

---

# ECCCoS from the Black Box: Letting Models speak for Themselves

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Counterfactual Explanations offer an intuitive and straightforward way to explain  
2 Black Box Models but they are not unique. To address the need for plausible  
3 explanations, existing work has primarily relied on surrogate models to learn how  
4 the input data is distributed. This effectively reallocates the task of learning realistic  
5 representations of the data from the model itself to the surrogate. Consequently, the  
6 generated explanations may look plausible to humans but not necessarily conform  
7 with the behaviour of the Black Box Model. We formalise this notion of model  
8 conformity through the introduction of tailored evaluation measures and propose  
9 a novel algorithmic framework for generating **Energy-Constrained Conformal**  
10 **Counterfactuals** that are only as plausible as the model permits. To do so, **ECCCo**  
11 leverages recent advances in energy-based modelling and predictive uncertainty  
12 quantification through conformal inference. Through illustrative examples and  
13 extensive empirical studies, we demonstrate that ECCCoS reconcile the need for  
14 plausibility and model conformity.

## 15 1 Introduction

16 Counterfactual Explanations provide a powerful, flexible and intuitive way to not only explain Black  
17 Box Models but also enable affected individuals to challenge them through the means of Algorithmic  
18 Recourse. Instead of opening the black box, Counterfactual Explanations work under the premise  
19 of strategically perturbing model inputs to understand model behaviour [29]. Intuitively speaking,  
20 we generate explanations in this context by asking simple what-if questions of the following nature:  
21 ‘Our credit risk model currently predicts that this individual’s credit profile is too risky to offer them a  
22 loan. What if they reduced their monthly expenditures by 10%? Will our model then predict that the  
23 individual is credit-worthy?’

24 This is typically implemented by defining a target outcome  $\mathbf{y}^* \in \mathcal{Y}$  for some individual  $\mathbf{x} \in \mathcal{X} = \mathbb{R}^D$   
25 described by  $D$  attributes, for which the model  $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$  initially predicts a different outcome:  
26  $M_\theta(\mathbf{x}) \neq \mathbf{y}^*$ . Counterfactuals are then searched by minimizing a loss function that compares the  
27 predicted model output to the target outcome:  $\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}^*)$ . Since Counterfactual Explanations  
28 (CE) work directly with the Black Box Model, valid counterfactuals always have full local fidelity by  
29 construction [17]. Fidelity is defined as the degree to which explanations approximate the predictions  
30 of the Black Box Model. This is arguably one of the most important evaluation metrics for model  
31 explanations, since any explanation that explains a prediction not actually made by the model is  
32 useless [16].

33 In situations where full fidelity is a requirement, CE therefore offers a more appropriate solution to  
34 Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [22] and SHAP  
35 [12], which involve local surrogate models. But even full fidelity is not a sufficient condition for  
36 ensuring that an explanation adequately describes the behaviour of a model. That is because two

37 very distinct explanations can both lead to the same model prediction, especially when dealing with  
 38 heavily parameterized models:

39           [. . .] deep neural networks are typically very underspecified by the available  
 40           data, and [. . .] parameters [therefore] correspond to a diverse variety of compelling  
 41           explanations for the data. — Wilson [30]

42 When people talk about Black Box Models, this is usually the type of model they have in mind.

43 In the context of CE, the idea that no two explanations are the same arises almost naturally. Even  
 44 the baseline approach proposed by Wachter et al. [29] can yield a diverse set of explanations  
 45 if counterfactuals are initialised randomly. This multiplicity of explanations has not only been  
 46 acknowledged in the literature but positively embraced: since individuals seeking Algorithmic  
 47 Recourse (AR) have unique preferences, Mothilal et al. [17], for example, have prescribed *diversity*  
 48 as an explicit goal for counterfactuals. More generally, the literature on CE and AR has brought  
 49 forward a myriad of desiderata for explanations, which we will discuss in more detail in the following  
 50 section.

## 51 2 Background and Related Work

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

### 55 2.1 Gradient-Based Counterfactual Search

56 While Counterfactual Explanations can be generated for arbitrary regression models [24], existing  
 57 work has primarily focused on classification problems. Let  $\mathcal{Y} = (0, 1)^K$  denote the one-hot-encoded  
 58 output domain with  $K$  classes. Then most SOTA counterfactual generators rely on gradient descent  
 59 to optimize different flavours of the following counterfactual search objective:

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

60 Here  $\text{yloss}$  denotes the primary loss function already introduced above and  $\text{cost}$  is either a single  
 61 penalty or a collection of penalties that are used to impose constraints through regularization. Follow-  
 62 ing the convention in Altmeyer et al. [2] we use  $\mathbf{Z}' = \{\mathbf{z}_m\}_M$  to denote the vector  $M$ -dimensional  
 63 array of counterfactual states. This is to explicitly account for the fact that we can generate multiple  
 64 counterfactuals  $M$ , as with DiCE [17], and may choose to traverse a latent representation  $\mathcal{Z}$  of the  
 65 feature space  $\mathcal{X}$ , as we will discuss further below.

66 Solutions to Equation 1 are considered valid as soon as the predicted label matches the target label. A  
 67 stripped-down counterfactual explanation is therefore little different from an adversarial example.  
 68 In Figure 1, for example, we have the baseline approach proposed in Wachter et al. [29] to MNIST  
 69 data (centre panel). This approach solves Equation 1 through gradient-descent in the feature space  
 70 with a penalty for the distance between the factual  $\mathbf{x}$  and the counterfactual  $\mathbf{x}'$ . The underlying  
 71 classifier  $M_\theta$  is a simple Multi-Layer Perceptron (MLP) with good test accuracy. For the generated  
 72 counterfactual  $\mathbf{x}'$  the model predicts the target label with high confidence (centre panel in Figure 1).  
 73 The explanation is valid by definition, even though it looks a lot like an Adversarial Example [6].  
 74 Schut et al. [23] make the connection between Adversarial Examples and Counterfactual Explanations  
 75 explicit and propose using a Jacobian-Based Saliency Map Attack (JSMA) to solve Equation 1. They  
 76 demonstrate that this approach yields realistic and sparse counterfactuals for Bayesian, adversarially  
 77 robust classifiers. Applying their approach to our simple MNIST classifier does not yield a realistic  
 78 counterfactual but this one, too, is valid (right panel in Figure 1).

### 79 2.2 From Adversial Examples to Plausible Explanations

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

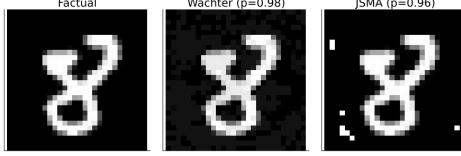


Figure 1: You may not like it, but this is what stripped-down counterfactuals look like. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using Wachter et al. [29] (centre); and a counterfactual produced using JSMA-based approach introduced by [23].

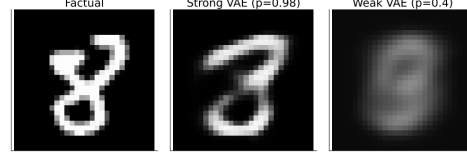


Figure 2: Using surrogates can improve plausibility, but also increases vulnerability. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using REVISE [9] with a well-specified surrogate (centre); and a counterfactual produced using REVISE [9] with a poorly specified surrogate (right).

both AI practitioners and individuals affected by AI decision-making systems, counterfactuals should be sparse, proximate [29], actionable [27], diverse [17], plausible [9, 21, 23], robust [26, 20, 2] and causal [11] among other things.

Researchers have come up with various ways to meet these desiderata, which have been extensively surveyed and evaluated in various studies [28, 10, 19, 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 [13].

### 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  $\mathcal{Z}$  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 compelling counterfactual 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. [21] use density estimators ( $\hat{p} : \mathcal{X} \mapsto [0, 1]$ ) to constrain the counterfactual paths. Karimi et al. [11] 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:  $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^*$ .*

Note that Definition 2.1 is consistent with the notion of plausible counterfactual paths, since we can simply apply it to each counterfactual state along the path.

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. [23] 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  $M_\theta$  predicts the target label  $\mathbf{y}^*$  with high confidence. Provided the model is well-specified, these counterfactuals are plausible. Unfortunately, this idea hinges on the assumption that the Black Box Model provides well-calibrated predictive uncertainty estimates.

## 2.3 From Fidelity to Model Conformity

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 underlying data’? To introduce this section, we posit a related, slightly more nuanced question: ‘Is this counterfactual consistent with what the model has learned about the underlying data’? We will argue that fidelity is not a sufficient evaluation measure to answer this question and propose a novel way to assess if Counterfactual Explanations conform with model behaviour.

The word *fidelity* stems from the Latin word ‘fidelis’, which means ‘faithful, loyal, trustworthy’ [15]. 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 Counterfactual Explanations, we propose an adapted version:

**Definition 2.2** (Conformal Counterfactuals). *Let  $\mathcal{X}_\theta|\mathbf{y}^* = p_\theta(x|\mathbf{y}^*)$  denote the conditional distribution of  $\mathbf{x}$  in the target class  $\mathbf{y}^*$ , where  $\theta$  denotes the parameters of model  $M_\theta$ . Then for  $\mathbf{x}'$  to be considered a conformal counterfactual, we need:  $\mathbf{x}' \sim \mathcal{X}_\theta|\mathbf{y}^*$ .*

In words, conformal counterfactuals conform with what the predictive model has learned about the input data  $\mathbf{x}$ . Since this definition works with distributional properties, it explicitly accounts for the multiplicity of explanations we discussed earlier. 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 reconciles the notions of plausibility and model conformity and 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 pointed out that there is a ‘generative model hidden within every standard discriminative model’. The authors show that we can draw samples from the posterior conditional distribution  $p_\theta(\mathbf{x}|\mathbf{y})$  using Stochastic Gradient Langevin Dynamics (SGLD). The authors use this insight to train classifiers jointly for the discriminative task using standard cross-entropy and the generative task using SGLD. They demonstrate empirically that among other things this improves predictive uncertainty quantification for discriminative models. Our findings in this work suggest that Joint Energy Models (JEM) also tend to yield more plausible

Counterfactual Explanations. Based on the definition of plausible counterfactuals (Definition 2.1) this is not surprising.

Crucially for our purpose, one can apply their proposed sampling strategy 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 the Black Box model itself. 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,

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

where  $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  is the stochastic term and the step-size  $\epsilon$  is typically polynomially decayed. The term  $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*)$  denotes the energy function where we use  $\mathcal{E}(\mathbf{x}_j|\mathbf{y}^*) = -M_\theta(\mathbf{x}_j)[\mathbf{y}^*]$ , that is the negative logit corresponding to the target class label  $\mathbf{y}^*$ . Generating multiple samples in this manner yields an empirical distribution  $\hat{\mathcal{X}}_\theta|\mathbf{y}^*$  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, 14]. 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. [25] in the context of conformal training:

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

Here,  $\kappa \in \{0, 1\}$  is a hyper-parameter and  $C_{\theta, \mathbf{y}}(\mathbf{x}; \alpha)$  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. 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,

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

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 Manhattan Distance since it enforces sparsity.

207 The first two terms in Equation 4 correspond to the counterfactual search objective defined in Wachter  
 208 et al. [29] which merely penalises the distance of counterfactuals from their factual values. The  
 209 additional two penalties in ECCCo ensure that counterfactuals conform with the model’s generative  
 210 property and lead to minimally uncertain predictions, respectively. The hyperparameters  $\lambda_1, \dots, \lambda_3$   
 211 can be used to balance the different objectives: for example, we may choose to incur larger deviations  
 212 from the factual in favour of conformity with the model’s generative property by choosing lower  
 213 values of  $\lambda_1$  and relatively higher values of  $\lambda_2$ .

---

**Algorithm 1** An algorithm with caption

---

**Require:**  $n \geq 0$   
**Ensure:**  $y = x^n$   
 $y \leftarrow 1$   
 $X \leftarrow x$   
 $N \leftarrow n$   
**while**  $N \neq 0$  **do**  
  **if**  $N$  is even **then**  
     $X \leftarrow X \times X$   
     $N \leftarrow \frac{N}{2}$    ▷ This is a comment  
  **else if**  $N$  is odd **then**  
     $y \leftarrow y \times X$   
     $N \leftarrow N - 1$   
  **end if**  
**end while**

---

## 214 4 Evaluation Framework

### 215 4.1 Evaluation Measures

216 Above we have defined plausibility (2.1) and conformity (2.2) for Counterfactual Explanations.  
 217 In this subsection, we introduce evaluation measures that facilitate a quantitative evaluation of  
 218 counterfactuals for these objectives.

219 Firstly, in order to assess the plausibility of counterfactuals we adapt the implausibility metric  
 220 proposed in Guidotti [8]. The authors propose to evaluate plausibility in terms of the distance of the  
 221 counterfactual  $\mathbf{x}'$  from its nearest neighbour in the target class  $\mathbf{y}^*$ : the smaller this distance, the more  
 222 plausible the counterfactual. Instead of focusing only on the nearest neighbour of  $\mathbf{x}'$ , we suggest  
 223 computing the average over distances from multiple (possibly all) observed instances in the target  
 224 class. Formally, for a single counterfactual, we have:

$$\text{impl} = \frac{1}{|\mathbf{x} \in \mathcal{X}|\mathbf{y}^*|} \sum_{\mathbf{x} \in \mathcal{X}|\mathbf{y}^*} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (5)$$

225 This measure is straightforward to compute and should be less sensitive to outliers in the target class  
 226 than the one based on the nearest neighbour. It also gives rise to a very similar evaluation measure for  
 227 conformity. We merely swap out the subsample of individuals in the target class for the empirical  
 228 distribution of generated conditional samples:

$$\text{conf} = \frac{1}{|\mathbf{x} \in \mathcal{X}_\theta|\mathbf{y}^*|} \sum_{\mathbf{x} \in \mathcal{X}_\theta|\mathbf{y}^*} \text{dist}(\mathbf{x}', \mathbf{x}) \quad (6)$$

229 As noted by Guidotti [8], these distance-based measures are simplistic and more complex alternative  
 230 measures may ultimately be more appropriate for the task. For example, we considered using statisti-  
 231 cal divergence measures instead. This would involve generating not one but many counterfactuals and  
 232 comparing the generated empirical distribution to the target distributions in Definitions 2.1 and 2.2.  
 233 While this approach is potentially more rigorous, generating enough counterfactuals is not always  
 234 practical.

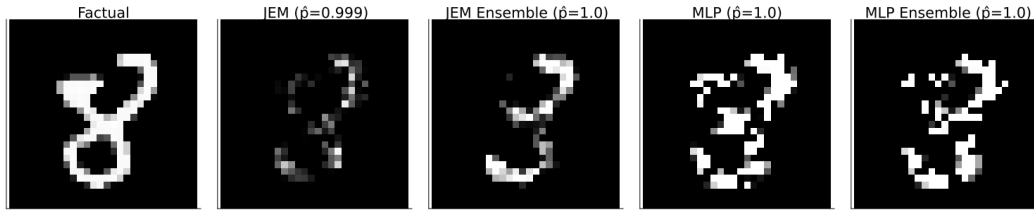


Figure 3: ECCCos from Black Boxes. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left);

## 5 Experiments

- BatchNorm does not seem compatible with JEM
- Coverage and temperature impacts CCE in somewhat unpredictable ways
- It seems that models that are not explicitly trained for generative task, still learn it implicitly
- Batch size seems to impact quality of generated samples (at inference, but not so much during JEM training)
- ECCCo is sensitive to optimizer (Adam works well), learning rate and distance metric (l1 currently only one that works)
- SGLD takes time
- REVISE has benefit of lower dimensional space

## 6 Discussion

Consistent with the findings in Schut et al. [23], we have demonstrated that predictive uncertainty estimates can be leveraged to generate plausible counterfactuals. Interestingly, Schut et al. [23] point out that this finding — as intuitive as it is — may be linked to a positive connection between the generative task and predictive uncertainty quantification. In particular, Grathwohl et al. [7] demonstrate that their proposed method for integrating the generative objective in training yields models that have improved predictive uncertainty quantification. Since neither Schut et al. [23] nor we have employed any surrogate generative models, our findings seem to indicate that the positive connection found in Grathwohl et al. [7] is bidirectional.

## References

- [1] Patrick Altmeyer. Conformal Prediction in Julia. URL <https://www.paltmeyer.com/blog/posts/conformal-prediction/>.
- [2] 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.
- [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=HkxzxONtDB>.
- [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] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] Valery Manokhin. Awesome conformal prediction.
- [15] Merriam-Webster. "fidelity". URL <https://www.merriam-webster.com/dictionary/fidelity>.
- [16] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.
- [17] 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.
- [18] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Thomas Spooner, Danial Dervovic, Jason Long, Jon Shepard, Jiahao Chen, and Daniele Magazzeni. Counterfactual Explanations for Arbitrary Regression Models. 2021.



- [25] 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>.
- [26] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. 2021.
- [27] 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.
- [28] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. 2020.
- [29] 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.
- [30] 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  $\epsilon \rightarrow 0$  and  $J \rightarrow \infty$ , the bias introduced for a small finite  $\epsilon$  is negligible in practice [18, 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. [25] 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  $\alpha$ , 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

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  $\sigma$  is the sigmoid function and  $T$  is a hyper-parameter used for temperature scaling [25].

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,

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

where  $\hat{q}$  denotes the  $(1 - \alpha)$ -quantile of  $\mathcal{S}$  and  $\alpha$  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  $\alpha$  [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  $\alpha$  consider the case of  $\alpha = 0$ , which requires that the true label is covered by the prediction set with probability equal to one.

## A Submission of papers to NeurIPS 2023

Please read the instructions below carefully and follow them faithfully.

### A Style

Papers to be submitted to NeurIPS 2023 must be prepared according to the instructions presented here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2023 are the same as those in previous years.

Authors are required to use the NeurIPS L<sup>A</sup>T<sub>E</sub>X style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

### B Retrieval of style files

The style files for NeurIPS and other conference information are available on the website at

<http://www.neurips.cc/>

The file `neurips_2023.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2023 is `neurips_2023.sty`, rewritten for L<sup>A</sup>T<sub>E</sub>X 2<sub>ε</sub>. **Previous style files for L<sup>A</sup>T<sub>E</sub>X 2.09, Microsoft Word, and RTF are no longer supported!**

The L<sup>A</sup>T<sub>E</sub>X style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

**Preprint option** If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text “Preprint. Work in progress.” in the footer. This version may be distributed as you see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please *do not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2023.tex` may be used as a “shell” for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections B, C, and D below.

## 403 **B General formatting instructions**

404 The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long.  
405 The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points.  
406 Times New Roman is the preferred typeface throughout, and will be selected for you by default.  
407 Paragraphs are separated by 1/2 line space (5.5 points), with no indentation.

408 The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal  
409 rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow 1/4 inch  
410 space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the  
411 page.

412 For the final version, authors' names are set in boldface, and each name is centered above the  
413 corresponding address. The lead author's name is to be listed first (left-most), and the co-authors'  
414 names (if different address) are set to follow. If there is only one co-author, list both author and  
415 co-author side by side.

416 Please pay special attention to the instructions in Section D regarding figures, tables, acknowledg-  
417 ments, and references.

## 418 **C Headings: first level**

419 All headings should be lower case (except for first word and proper nouns), flush left, and bold.

420 First-level headings should be in 12-point type.

### 421 **A Headings: second level**

422 Second-level headings should be in 10-point type.

#### 423 **A.1 Headings: third level**

424 Third-level headings should be in 10-point type.

425 **Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush  
426 left, and inline with the text, with the heading followed by 1 em of space.

## 427 **D Citations, figures, tables, references**

428 These instructions apply to everyone.

### 429 **A Citations within the text**

430 The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as  
431 long as you maintain internal consistency. As to the format of the references themselves, any style is  
432 acceptable as long as it is used consistently.

433 The documentation for `natbib` may be found at

434 `http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

435 Of note is the command `\citet`, which produces citations appropriate for use in inline text. For  
436 example,

437 `\citet{hasselmo}` investigated\dots

438 produces

439 Hasselmo, et al. (1995) investigated...

440 If you wish to load the `natbib` package with options, you may add the following before loading the  
441 `neurips_2023` package:

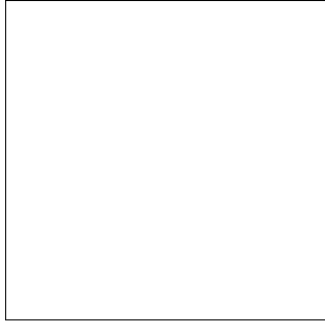


Figure 4: Sample figure caption.

442 `\PassOptionsToPackage{options}{natbib}`

443 If `natbib` clashes with another package you load, you can add the optional argument `nonatbib`  
444 when loading the style file:

445 `\usepackage[nonatbib]{neurips_2023}`

446 As submission is double blind, refer to your own published work in the third person. That is, use “In  
447 the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers  
448 that are not widely available (e.g., a journal paper under review), use anonymous author names in the  
449 citation, e.g., an author of the form “A. Anonymous” and include a copy of the anonymized paper in  
450 the supplementary material.

## 451 **B Footnotes**

452 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number<sup>1</sup>  
453 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote  
454 with a horizontal rule of 2 inches (12 picas).

455 Note that footnotes are properly typeset *after* punctuation marks.<sup>2</sup>

## 456 **C Figures**

457 All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction.  
458 The figure number and caption always appear after the figure. Place one line space before the figure  
459 caption and one line space after the figure. The figure caption should be lower case (except for first  
460 word and proper nouns); figures are numbered consecutively.

461 You may use color figures. However, it is best for the figure captions and the paper body to be legible  
462 if the paper is printed in either black/white or in color.

## 463 **D Tables**

464 All tables must be centered, neat, clean and legible. The table number and title always appear before  
465 the table. See Table 1.

466 Place one line space before the table title, one line space after the table title, and one line space after  
467 the table. The table title must be lower case (except for first word and proper nouns); tables are  
468 numbered consecutively.

469 Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the  
470 `booktabs` package, which allows for typesetting high-quality, professional tables:

471 <https://www.ctan.org/pkg/booktabs>

472 This package was used to typeset Table 1.

---

<sup>1</sup>Sample of the first footnote.

<sup>2</sup>As in this example.

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

## 473 E Math

474 Note that display math in bare TeX commands will not create correct line numbers for sub-  
 475 mission. Please use LaTeX (or AMSTeX) commands for unnumbered display math. (You  
 476 really shouldn't be using \$\$ anyway; see <https://tex.stackexchange.com/questions/503/why-is-preferable-to> and [https://tex.stackexchange.com/questions/40492/](https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath)  
 477 [what-are-the-differences-between-align-equation-and-displaymath](https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath) for more infor-  
 478 mation.)

## 480 F Final instructions

481 Do not change any aspects of the formatting parameters in the style files. In particular, do not modify  
 482 the width or length of the rectangle the text should fit into, and do not change font sizes (except  
 483 perhaps in the **References** section; see below). Please note that pages should be numbered.

## 484 E Preparing PDF files

485 Please prepare submission files with paper size "US Letter," and not, for example, "A4."

486 Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or  
 487 Embedded TrueType fonts. Here are a few instructions to achieve this.

- 488 • You should directly generate PDF files using `pdflatex`.
- 489 • You can check which fonts a PDF files uses. In Acrobat Reader, select the menu  
 490 Files>Document Properties>Fonts and select Show All Fonts. You can also use the program  
 491 `pdf fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- 492 • `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- 493 • The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS  
 494 Fonts:

495 `\usepackage{amsfonts}`  
 496 followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for  $\mathbb{R}$ ,  $\mathbb{N}$  or  $\mathbb{C}$ . You can also  
 497 use the following workaround for reals, natural and complex:

```
498 \newcommand{\RR}{\mathbb{R}} %real numbers
499 \newcommand{\Nat}{\mathbb{N}} %natural numbers
500 \newcommand{\CC}{\mathbb{C}} %complex numbers
```

501 Note that `amsfonts` is automatically loaded by the `amssymb` package.

502 If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

## 503 A Margins in L<sup>A</sup>T<sub>E</sub>X

504 Most of the margin problems come from figures positioned by hand using `\special` or other  
 505 commands. We suggest using the command `\includegraphics` from the `graphicx` package.  
 506 Always specify the figure width as a multiple of the line width as in the example below:

```
507 \usepackage[pdftex]{graphicx} ...
508 \includegraphics[width=0.8\linewidth]{myfile.pdf}
```

509 See Section 4.4 in the graphics bundle documentation ([http://mirrors.ctan.org/macros/](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf)  
510 [latex/required/graphics/grfguide.pdf](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf))  
511 A number of width problems arise when L<sup>A</sup>T<sub>E</sub>X cannot properly hyphenate a line. Please give LaTeX  
512 hyphenation hints using the \- command when necessary.

## 513 **F Supplementary Material**

514 Authors may wish to optionally include extra information (complete proofs, additional experiments  
515 and plots) in the appendix. All such materials should be part of the supplemental material (submitted  
516 separately) and should NOT be included in the main submission.

## 517 **References**

518 References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level  
519 heading for the references. Any choice of citation style is acceptable as long as you are consistent. It  
520 is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the  
521 Reference section does not count towards the page limit.

522 [1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In  
523 G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp.  
524 609–616. Cambridge, MA: MIT Press.

525 [2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the*  
526 *GENeral NEural Simulation System*. New York: TELOS/Springer-Verlag.

527 [3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent  
528 synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.