6).

Thus, 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:

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

Here yloss denotes the primary loss function, $f(\cdot)$ is a function that maps from the counterfactual state space to the feature space 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. [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:

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: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.*

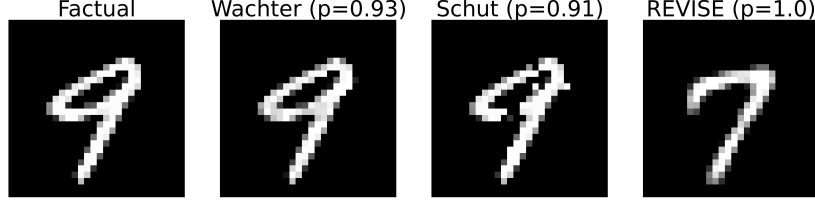


Figure 1: Counterfactuals for turning a 9 (nine) into a 7 (seven): original image (left); then from left to right the counterfactuals generated using *Wachter*, *Schut* and *REVISE*.

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

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

104 3 Why Fidelity is not Enough

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

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

119 Since *Wachter* is only concerned with proximity, the generated counterfactual is almost indistinguish-
120 able from the factual. The approach by Schut et al. [12] expects a well-calibrated model that can
121 generate predictive uncertainty estimates. Since this is not the case, the generated counterfactual
122 looks like an adversarial example. Finally, the counterfactual generated by *REVISE* looks much more
123 plausible than the other two. But is it also more faithful to the behaviour of our *MNIST* classifier?
124 That is much less clear because the surrogate used by *REVISE* introduces friction: the generated
125 explanations no longer depend exclusively on the black-box model itself.

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

129 4 A New Notion of Faithfulness

130 Considering the limitations of fidelity as demonstrated in the previous section, analogous to Defini-
 131 tion 2.1, we introduce a new notion of faithfulness in the context of CE:

132 **Definition 4.1** (Faithful Counterfactuals). *Let $\mathcal{X}_\theta|\mathbf{y}^+ = p_\theta(\mathbf{x}|\mathbf{y}^+)$ denote the conditional distribution*
 133 *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*
 134 *a faithful counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^+$.*

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

137 4.1 Quantifying the Model’s Generative Property

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

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

143 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].
 144 The term $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^+)$ denotes the model energy conditioned on the target class label \mathbf{y}^+ which we
 145 specify as the negative logit corresponding to the target class label \mathbf{y}^* . To allow for faster sampling,
 146 we follow the common practice of choosing the step-size ϵ and the standard deviation of \mathbf{r}_j separately.
 147 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
 148 a small finite ϵ is negligible in practice [27, 24]. Appendix A provides additional implementation
 149 details for any tasks related to energy-based modelling.

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

154 4.2 Evaluating Plausibility and Faithfulness

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

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

162 where \mathbf{x}' denotes the counterfactual and $\mathbf{X}_{\mathbf{y}^+}$ is a subsample of the training data in the target class
 163 \mathbf{y}^+ . By averaging over multiple samples in this manner, we avoid the need to make any assumption
 164 about the nearest neighbour of \mathbf{x}' .

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

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

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:

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

The first penalty term involving λ_1 induces proximity like in Wachter et al. [1]. Our default choice for $\text{dist}(\cdot)$ is the L1 Norm due to its sparsity-inducing properties. The second penalty term involving λ_2 constrains the energy of the generated counterfactual by penalising its distance from the lowest-energy conditional samples as defined in Equation 4. Intuitively, this component induces faithfulness which coincides with plausibility to the extent that the model M_θ has learned the true posterior conditional distribution of inputs: $p_\theta(\mathbf{x}_{\mathbf{y}^+}) \rightarrow p(\mathbf{x}_{\mathbf{y}^+})$.

The third and final penalty term involving λ_3 introduces a new but familiar 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:

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

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}_{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 $\text{yloss}(\cdot)$ 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.

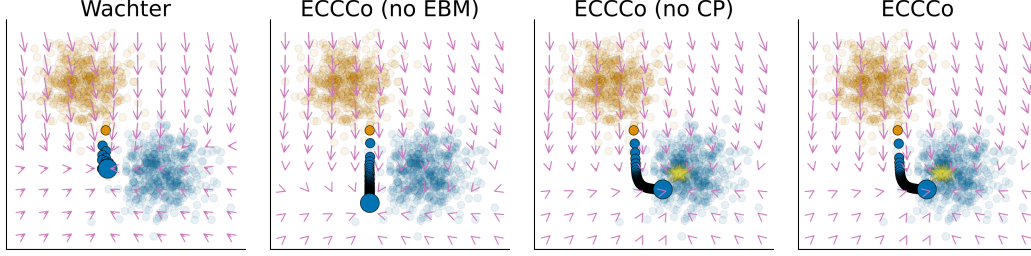


Figure 2: Gradient fields and counterfactual paths for different generators. The objective is to generate a counterfactual in the ‘blue’ class for a sample from the ‘orange’ class. Bright yellow stars indicate conditional samples generated through SGLD. The underlying classifier is a Joint Energy Model.

Algorithm 1 The *ECCCo* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T, \eta, n_B, n_E$ 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 $\{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B} \leftarrow p_\theta(\mathbf{x}_{\mathbf{y}^+})$ ▷ Generate n_B samples using SGLD (Equation 2).
- 3: Store $\hat{\mathbf{X}}^{n_E} \leftarrow \{\hat{\mathbf{x}}_{\theta, \mathbf{y}^+}\}_{n_B}$ ▷ Choose n_E lowest-energy samples.
- 4: Run *SCP* for M_θ 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: $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \hat{\mathbf{X}}^{n_E}; \Lambda, \alpha)$ ▷ Take gradient step of size η .
- 8: $t \leftarrow t + 1$
- 9: **end while**
- 10: $\mathbf{x}' \leftarrow f(\mathbf{z}')$ ▷ Map back to feature space.

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

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

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

231 6 Empirical Analysis

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

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

235 **Research Question 6.2** (Joint Objective). *Compared to our benchmark generators, how do ECCCoS*
 236 *balance the two key objectives of faithfulness and plausibility?*

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

240 6.1 Experimental Setup

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

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

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

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

266 6.2 Results for Synthetic Data

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

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

Table 1: Results for synthetic datasets: sample averages plus/minus one standard deviation. 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	Generator	Linearly Separable		Moons		Circles	
		Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓
JEM	ECCTCo	0.03 ± 0.06**	0.20 ± 0.08**	0.31 ± 0.30*	1.20 ± 0.15**	0.52 ± 0.36	1.22 ± 0.46
	ECCTCo (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
	ECCTCo (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	ECCTCo	0.29 ± 0.05**	0.23 ± 0.06**	0.80 ± 0.62	1.69 ± 0.40	0.65 ± 0.53	1.17 ± 0.41
	ECCTCo (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
	ECCTCo (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*, *ECCTCo* yields significantly 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 other more complex datasets and supportive of our narrative, so it is worth highlighting: *ECCTCos* 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 *ECCTCo* 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 of all evaluation metrics). It turns out that 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 strongly on the quality of the surrogate model. Independent of the outcome for *REVISE*, however, the results do not seem to indicate that *ECCTCo* significantly improves faithfulness and plausibility for the *Circles* data. We think this points to a limitation of our evaluation metrics rather than *ECCTCo* itself: computing average distances fails to account for the ‘wraparound’ effect associated with circular data [32].

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 *ECCTCos* is positively correlated with the capacity of the black-box model to distinguish plausible from implausible inputs. The case is very clear for *MNIST*: *ECCTCos* 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*, *ECCTCo* is essentially on par with *REVISE* and does significantly better than the baseline generator. We also note that *ECCTCo* is the only generator that consistently achieves full validity for all models (Appendix E).

For the tabular credit dataset (*GMSC*) it is inherently challenging to use deep neural networks in order to achieve good discriminative performance [33, 34] and discriminative performance [35], respectively. In order to achieve high plausibility, *ECCTCo* 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 *ECCTCo* once again achieves state-of-the-art faithfulness, it is outperformed by *REVISE* and *Schut* with respect to plausibility.

6.4 Key Takeways

To conclude this section, we summarize our findings with reference to the opening questions. The results have clearly demonstrated that *ECCTCo* consistently achieves state-of-the-art faithfulness, as it was designed to do (Research Question 6.1). A related important finding is that *ECCTCo* yields highly

Table 2: Results for real-world datasets: sample averages plus/minus one standard deviation. 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	Generator	MNIST		GMSC	
		Unfaithfulness ↓	Implausibility ↓	Unfaithfulness ↓	Implausibility ↓
JEM	ECCCo	19.28 ± 5.01**	314.76 ± 32.36*	79.45 ± 11.98**	22.05 ± 10.58**
	REVISE	188.70 ± 26.18*	255.26 ± 41.50**	187.06 ± 31.29	7.06 ± 7.73**
	Schut	211.00 ± 27.21	286.61 ± 39.85*	185.64 ± 37.42	8.47 ± 8.68**
	Wachter	222.90 ± 26.56	361.88 ± 39.74	186.20 ± 42.26	70.79 ± 58.72
JEM Ensemble	ECCCo	15.99 ± 3.06**	294.72 ± 30.75**	79.65 ± 11.83**	17.81 ± 5.44**
	REVISE	173.59 ± 20.65**	246.32 ± 37.46**	204.14 ± 36.13	4.90 ± 0.95**
	Schut	205.33 ± 24.07	287.39 ± 39.33*	186.24 ± 36.18	6.35 ± 1.22**
	Wachter	217.67 ± 23.78	363.23 ± 39.24	184.05 ± 23.11	61.40 ± 48.29
MLP	ECCCo	41.95 ± 6.50**	591.58 ± 36.24	79.84 ± 15.97**	26.78 ± 11.64**
	REVISE	365.82 ± 15.35*	249.49 ± 41.55**	180.18 ± 30.75	5.05 ± 1.05**
	Schut	382.44 ± 17.81	285.98 ± 42.48*	196.86 ± 45.07	11.16 ± 12.19**
	Wachter	386.05 ± 16.60	361.83 ± 42.18	196.51 ± 31.36	81.50 ± 54.31
MLP Ensemble	ECCCo	31.43 ± 3.91**	490.88 ± 27.19	76.32 ± 14.56**	22.99 ± 8.31**
	REVISE	337.74 ± 11.89*	247.67 ± 38.36**	184.04 ± 29.13*	5.25 ± 1.31**
	Schut	359.54 ± 14.52	283.99 ± 41.08*	214.74 ± 34.33	6.18 ± 1.17**
	Wachter	360.79 ± 14.39	357.73 ± 42.55	216.50 ± 41.31	64.04 ± 52.79

plausible explanations provided that they faithfully describe model behaviour (Research Question 6.2). Our findings here also indicate that *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. This choice is associated with relatively higher costs to individuals since the proposed recourse typically involves more substantial feature changes than for our benchmark generators.

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

Furthermore, 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, common challenges associated with energy-based modelling during our experiments 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 representations of the data, which we demonstrate 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 CE.

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 Buszydlík, 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 Velicoglu, 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=HkxxzXONtDB>.
- [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=t80-4LKfVx>.
- [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] 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>.
- [33] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. 2021.
- [34] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? 2022.
- [35] 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>.
- [36] Patrick Altmeyer. Conformal Prediction in Julia. URL <https://www.paltmeyer.com/blog/posts/conformal-prediction/>.

Appendices

A JEM

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]. While Grathwohl et al. [24] 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. [30] 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 6 in the body of this paper

Formally, it is defined as $C_{\theta, \mathbf{y}}(\mathbf{x}_i; \alpha) := \sigma((s(\mathbf{x}_i, \mathbf{y}) - \alpha)T^{-1})$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling [30].

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 [36]. 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}_{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 [28].

Finally, classification sets are formed as follows,

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

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}_{test} does not cover the true test label \mathbf{y}_{test} approaches α [28].

Observe from Equation 7 that Conformal Prediction works on an instance-level basis, much like CE 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 ↑
Circles	JEM	ECCCo	0.74 (0.21)	0.52 (0.36)	1.22 (0.46)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)**
		ECCCo (no CP)	0.72 (0.21)	0.54 (0.39)	1.21 (0.46)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)**
		ECCCo (no EBM)	0.52 (0.15)	0.70 (0.33)	1.30 (0.37)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)**
		REVISE	0.97 (0.34)	0.48 (0.16)*	0.95 (0.32)*	0.00 (0.00)	0.00 (0.00)	0.50 (0.51)
		Schut	1.06 (0.43)	0.54 (0.43)	1.28 (0.53)	0.26 (0.25)*	0.00 (0.00)	1.00 (0.00)**
		Wachter	0.44 (0.16)	0.68 (0.34)	1.33 (0.32)	0.00 (0.00)	0.00 (0.00)	0.98 (0.14)
	MLP	ECCCo	0.67 (0.19)	0.65 (0.53)	1.17 (0.41)	0.00 (0.00)	0.09 (0.19)**	1.00 (0.00)
		ECCCo (no CP)	0.71 (0.16)	0.49 (0.35)	1.19 (0.44)	0.00 (0.00)	0.05 (0.16)**	1.00 (0.00)
		ECCCo (no EBM)	0.45 (0.11)	0.84 (0.51)	1.23 (0.31)	0.00 (0.00)	0.15 (0.23)*	1.00 (0.00)
		REVISE	0.96 (0.31)	0.58 (0.52)	0.95 (0.32)	0.00 (0.00)	0.00 (0.00)**	0.50 (0.51)
		Schut	0.57 (0.11)	0.58 (0.37)	1.23 (0.43)	0.43 (0.18)**	0.00 (0.00)**	1.00 (0.00)
		Wachter	0.40 (0.09)	0.83 (0.50)	1.24 (0.29)	0.00 (0.00)	0.53 (0.01)	1.00 (0.00)
GMSC	JEM	ECCCo	19.32 (4.51)**	79.45 (11.98)**	22.05 (10.58)**	0.00 (0.00)	0.07 (0.03)	0.85 (0.37)
		REVISE	3.66 (2.25)**	187.06 (31.29)	7.06 (7.73)**	0.00 (0.00)	0.37 (0.21)	0.95 (0.22)
		Schut	1.56 (1.75)**	185.64 (37.42)	8.47 (8.68)**	0.69 (0.19)**	0.08 (0.02)	0.95 (0.22)
		Wachter	65.38 (61.49)	186.20 (42.26)	70.79 (58.72)	0.00 (0.00)	0.08 (0.02)	0.95 (0.22)
	JEM Ensemble	ECCCo	16.90 (4.81)**	79.65 (11.83)**	17.81 (5.44)**	0.00 (0.00)	0.17 (0.19)	1.00 (0.00)
		REVISE	2.97 (0.95)**	204.14 (36.13)	4.90 (0.95)**	0.00 (0.00)	0.35 (0.18)	1.00 (0.00)
		Schut	1.23 (0.30)**	186.24 (36.18)	6.35 (1.22)**	0.66 (0.06)**	0.13 (0.06)	1.00 (0.00)
		Wachter	57.72 (49.41)	184.05 (23.11)	61.40 (48.29)	0.01 (0.02)	0.11 (0.02)	1.00 (0.00)
	MLP	ECCCo	22.47 (6.06)**	79.84 (15.97)**	26.78 (11.64)**	0.00 (0.00)	0.11 (0.05)	0.85 (0.37)
		REVISE	7.29 (12.81)**	180.18 (30.75)	5.05 (1.05)**	0.00 (0.00)	0.31 (0.14)	1.00 (0.00)**
		Schut	2.67 (2.71)**	196.86 (45.07)	11.16 (12.19)**	0.67 (0.25)**	0.12 (0.04)	0.90 (0.31)
		Wachter	81.98 (54.19)	196.51 (31.36)	81.50 (54.31)	0.00 (0.00)	0.12 (0.04)	0.90 (0.31)
	MLP Ensemble	ECCCo	22.45 (8.45)**	76.32 (14.56)**	22.99 (8.31)**	0.00 (0.00)	0.13 (0.00)	1.00 (0.00)**
		REVISE	3.16 (0.91)**	184.04 (29.13)*	5.25 (1.31)**	0.00 (0.00)	0.27 (0.11)	1.00 (0.00)**
		Schut	0.61 (0.24)**	214.74 (34.33)	6.18 (1.17)**	0.89 (0.03)**	0.13 (0.00)	1.00 (0.00)**
		Wachter	60.72 (53.52)	216.50 (41.31)	64.04 (52.79)	0.00 (0.00)	0.06 (0.06)	0.50 (0.51)
Linearly Separable	JEM	ECCCo	0.75 (0.17)	0.03 (0.06)**	0.20 (0.08)**	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no CP)	0.75 (0.17)	0.03 (0.06)**	0.20 (0.08)**	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no EBM)	0.70 (0.16)	0.16 (0.11)	0.34 (0.19)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		REVISE	0.41 (0.15)	0.19 (0.03)	0.41 (0.01)**	0.00 (0.00)	0.36 (0.36)	0.50 (0.51)
		Schut	1.15 (0.35)	0.39 (0.07)	0.73 (0.17)	0.25 (0.25)	0.00 (0.00)	1.00 (0.00)
		Wachter	0.50 (0.13)	0.18 (0.10)	0.44 (0.17)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
	MLP	ECCCo	0.95 (0.16)	0.29 (0.05)**	0.23 (0.06)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no CP)	0.94 (0.16)	0.29 (0.05)**	0.23 (0.07)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no EBM)	0.60 (0.15)	0.46 (0.05)	0.28 (0.04)**	0.00 (0.00)	0.02 (0.10)**	1.00 (0.00)
		REVISE	0.42 (0.14)	0.56 (0.05)	0.41 (0.01)	0.00 (0.00)	0.47 (0.50)	0.48 (0.50)
		Schut	0.77 (0.17)	0.43 (0.06)*	0.47 (0.36)	0.20 (0.25)	0.00 (0.00)**	1.00 (0.00)
		Wachter	0.51 (0.15)	0.51 (0.04)	0.40 (0.08)	0.00 (0.00)	0.59 (0.02)	1.00 (0.00)
MNIST	JEM	ECCCo	334.61 (46.37)	19.28 (5.01)**	314.76 (32.36)*	0.00 (0.00)	4.43 (0.56)	0.98 (0.12)
		REVISE	170.68 (63.26)	188.70 (26.18)*	255.26 (41.50)**	0.00 (0.00)	4.39 (0.91)	0.96 (0.20)
		Schut	9.44 (1.60)**	211.00 (27.21)	286.61 (39.85)*	0.99 (0.00)**	1.08 (1.95)*	0.24 (0.43)
		Wachter	128.36 (14.95)	222.90 (26.56)	361.88 (39.74)	0.00 (0.00)	4.37 (0.98)	0.95 (0.21)
	JEM Ensemble	ECCCo	342.64 (41.14)	15.99 (3.06)**	294.72 (30.75)**	0.00 (0.00)	2.07 (0.06)**	1.00 (0.00)**
		REVISE	170.21 (58.02)	173.59 (20.65)**	246.32 (37.46)**	0.00 (0.00)	2.56 (0.83)	0.93 (0.26)
		Schut	9.78 (1.02)**	205.33 (24.07)	287.39 (39.33)*	0.99 (0.00)**	0.32 (0.94)**	0.11 (0.31)
		Wachter	135.07 (16.79)	217.67 (23.78)	363.23 (39.24)	0.00 (0.00)	2.93 (0.77)	0.94 (0.23)
	MLP	ECCCo	605.17 (44.78)	41.95 (6.50)**	591.58 (36.24)	0.00 (0.00)	0.57 (0.00)**	1.00 (0.00)**
		REVISE	146.61 (36.96)	365.82 (15.35)*	249.49 (41.55)**	0.00 (0.00)	0.62 (0.30)	0.87 (0.34)
		Schut	9.95 (0.37)**	382.44 (17.81)	285.98 (42.48)*	0.99 (0.00)**	0.05 (0.19)**	0.06 (0.24)
		Wachter	136.08 (16.09)	386.05 (16.60)	361.83 (42.18)	0.00 (0.00)	0.68 (0.36)	0.84 (0.36)
	MLP Ensemble	ECCCo	525.87 (34.00)	31.43 (3.91)**	490.88 (27.19)	0.00 (0.00)	0.29 (0.00)**	1.00 (0.00)**
		REVISE	146.60 (35.64)	337.74 (11.89)*	247.67 (38.36)**	0.00 (0.00)	0.39 (0.22)	0.85 (0.36)
		Schut	9.98 (0.25)**	359.54 (14.52)	283.99 (41.08)*	0.99 (0.00)**	0.03 (0.14)**	0.06 (0.24)
		Wachter	137.53 (18.95)	360.79 (14.39)	357.73 (42.55)	0.00 (0.00)	0.47 (0.64)	0.80 (0.40)
Moons	JEM	ECCCo	1.56 (0.44)	0.31 (0.30)*	1.20 (0.15)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)**
		ECCCo (no CP)	1.56 (0.46)	0.37 (0.30)*	1.21 (0.17)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)**
		ECCCo (no EBM)	0.80 (0.25)	0.91 (0.32)	1.71 (0.25)	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)**
		REVISE	1.04 (0.43)	0.78 (0.23)	1.57 (0.26)	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)**
		Schut	1.12 (0.31)	0.67 (0.27)	1.50 (0.22)*	0.08 (0.19)	0.00 (0.00)**	0.98 (0.14)
		Wachter	0.72 (0.24)	0.80 (0.27)	1.78 (0.24)	0.00 (0.00)	0.02 (0.10)	0.98 (0.14)
	MLP	ECCCo	2.18 (1.05)	0.80 (0.62)	1.69 (0.40)	0.00 (0.00)	0.15 (0.24)*	1.00 (0.00)
		ECCCo (no CP)	2.07 (1.15)	0.79 (0.62)	1.68 (0.42)	0.00 (0.00)	0.15 (0.24)*	1.00 (0.00)
		ECCCo (no EBM)	1.25 (0.92)	1.34 (0.47)	1.68 (0.47)	0.00 (0.00)	0.43 (0.18)	1.00 (0.00)
		REVISE	0.79 (0.19)*	1.45 (0.44)	1.64 (0.31)	0.00 (0.00)	0.40 (0.22)	1.00 (0.00)
		Schut	0.73 (0.25)*	1.45 (0.55)	1.73 (0.48)	0.31 (0.28)*	0.00 (0.00)**	0.90 (0.30)
		Wachter	1.08 (0.83)	1.32 (0.41)	1.69 (0.32)	0.00 (0.00)	0.52 (0.08)	1.00 (0.00)

Table 4: All results for all datasets including only valid counterfactuals. 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 ↑
Circles	JEM	ECCCo	0.74 (0.21)	0.52 (0.36)	1.22 (0.46)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no CP)	0.72 (0.21)	0.54 (0.39)	1.21 (0.46)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no EBM)	0.52 (0.15)	0.70 (0.33)	1.30 (0.37)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		REVISE	0.97 (0.34)	0.48 (0.16)*	0.95 (0.32)*	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		Schut	1.06 (0.43)	0.54 (0.43)	1.28 (0.53)	0.26 (0.25)*	0.00 (0.00)	1.00 (0.00)
		Wachter	0.44 (0.16)	0.68 (0.34)	1.33 (0.32)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
	MLP	ECCCo	0.67 (0.19)	0.65 (0.53)	1.17 (0.41)	0.00 (0.00)	0.09 (0.19)**	1.00 (0.00)
		ECCCo (no CP)	0.71 (0.16)	0.49 (0.35)	1.19 (0.44)	0.00 (0.00)	0.05 (0.16)**	1.00 (0.00)
		ECCCo (no EBM)	0.45 (0.11)	0.84 (0.51)	1.23 (0.31)	0.00 (0.00)	0.15 (0.23)*	1.00 (0.00)
		REVISE	0.96 (0.31)	0.58 (0.52)	0.95 (0.32)	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		Schut	0.57 (0.11)	0.58 (0.37)	1.23 (0.43)	0.43 (0.18)**	0.00 (0.00)**	1.00 (0.00)
		Wachter	0.40 (0.09)	0.83 (0.50)	1.24 (0.29)	0.00 (0.00)	0.53 (0.01)	1.00 (0.00)
GMSC	JEM	ECCCo	19.32 (4.51)**	79.45 (11.98)**	22.05 (10.58)**	0.00 (0.00)	0.07 (0.03)	1.00 (0.00)
		REVISE	3.66 (2.25)**	187.06 (31.29)	7.06 (7.73)**	0.00 (0.00)	0.37 (0.21)	1.00 (0.00)
		Schut	1.56 (1.75)**	185.64 (37.42)	8.47 (8.68)**	0.69 (0.19)**	0.08 (0.02)	1.00 (0.00)
		Wachter	65.38 (61.49)	186.20 (42.26)	70.79 (58.72)	0.00 (0.00)	0.08 (0.02)	1.00 (0.00)
	JEM Ensemble	ECCCo	16.90 (4.81)**	79.65 (11.83)**	17.81 (5.44)**	0.00 (0.00)	0.17 (0.19)	1.00 (0.00)
		REVISE	2.97 (0.95)**	204.14 (36.13)	4.90 (0.95)**	0.00 (0.00)	0.35 (0.18)	1.00 (0.00)
		Schut	1.23 (0.30)**	186.24 (36.18)	6.35 (1.22)**	0.66 (0.06)**	0.13 (0.06)	1.00 (0.00)
		Wachter	57.72 (49.41)	184.05 (23.11)	61.40 (48.29)	0.01 (0.02)	0.11 (0.02)	1.00 (0.00)
	MLP	ECCCo	22.47 (6.06)**	79.84 (15.97)**	26.78 (11.64)**	0.00 (0.00)	0.11 (0.05)	1.00 (0.00)
		REVISE	7.29 (12.81)**	180.18 (30.75)	5.05 (1.05)**	0.00 (0.00)	0.31 (0.14)	1.00 (0.00)
		Schut	2.67 (2.71)**	196.86 (45.07)	11.16 (12.19)**	0.67 (0.25)**	0.12 (0.04)	1.00 (0.00)
		Wachter	81.98 (54.19)	196.51 (31.36)	81.50 (54.31)	0.00 (0.00)	0.12 (0.04)	1.00 (0.00)
	MLP Ensemble	ECCCo	22.45 (8.45)**	76.32 (14.56)**	22.99 (8.31)**	0.00 (0.00)	0.13 (0.00)	1.00 (0.00)
		REVISE	3.16 (0.91)**	184.04 (29.13)*	5.25 (1.31)**	0.00 (0.00)	0.27 (0.11)	1.00 (0.00)
		Schut	0.61 (0.24)**	214.74 (34.33)	6.18 (1.17)**	0.89 (0.03)**	0.13 (0.00)	1.00 (0.00)
		Wachter	60.72 (53.52)	216.50 (41.31)	64.04 (52.79)	0.00 (0.00)	0.06 (0.06)	1.00 (0.00)
Linearly Separable	JEM	ECCCo	0.75 (0.17)	0.03 (0.06)**	0.20 (0.08)**	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no CP)	0.75 (0.17)	0.03 (0.06)**	0.20 (0.08)**	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		ECCCo (no EBM)	0.70 (0.16)	0.16 (0.11)	0.34 (0.19)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
		REVISE	0.41 (0.15)	0.19 (0.03)	0.41 (0.01)**	0.00 (0.00)	0.36 (0.36)	1.00 (0.00)
		Schut	1.15 (0.35)	0.39 (0.07)	0.73 (0.17)	0.25 (0.25)	0.00 (0.00)	1.00 (0.00)
		Wachter	0.50 (0.13)	0.18 (0.10)	0.44 (0.17)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
	MLP	ECCCo	0.95 (0.16)	0.29 (0.05)**	0.23 (0.06)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no CP)	0.94 (0.16)	0.29 (0.05)**	0.23 (0.07)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no EBM)	0.60 (0.15)	0.46 (0.05)	0.28 (0.04)**	0.00 (0.00)	0.02 (0.10)**	1.00 (0.00)
		REVISE	0.42 (0.14)	0.56 (0.05)	0.41 (0.01)	0.00 (0.00)	0.47 (0.50)	1.00 (0.00)
		Schut	0.77 (0.17)	0.43 (0.06)*	0.47 (0.36)	0.20 (0.25)	0.00 (0.00)**	1.00 (0.00)
		Wachter	0.51 (0.15)	0.51 (0.04)	0.40 (0.08)	0.00 (0.00)	0.59 (0.02)	1.00 (0.00)
	JEM Ensemble	ECCCo	334.61 (46.37)	19.28 (5.01)**	314.76 (32.36)*	0.00 (0.00)	4.43 (0.56)	1.00 (0.00)
		REVISE	170.68 (63.26)	188.70 (26.18)*	255.26 (41.50)**	0.00 (0.00)	4.39 (0.91)	1.00 (0.00)
		Schut	9.44 (1.60)**	211.00 (27.21)	286.61 (39.85)*	0.99 (0.00)**	1.08 (1.95)*	1.00 (0.00)
		Wachter	128.36 (14.95)	222.90 (26.56)	361.88 (39.74)	0.00 (0.00)	4.37 (0.98)	1.00 (0.00)
	MLP Ensemble	ECCCo	342.64 (41.14)	15.99 (3.06)**	294.72 (30.75)**	0.00 (0.00)	2.07 (0.06)**	1.00 (0.00)
		REVISE	170.21 (58.02)	173.59 (20.65)**	246.32 (37.46)**	0.00 (0.00)	2.56 (0.83)	1.00 (0.00)
		Schut	9.78 (1.02)**	205.33 (24.07)	287.39 (39.33)*	0.99 (0.00)**	0.32 (0.94)**	1.00 (0.00)
		Wachter	135.07 (16.79)	217.67 (23.78)	363.23 (39.24)	0.00 (0.00)	2.93 (0.77)	1.00 (0.00)
MNIST	JEM	ECCCo	605.17 (44.78)	41.95 (6.50)**	591.58 (36.24)	0.00 (0.00)	0.57 (0.00)**	1.00 (0.00)
		REVISE	146.61 (36.96)	365.82 (15.35)*	249.49 (41.55)**	0.00 (0.00)	0.62 (0.30)	1.00 (0.00)
		Schut	9.95 (0.37)**	382.44 (17.81)	285.98 (42.48)*	0.99 (0.00)**	0.05 (0.19)**	1.00 (0.00)
		Wachter	136.08 (16.09)	386.05 (16.60)	361.83 (42.18)	0.00 (0.00)	0.68 (0.36)	1.00 (0.00)
	JEM Ensemble	ECCCo	525.87 (34.00)	31.43 (3.91)**	490.88 (27.19)	0.00 (0.00)	0.29 (0.00)**	1.00 (0.00)
		REVISE	146.60 (35.64)	337.74 (11.89)*	247.67 (38.36)**	0.00 (0.00)	0.39 (0.22)	1.00 (0.00)
		Schut	9.98 (0.25)**	359.54 (14.52)	283.99 (41.08)*	0.99 (0.00)**	0.03 (0.14)**	1.00 (0.00)
		Wachter	137.53 (18.95)	360.79 (14.39)	357.73 (42.55)	0.00 (0.00)	0.47 (0.64)	1.00 (0.00)
	MLP	ECCCo	1.56 (0.44)	0.31 (0.30)*	1.20 (0.15)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no CP)	1.56 (0.46)	0.37 (0.30)*	1.21 (0.17)**	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		ECCCo (no EBM)	0.80 (0.25)	0.91 (0.32)	1.71 (0.25)	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
		REVISE	1.04 (0.43)	0.78 (0.23)	1.57 (0.26)	0.00 (0.00)	0.00 (0.00)**	1.00 (0.00)
	MLP Ensemble	ECCCo	1.12 (0.31)	0.67 (0.27)	1.50 (0.22)*	0.08 (0.19)	0.00 (0.00)**	1.00 (0.00)
		Schut	0.72 (0.24)	0.80 (0.27)	1.78 (0.24)	0.00 (0.00)	0.02 (0.10)	1.00 (0.00)
Moons	JEM	ECCCo	2.18 (1.05)	0.80 (0.62)	1.69 (0.40)	0.00 (0.00)	0.15 (0.24)*	1.00 (0.00)
		ECCCo (no CP)	2.07 (1.15)	0.79 (0.62)	1.68 (0.42)	0.00 (0.00)	0.15 (0.24)*	1.00 (0.00)
		ECCCo (no EBM)	1.25 (0.92)	1.34 (0.47)	1.68 (0.47)	0.00 (0.00)	0.43 (0.18)	1.00 (0.00)
		REVISE	0.79 (0.19)*	1.45 (0.44)	1.64 (0.31)	0.00 (0.00)	0.40 (0.22)	1.00 (0.00)
		Schut	0.73 (0.25)*	1.45 (0.55)	1.73 (0.48)	0.31 (0.28)*	0.00 (0.00)**	1.00 (0.00)
		Wachter	1.08 (0.83)	1.32 (0.41)	1.69 (0.32)	0.00 (0.00)	0.52 (0.08)	1.00 (0.00)