Appendices

The following appendices provide additional details that are relevant to the paper. Appendices A and B explain any tasks related to Energy-Based Modelling and Predictive Uncertainty Quantification through Conformal Prediction, respectively. Appendix C provides additional technical and implementation details about our proposed generator, *ECCCo*, including references to our open-sourced code base. A complete overview of our experimental setup detailing our parameter choices, training procedures and initial black-box model performance can be found in Appendix D. Finally, Appendix F reports all of our experimental results in more detail.

A Energy-Based Modelling

Since we were not able to identify any existing open-source software for Energy-Based Modelling that would be flexible enough to cater to our needs, we have developed a `Julia` package from scratch. The package has been open-sourced, but to avoid compromising the double-blind review process, we refrain from providing more information at this stage. In our development we have heavily drawn on the existing literature: Du and Mordatch (2019) describe best practices for using EBM for generative modelling; Grathwohl et al. (2020) explain how EBM can be used to train classifiers jointly for the discriminative and generative tasks. We have used the same package for training and inference, but there are some important differences between the two cases that are worth highlighting here.

Training: Joint Energy Models To train our Joint Energy Models we broadly follow the approach outlined in Grathwohl et al. (2020). Formally, JEMs are defined by the following joint distribution:

$$\log p_\theta(\mathbf{x}, \mathbf{y}) = \log p_\theta(\mathbf{y}|\mathbf{x}) + \log p_\theta(\mathbf{x}) \tag{10}$$

Training therefore involves a standard classification loss component $L_{\text{clf}}(\theta) = -\log p_\theta(\mathbf{y}|\mathbf{x})$ (e.g. cross-entropy loss) as well as a generative loss component $L_{\text{gen}}(\theta) = -\log p_\theta(\mathbf{x})$. Analogous to how we defined the conditional distribution over inputs in Definition 4.1, $p_\theta(\mathbf{x})$ denotes the unconditional distribution over inputs. The model gradient of this component of the loss function can be expressed as follows:

$$\nabla_\theta L_{\text{gen}}(\theta) = -\nabla_\theta \log p_\theta(\mathbf{x}) = -\left(\mathbb{E}_{p(\mathbf{x})}\left\{\nabla_\theta \mathcal{E}_\theta(\mathbf{x})\right\} - \mathbb{E}_{p_\theta(\mathbf{x})}\left\{\nabla_\theta \mathcal{E}_\theta(\mathbf{x})\right\}\right) \tag{11}$$

To draw samples from $p_\theta(\mathbf{x})$, we rely exclusively on the conditional sampling approach described in Grathwohl et al. (2020) for both training and inference: we first draw $\mathbf{y} \sim p(\mathbf{y})$ and then sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{y})$ (Grathwohl et al. 2020) via Equation 2 with energy $\mathcal{E}_\theta(\mathbf{x}|\mathbf{y}) = \mu_\theta(\mathbf{x})[\mathbf{y}]$ where $\mu_\theta : \mathcal{X} \mapsto \mathbb{R}^K$ returns the linear predictions (logits) of our classifier M_θ. While our package also supports unconditional sampling, we found conditional sampling to work well. It is also well aligned with CE, since in this context we are interested in conditioning on the target class.

As mentioned in the body of the paper, we rely on a biased sampler involving separately specified values for the step size ϵ and the standard deviation σ of the stochastic term involving $\mathbf{r}$. Formally, our biased sampler performs updates as follows:

$$\hat{\mathbf{x}}_{j+1} \leftarrow \hat{\mathbf{x}}_j - \frac{\phi}{2}\mathcal{E}_\theta(\hat{\mathbf{x}}_j|\mathbf{y}^+) + \sigma \mathbf{r}_j, \quad j = 1, ..., J \tag{12}$$

Consistent with Grathwohl et al. (2020), we have specified $\phi = 2$ and $\sigma = 0.01$ as the default values for all of our experiments. Here we have deliberately departed slightly from the notation in Equation 2 to emphasize that we use fixed values for ϕ and σ, consistent with the related literature. The number of total SGLD steps J varies by dataset (Table 3). Following best practices, we initialize $\mathbf{x}_0$ randomly in 5% of all cases and sample from a buffer in all other cases. The buffer itself is randomly initialised and gradually grows to a maximum of 10,000 samples during training as $\hat{\mathbf{x}}_J$ is stored in each epoch (Du and Mordatch 2019; Grathwohl et al. 2020).

It is important to realise that sampling is done during each training epoch, which makes training Joint Energy Models significantly harder than conventional neural classifiers. In each epoch the generated (batch of) sample(s) $\hat{\mathbf{x}}_J$ is used as part of the generative loss component, which compares its energy to that of observed samples $\mathbf{x}$:

$$L_{\text{gen}}(\theta) \approx \mu_\theta(\mathbf{x})[\mathbf{y}] - \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}] \tag{13}$$

Our full training objective can be summarized as follows,

$$L_{\text{JEM}}(\theta) = L_{\text{clf}}(\theta) + L_{\text{gen}}(\theta) + \lambda L_{\text{reg}}(\theta) \tag{14}$$

where $L_{\text{reg}}(\theta)$ is a Ridge penalty (L2 norm) that regularises energy magnitudes for both observed and generated samples (Du and Mordatch 2019). We have used varying degrees of regularization depending on the dataset (λ in Table 3).

Contrary to existing work, we have not typically used the entire minibatch of training data for the generative loss component but found that using a subset of the minibatch was often sufficient in attaining decent generative performance (Table 3). This has helped to reduce the computational burden for our models, which should make it easier for others to reproduce our findings. Figures 4 and 5 show generated samples for our *MNIST* and *Moons* data, to provide a sense of their generative property.

Table 3: EBM hyperparemeter choices for our experiments.

Dataset	SGLD Steps	Batch Size	λ
Linearly Separable	50	50	0.10
Moons	30	10	0.10
Circles	30	50	0.01
California Housing	30	10	0.10
GMSC	30	10	0.10
German Credit	30	10	0.10
MNIST	25	10	0.01
Fashion MNIST	25	10	0.01

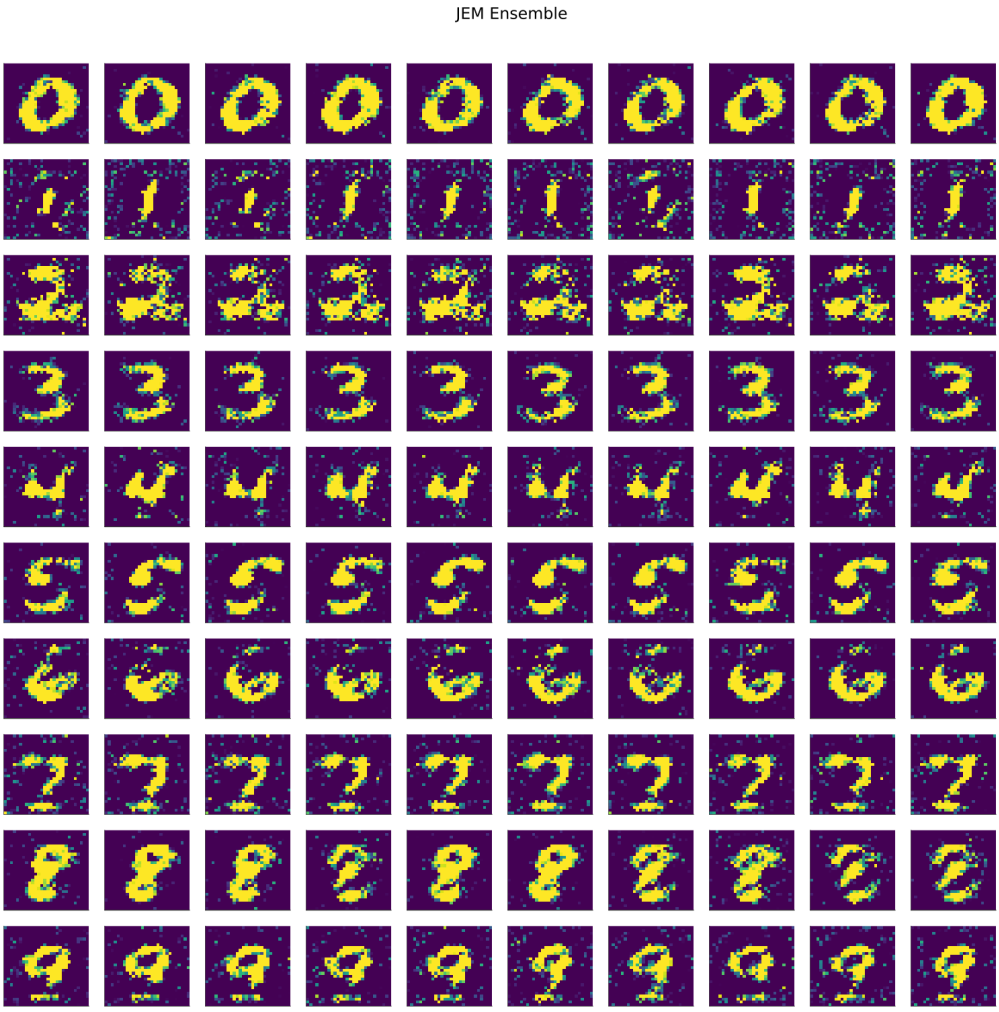

Figure 4: Conditionally generated *MNIST* images for our JEM Ensemble.

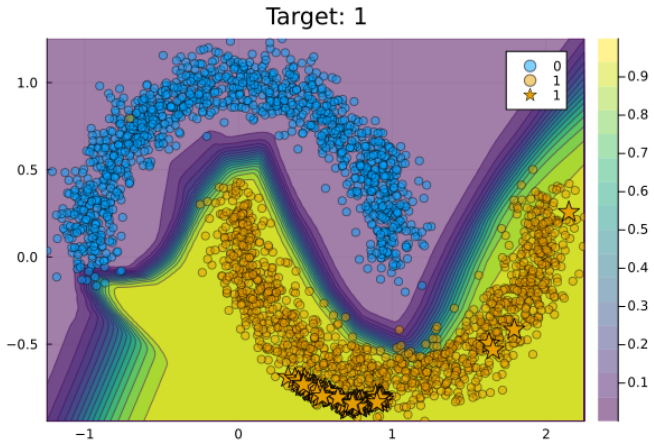

Figure 5: Conditionally generated samples (stars) for our *Moons* data using a JEM.

Inference: Quantifying Models' Generative Property At inference time, we assume no prior knowledge about the model's generative property. This means that we do not tab into the existing buffer of generated samples for our Joint Energy Models, but instead generate conditional samples from scratch. While we have relied on the default values $\epsilon = 2$ and $\sigma = 0.01$ also during inference, the number of total SGLD steps was set to $J = 500$ in all cases, so significantly higher than during training. For all of our synthetic datasets and models, we generated 50 conditional samples and then formed subsets containing the $n_E = 25$ lowest-energy samples. While in practice it would be sufficient to do this once for each model and dataset, we have chosen to perform sampling separately for each individual counterfactual in our experiments to account for stochasticity. To help reduce the computational burden for our real-world datasets we have generated only 10 conditional samples each time and used all of them in our counterfactual search. Using more samples, as we originally did, had no substantial impact on our results.

B Conformal Prediction

In this Appendix B we provide some more background on CP and explain in some more detail how we have used recent advances in Conformal Training for our purposes.

Background on CP 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. 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 $\mathcal{D}_n = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\ldots,n}$ 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 $\mathbf{y}_i$ (Angelopoulos and Bates 2021).

Finally, classification sets are formed as follows,

$$C_\theta(\mathbf{x}_i; \alpha) = \{\mathbf{y} : s(\mathbf{x}_i, \mathbf{y}) \leq \hat{q}\} \tag{15}$$

where $\hat{q}$ denotes the $(1-\alpha)$-quantile of $\mathcal{S}$ and α is a predetermined error rate. As the size of the calibration set increases, the probability that the classification set $C(\mathbf{x}_{\text{test}})$ for a newly arrived sample $\mathbf{x}_{\text{test}}$ does not cover the true test label $\mathbf{y}_{\text{test}}$ approaches α (Angelopoulos and Bates 2021).

Observe from Equation 15 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.

Differentiability 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. (2021) have recently proposed a framework for Conformal

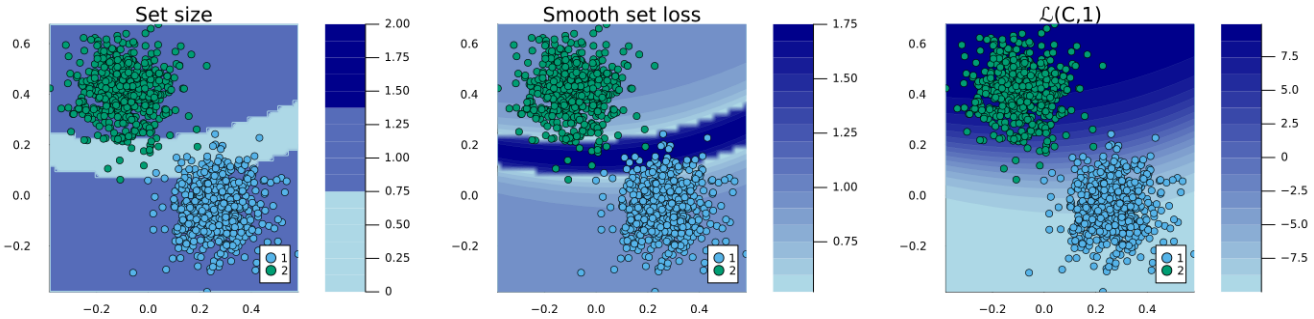

Figure 6: Prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a JEM trained on our *Linearly Separable* data.

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 4 in the body of this paper. Formally, it is defined as $C_{\theta,\mathbf{y}}(\mathbf{x}_i;\alpha) := \sigma\left((s(\mathbf{x}_i,\mathbf{y})-\alpha)T^{-1}\right)$ for $\mathbf{y} \in \mathcal{Y}$, where σ is the sigmoid function and T is a hyper-parameter used for temperature scaling (Stutz et al. 2021).

In addition to the smooth set size penalty, Stutz et al. (2021) also propose a configurable classification loss function, that can be used to enforce coverage. For *MNIST* data, we found that using this function generally improved the visual quality of the generated counterfactuals, so we used it in our experiments involving real-world data. For the synthetic dataset, visual inspection of the counterfactuals showed that using the configurable loss function sometimes led to overshooting: counterfactuals would end up deep inside the target domain but far away from the observed samples. For this reason, we instead relied on standard cross-entropy loss for our synthetic datasets. As we have noted in the body of the paper, more experimental work is certainly needed in this context. Figure 6 shows the prediction set size (left), smooth set size loss (centre) and configurable classification loss (right) for a *JEM* trained on our *Linearly Separable* data.

C *ECCCo*

In this section, we explain *ECCCo* in some more detail, briefly discuss convergence conditions for counterfactual explanations and provide details concerning the actual implementation of our framework in `Julia`.

Deriving the search objective The counterfactual search objective for *ECCCo* was introduced in Equation 9 in the body of the paper. We restate this equation here for reference:

$$\begin{aligned}\mathbf{Z}' = \arg\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{&\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) \\ &+ \lambda_2 \mathcal{E}_\theta(\mathbf{Z}', \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}');\alpha))\}\end{aligned} \tag{16}$$

We can make the connection to energy-based modeling more explicit by restating the counterfactual search objective in terms $L_{\text{JEM}}(\theta)$, which we defined in Equation 14. In particular, consider the following counterfactual search objective,

$$\mathbf{Z}' = \arg\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{L_{\text{JEM}}(\theta; M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}');\alpha))\} \tag{17}$$

where we have simply used the JEM loss function as $\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+)$.

Now note that aside from the additional penalties in Equation 16, the only key difference between our counterfactual search objective and the joint-energy training objective is the parameter that is being optimized. In joint-energy training we optimize the objective with respect to the network weights θ. Recall that $\mathcal{E}_\theta(\mathbf{x}|\mathbf{y}) = \mu_\theta(\mathbf{x})[\mathbf{y}]$. Then the partial gradient with respect to the generative loss component of $L_{\text{JEM}}(\theta)$ can be expressed as follows:

$$\nabla_\theta L_{\text{gen}}(\theta) = \nabla_\theta \mu_\theta(\mathbf{x})[\mathbf{y}] - \nabla_\theta \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}] \tag{18}$$

During the counterfactual search, we take the network parameters as fixed and instead optimize with respect to the counterfactual itself[2],

$$\nabla_{\mathbf{x}} L_{\text{gen}}(\theta) = \nabla_{\mathbf{x}} \mu_\theta(\mathbf{x})[\mathbf{y}^+] - \nabla_{\mathbf{x}} \mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}^+] = \nabla_{\mathbf{x}} \mu_\theta(\mathbf{x})[\mathbf{y}^+] = \nabla_{\mathbf{x}} \mathcal{E}_\theta(\mathbf{x}|\mathbf{y}^+) \tag{19}$$

[2]Here we omit the notion of a latent search space to make the comparison easier.

where the second term is equal to zero because $\mu_\theta(\hat{\mathbf{x}}_J)[\mathbf{y}]$ is invariant with respect to $\mathbf{x}$. Since this term has zero gradients, we can remove it from the loss function altogether. For the regularization loss component of $L_{\text{JEM}}(\theta)$ we can proceed analogously such that we can rewrite Equation 17 as follows:

$$\begin{aligned}\mathbf{Z}' = \arg\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{&\text{yloss}(M_\theta(f(\mathbf{Z}')), \mathbf{y}^+) + \mathcal{E}_\theta(f(\mathbf{Z}')|\mathbf{y}^+) + ||\mathcal{E}_\theta(f(\mathbf{Z}')|\mathbf{y}^+)||_2^2 \\ &+ \lambda_1 \text{dist}(f(\mathbf{Z}'), \mathbf{x}) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha))\}\end{aligned} \tag{20}$$

Now we notice that Equation 20 is equivalent to Equation 16 for $\lambda_2 = 1$. For the sake of simplicity, we omitted the regularization component from Equation 9 in the main text. Intuitively, taking iterative gradient steps according to Equation 19 has the effect of constraining the energy of the counterfactual until. The generative property of the underlying model implicitly enters this equation through θ.

The *ECCCo* algorithm Algorithm 1 describes how exactly *ECCCo* works. For the sake of simplicity and without loss of generality, we limit our attention to generating a single counterfactual $\mathbf{x}' = f(\mathbf{z}')$. The counterfactual state $\mathbf{z}'$ is initialized at the factual $\mathbf{x}$. Other forms of initialization are also suitable but not considered here. For example, one may choose at a small random perturbation to all features (Slack et al. 2021). Next, we calibrate the model M_θ through split conformal prediction. Finally, we search counterfactuals through gradient descent where $\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 9. The search terminates once the convergence criterium is met or the maximum number of iterations T has been exhausted. Note that the choice of convergence criterium has important implications on the final counterfactual which we explain below.

Algorithm 1 The *ECCCo* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: $\mathbf{x}'$

Initialize $\mathbf{z}' \leftarrow \mathbf{x}$
Run *SCP* for M_θ using $\mathcal{D}$ ▷ Calibrate model through split conformal prediction.
Initialize $t \leftarrow 0$
while *not converged* or $t < T$ **do** ▷ For convergence conditions see below.
    $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+; \Lambda, \alpha)$ ▷ Take gradient step of size η.
    $t \leftarrow t + 1$
end while
$\mathbf{x}' \leftarrow \mathbf{z}'$

The *ECCCo+* algorithm Algorithm 2 describes how exactly *ECCCo+* works. The only difference to *ECCCo* is that we encode and decode features using PCA. In particular, we let $f^{-1}(\mathbf{x})$ denote the projection of $\mathbf{x}$ to its first n_z principal components. Conversely, $f(\mathbf{z}')$ maps back from the projection to the feature space.

Algorithm 2 The *ECCCo+* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: $\mathbf{x}'$

Initialize $\mathbf{z}' \leftarrow f^{-1}(\mathbf{x})$ ▷ Map to counterfactual state space.
Run *SCP* for M_θ using $\mathcal{D}$ ▷ Calibrate model through split conformal prediction.
Initialize $t \leftarrow 0$
while *not converged* or $t < T$ **do** ▷ For convergence conditions see below.
    $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+; \Lambda, \alpha)$ ▷ Take gradient step of size η.
    $t \leftarrow t + 1$
end while
$\mathbf{x}' \leftarrow f(\mathbf{z}')$ ▷ Map back to feature space.

The *ECCCo-L1* algorithm Algorithm 3 describes a variation of *ECCCo* that we initally considered but ultimately discarded. For the sake of completeness we have included this approach here in the appendix. It generally yields very faithful counterfactuals but it is computationally much more expensive and struggles with plausibility.

Instead of constraining energy directly, this approach works under the premise of penalizing the distance between the counterfactual and samples generated through SGLD. The counterfactual state $\mathbf{z}'$ is initialized as in Algorithm 1. Next, we generate $n_{\mathcal{B}}$ conditional samples $\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}$ using SGLD (Equation 2) and store the n_E instances with the lowest energy. We then calibrate the model M_θ through split conformal prediction. Finally, we search counterfactuals through gradient descent where

$\mathcal{L}(\mathbf{z}', \mathbf{y}^+, \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}; \Lambda, \alpha)$ denotes our loss function defined in Equation 9, but instead of constraining energy directly we use Equaqtion 6 (unfaithfulness metric) as a penalty.

Algorithm 3 The *ECCCo-L1* generator

Input: $\mathbf{x}, \mathbf{y}^+, M_\theta, f, \Lambda = [\lambda_1, \lambda_2, \lambda_3], \alpha, \mathcal{D}, T, \eta, n_\mathcal{B}, n_E$ where $M_\theta(\mathbf{x}) \neq \mathbf{y}^+$
Output: $\mathbf{x}'$

Initialize $\mathbf{z}' \leftarrow \mathbf{x}$
Generate $\left\{\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}\right\}_{n_\mathcal{B}} \leftarrow p_\theta(\mathbf{x}_{\mathbf{y}^+})$ ▷ Generate $n_\mathcal{B}$ samples using SGLD (Equation 2).
Store $\widehat{\mathbf{X}}_{\theta,\mathbf{y}^+} \leftarrow \left\{\hat{\mathbf{x}}_{\theta,\mathbf{y}^+}\right\}_{n_\mathcal{B}}$ ▷ Choose n_E lowest-energy samples.
Run *SCP* for M_θ using $\mathcal{D}$ ▷ Calibrate model through split conformal prediction.
Initialize $t \leftarrow 0$
while *not converged* or $t < T$ **do** ▷ For convergence conditions see below.
    $\mathbf{z}' \leftarrow \mathbf{z}' - \eta \nabla_{\mathbf{z}'} \mathcal{L}(\mathbf{z}', \mathbf{y}^+, \widehat{\mathbf{X}}_{\theta,\mathbf{y}^+}; \Lambda, \alpha)$ ▷ Take gradient step of size η.
    $t \leftarrow t + 1$
end while
$\mathbf{x}' \leftarrow \mathbf{z}'$

A Note on Convergence Convergence is not typically discussed much in the context of CE, even though it has important implications on outcomes. One intuitive way to specify convergence is in terms of threshold probabilities: once the predicted probability $p(\mathbf{y}^+|\mathbf{x}')$ exceeds some user-defined threshold γ such that the counterfactual is valid, we could consider the search to have converged. In the binary case, for example, convergence could be defined as $p(\mathbf{y}^+|\mathbf{x}') > 0.5$ in this sense. Note, however, how this can be expected to yield counterfactuals in the proximity of the decision boundary, a region characterized by high aleatoric uncertainty. In other words, counterfactuals generated in this way would generally not be plausible. To avoid this from happening, we specify convergence in terms of gradients approaching zero for all our experiments and all of our generators. This is allows us to get a cleaner read on how the different counterfactual search objectives affect counterfactual outcomes.

`ECCCo.jl` The core part of our code base is integrated into a larger ecosystem of `Julia` packages that we are actively developing and maintaining. To avoid compromising the double-blind review process, we only provide a link to an anonymized repository at this stage: https://anonymous.4open.science/r/ECCCo-1252/README.md.

D Experimental Setup

In our experiments we always generate multiple counterfactuals for each model and generator. Each time the factual and target class is drawn randomly. For each generator and model we choose $n_f = 100$ factuals for all of our synthetic and vision data. For tabular data we choose $n_f = 25$ because larger values made grid search computationally prohibitive. For vision data, grid search was computationally prohibitve in any case, so hyperparameters were tuned manually. For all other datasets grid search was performed over different combinations of penalty strengths and optimizer steps sizes (details of which can be found in the code). To select the final hyperparameter setting we used the unfaithfulness metric as our criterion.

Table 4 provides an overview of all parameters related to our experiments. The *GMSC* data were randomly undersampled for balancing purposes and all features were standardized. *MNIST* data was also randomly undersampled for reasons outlined below. Pixel values were preprocessed to fall in the range of $[-1, 1]$ and a small Gaussian noise component ($\sigma = 0.03$) was added to training samples following common practice in the EBM literature. All of our models were trained through mini-batch training using the Adam optimiser (Kingma and Ba (2014)). Table 5 shows standard evaluation metrics measuring the predictive performance of our different models grouped by dataset. These measures were computed on test data.

Table 6 summarises our hyperparameter choices for the counterfactual generators where η denotes the learning rate used for Stochastic Gradient Descent (SGD) and λ_1, λ_2, λ_3 represent the chosen penalty strengths (Equations 1 and 9). Here λ_1 also refers to the chosen penalty for the distance from factual values that applies to both *Wachter* and *REVISE*, but not *Schut* which is penalty-free. *Schut* is also the only generator that uses JSMA instead of SGD for optimization.

E Compute

Research reported in this work was partially or completely facilitated by computational resources and support of the Delft-Blue (Delft High Performance Computing Centre DHPC) and the Delft AI Cluster (DAIC: https://doc.daic.tudelft.nl/) at TU Delft.

For grid search, we used 300 CPUs for tabular real-world datasets (<1.5 hours each), 150 CPUs for synthetic datasets (≈1 hour each) and 100 CPUs for vision datasets (<4 hours each), where we ran a smaller grid search for the latter. To generate the final results reported in the tables we used 300 CPUs for all datasets (<1.5 hours each) except the vision datasets. For the latter, we used 50 CPUs for smaller final experiments and at longer run times (≈5 hours each).

Table 4: Paremeter choices for our experiments.

		Network Architecture				Training	
Dataset	Sample Size	Hidden Units	Hidden Layers	Activation	Ensemble Size	Epochs	Batch Size
Linearly Separable	1000	16	3	swish	5	100	100
Moons	2500	32	3	relu	5	500	128
Circles	1000	32	1	swish	5	100	100
California Housing	16500	32	3	relu	5	100	128
GMSC	13370	32	3	relu	5	100	128
German Credit	800	32	3	relu	5	100	80
MNIST	10000	32	1	relu	5	100	128
Fashion MNIST	10000	32	2	relu	5	100	128

Table 5: Various standard performance metrics for our different models grouped by dataset.

		Performance Metrics		
Dataset	Model	Accuracy	Precision	F1-Score
Linearly Separable	JEM	0.98	0.98	0.98
	JEM Ensemble	0.99	0.99	0.99
	MLP	0.99	0.99	0.99
	MLP Ensemble	0.99	0.99	0.99
Moons	JEM	1.00	1.00	1.00
	JEM Ensemble	1.00	1.00	1.00
	MLP	1.00	1.00	1.00
	MLP Ensemble	1.00	1.00	1.00
Circles	JEM	1.00	1.00	1.00
	JEM Ensemble	1.00	1.00	1.00
	MLP	1.00	1.00	1.00
	MLP Ensemble	1.00	1.00	1.00
California Housing	JEM	0.87	0.87	0.87
	JEM Ensemble	0.87	0.87	0.87
	MLP	0.89	0.89	0.89
	MLP Ensemble	0.89	0.89	0.89
GMSC	JEM	0.75	0.76	0.74
	JEM Ensemble	0.74	0.75	0.74
	MLP	0.74	0.75	0.74
	MLP Ensemble	0.74	0.74	0.74
German Credit	JEM	0.54	0.60	0.47
	JEM Ensemble	0.55	0.68	0.46
	MLP	0.54	0.76	0.42
	MLP Ensemble	0.51	0.75	0.36
MNIST	JEM	0.84	0.85	0.84
	JEM Ensemble	0.90	0.90	0.90
	LeNet-5	0.98	0.98	0.98
	MLP	0.95	0.95	0.95
	MLP Ensemble	0.95	0.95	0.95
Fashion MNIST	JEM	0.62	0.70	0.62
	JEM Ensemble	0.78	0.78	0.78
	LeNet-5	0.83	0.84	0.82
	MLP	0.82	0.83	0.82
	MLP Ensemble	0.84	0.84	0.84

Table 6: Generator hyperparameters: the optimiser step size (η); penalty strengths where λ_1 applies to all generators but *Schut* and the other parameter are specific to *ECCCo*; finally, the strength for the Ridge penalty on energy for *ECCCo*.

Dataset	η	λ_1	λ_2	λ_3	Ridge penalty
Linearly Separable	0.01	0.10	0.1	0.05	0.0
Moons	0.01	0.10	0.1	0.50	0.0
Circles	0.05	0.10	0.1	0.05	1.0
California Housing	0.05	0.10	0.1	0.10	0.0
GMSC	0.05	0.10	0.1	0.10	0.0
German Credit	0.05	0.20	0.2	0.20	0.5
MNIST	0.10	0.01	0.1	0.30	0.0
Fashion MNIST	0.10	0.01	0.1	0.30	0.0

F Results

Figures 7 to 11 show examples of counterfactuals for *MNIST* generated by *ECCCo+* for our different models. Original images are shown on the diagonal and the corresponding counterfactuals are plotted across rows. Figures 12 to 16 show the same examples but for *REVISE*. Both counterfactual generators have access to the same optimizer. While the results for *REVISE* look fairly poor here, we have observed better results for optizers with higher step sizes. Note that the seemingly poor performance by *REVISE* upon visual inspection is not driven by a weak surrogate VAE: Figure 17 shows image reconstructions generated by the VAE.

Tables 7 to 14 reports all of the evaluation metrics we have computed. Tables 15 to 22 reports the same metrics for the subset of valid counterfactuals. The 'Unfaithfulness' and 'Implausibility' metrics have been discussed extensively in the body of the paper. The 'Cost' metric relates to the distance between the factual and the counterfactual and is measured using the L1 Norm. The 'Redundancy' metric measures sparsity in is defined as the percentage of features that remain unperturbed (higher is better). The 'Uncertainty' metric is just the average value of the smooth set size penalty (Equation 4). Finally, 'Validity' is the percentage of valid counterfactuals.

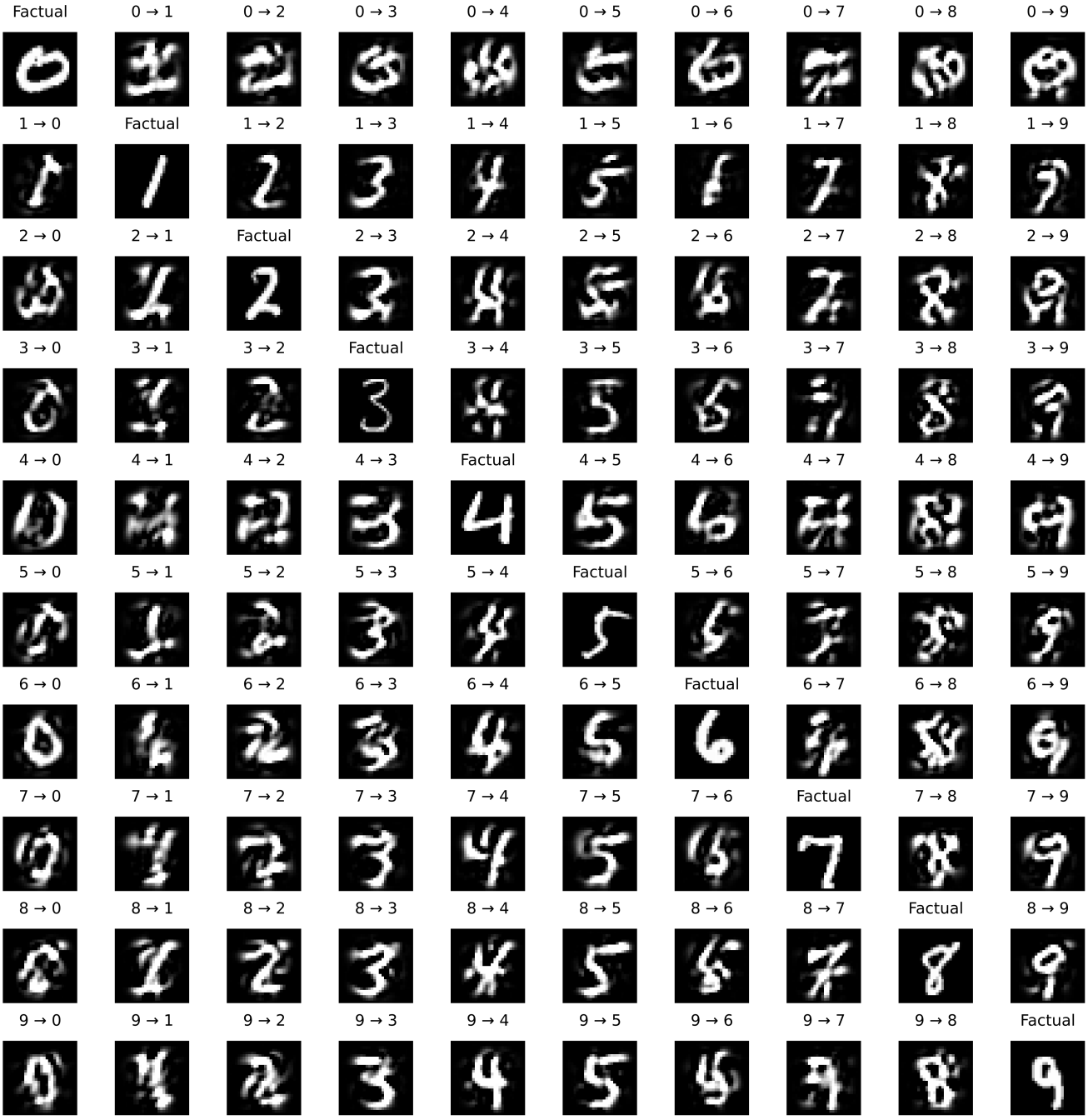

Figure 7: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a LeNet-5 *CNN*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

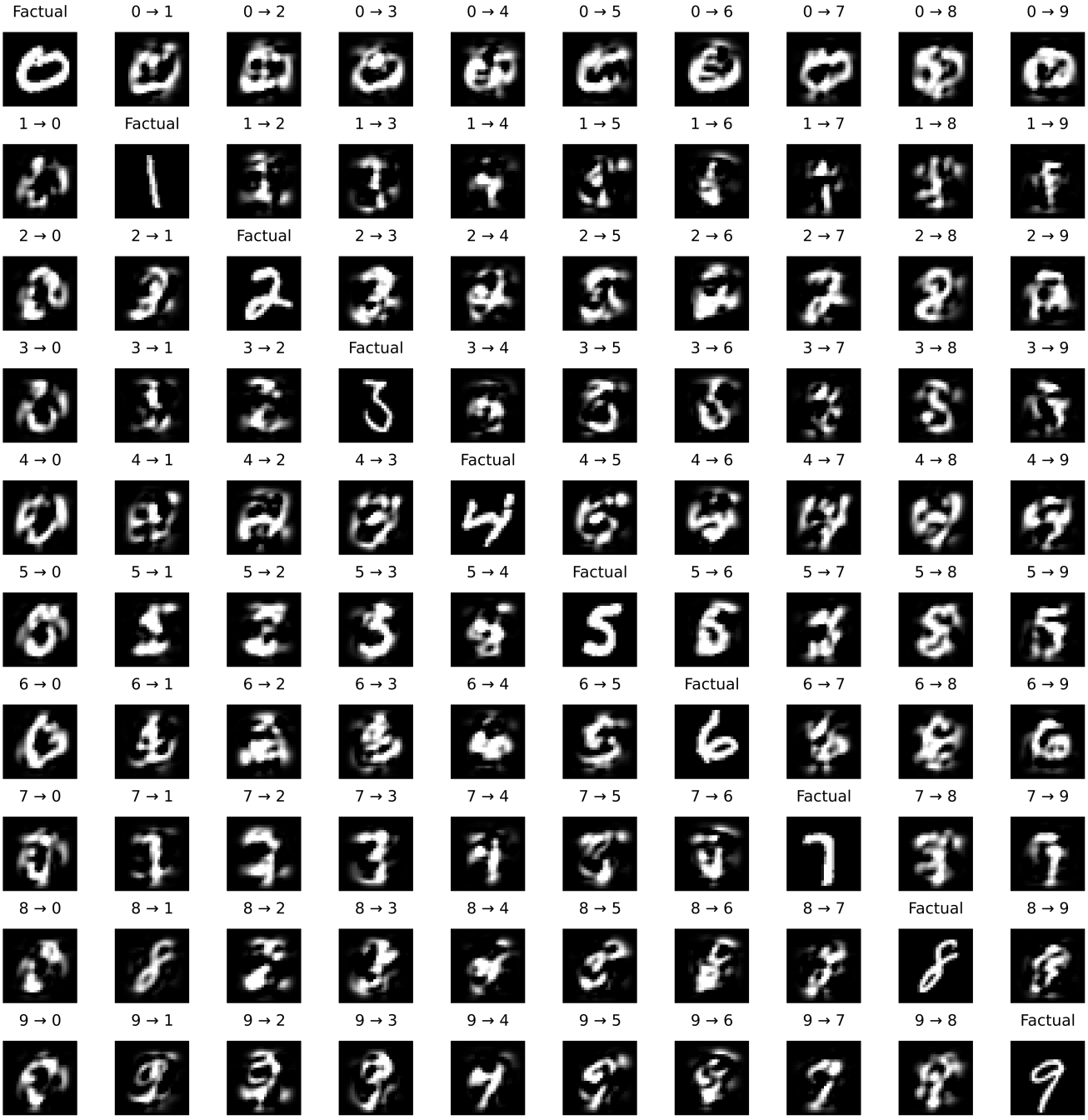

Figure 8: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is an *MLP*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

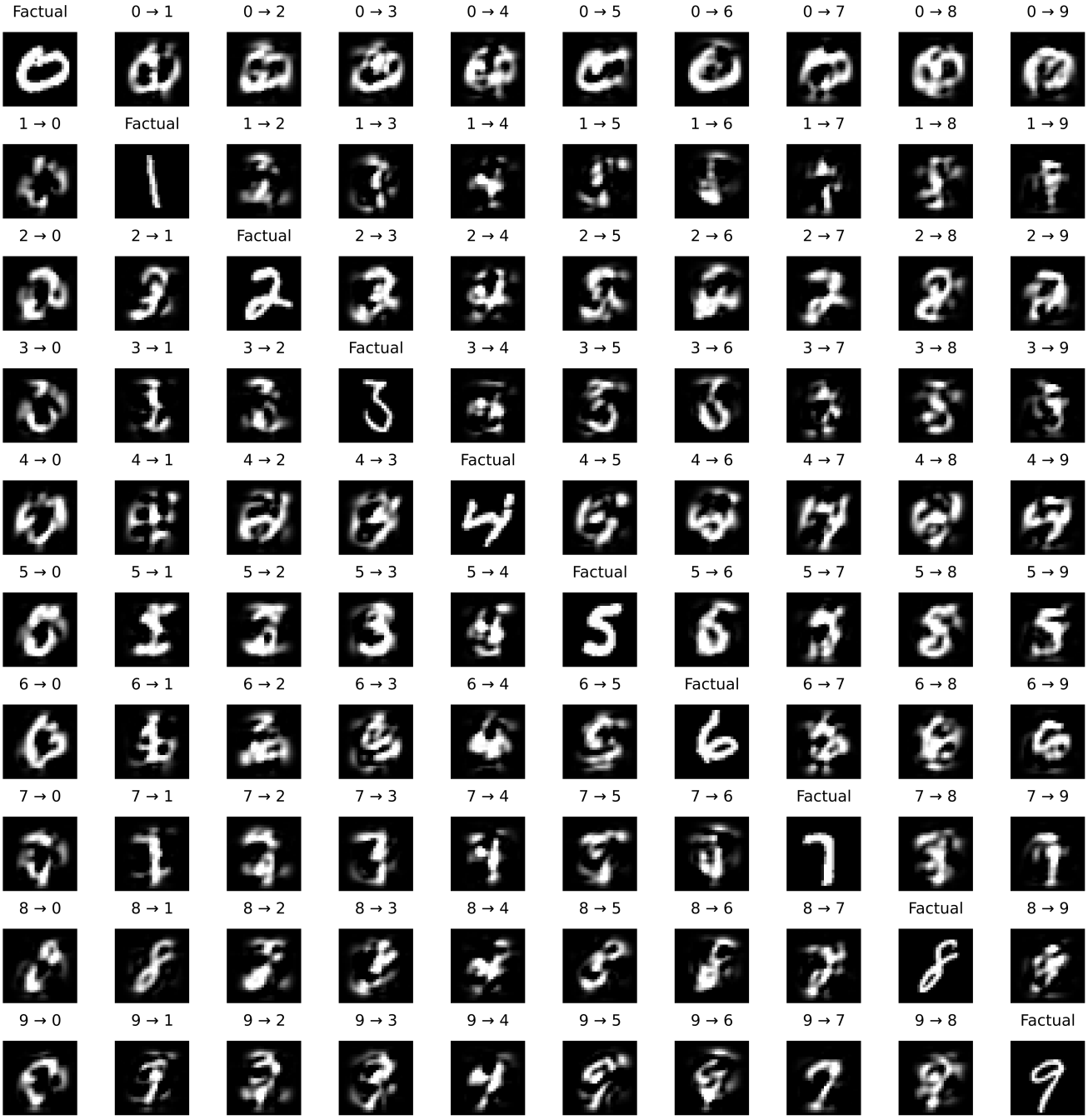

Figure 9: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is an *MLP* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

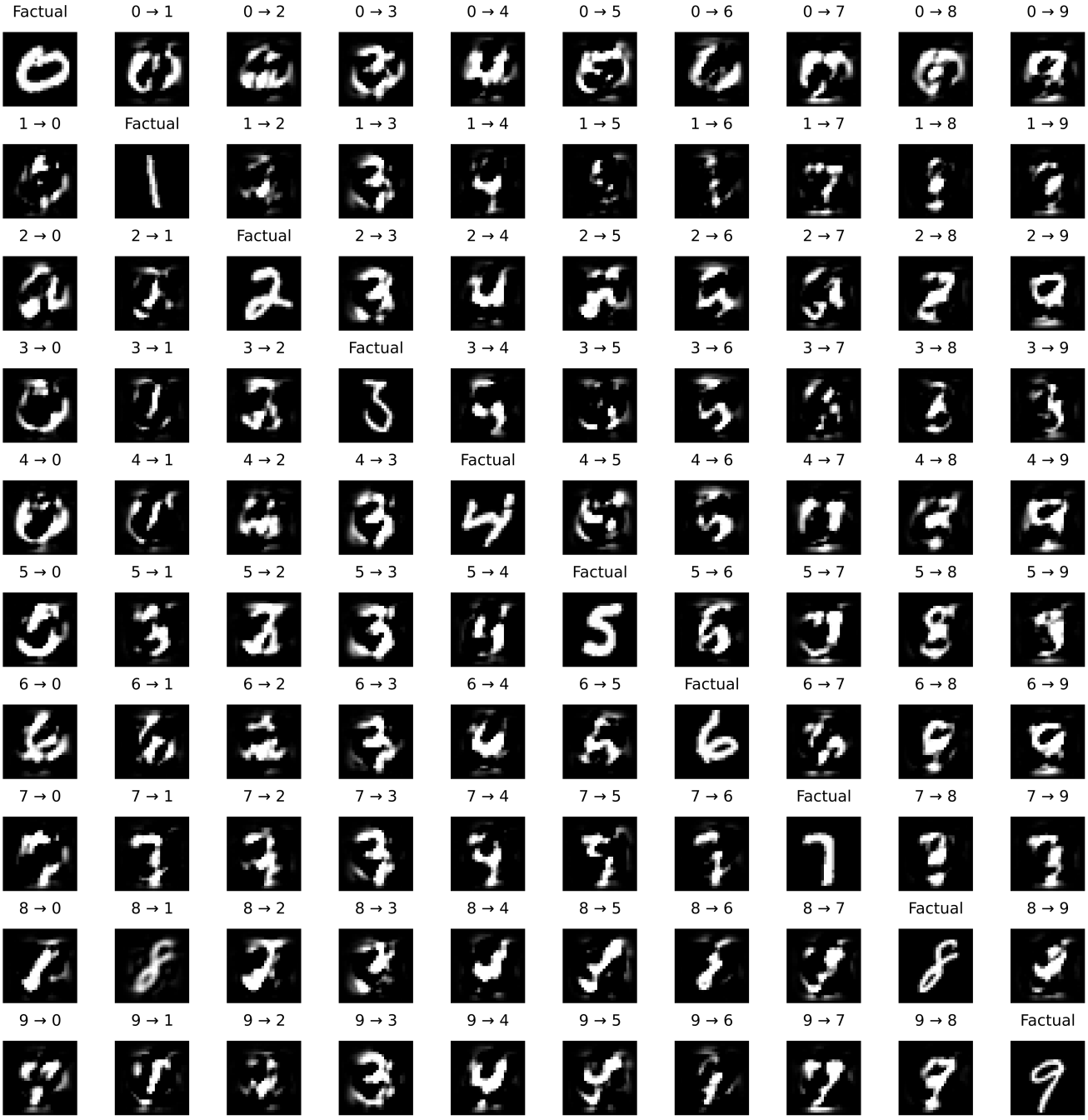

Figure 10: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a *JEM*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

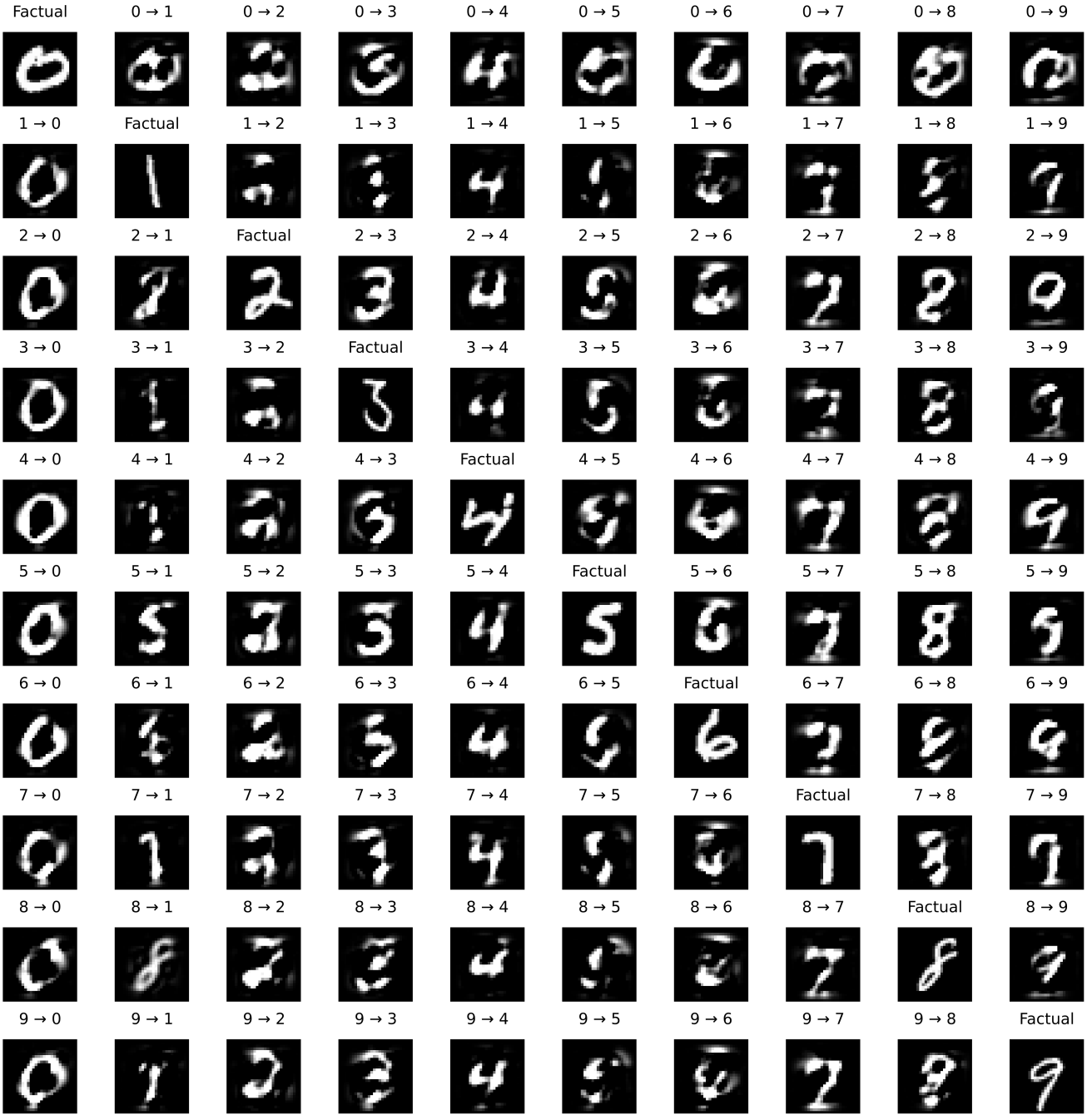

Figure 11: Counterfactuals for *MNIST* data generated by *ECCCo+*. The underlying model is a *JEM* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

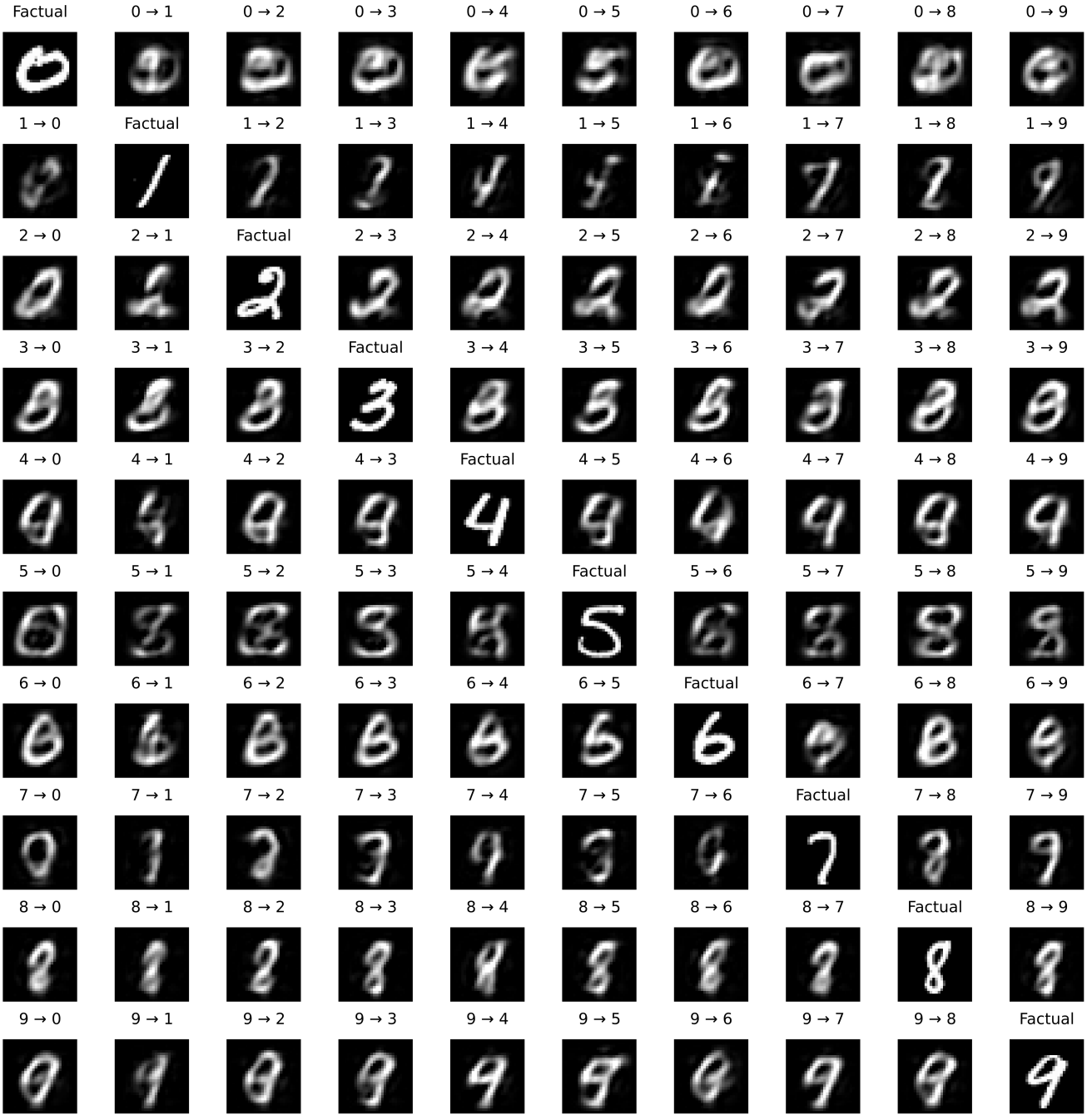

Figure 12: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a LeNet-5 *CNN*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

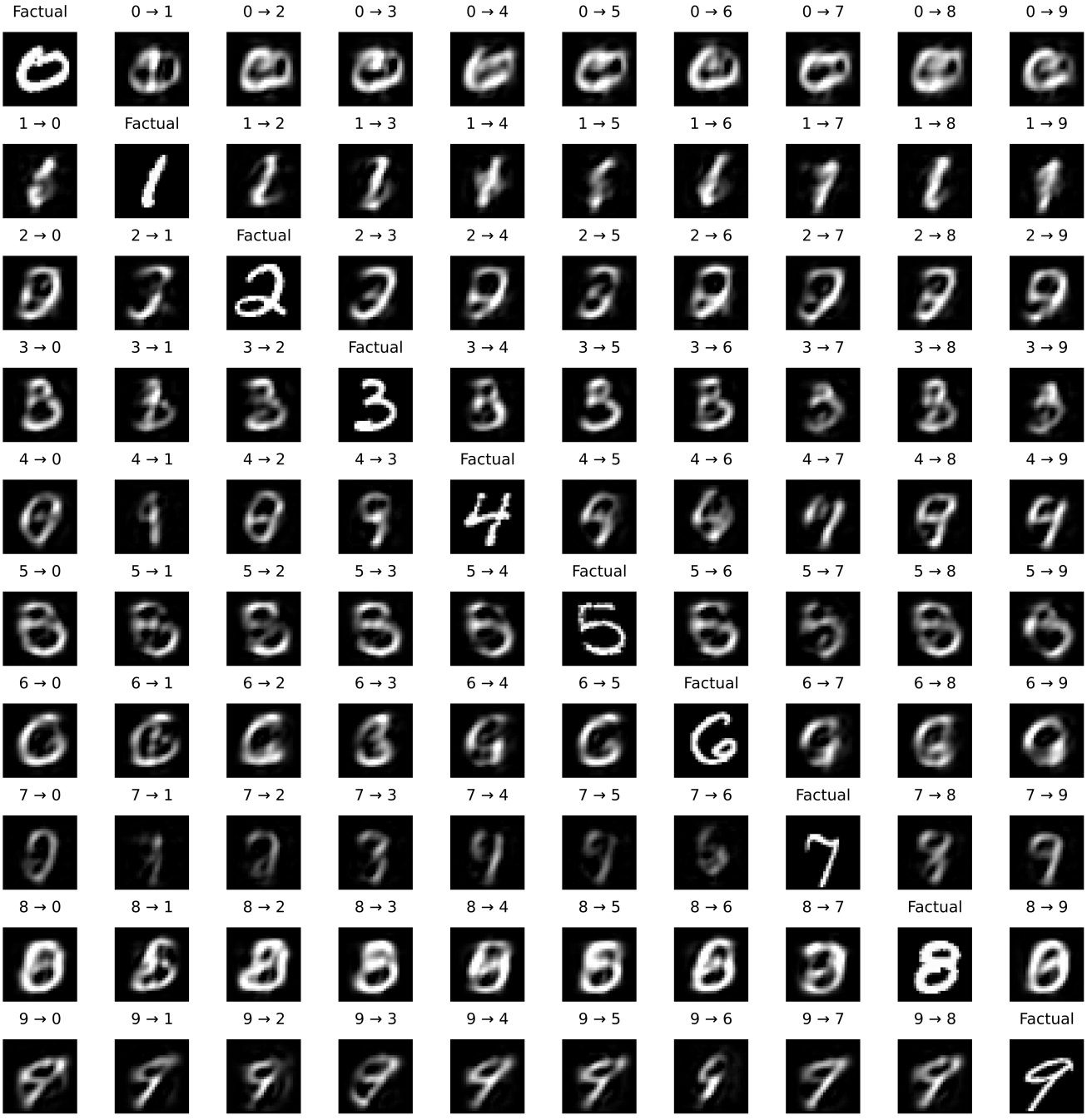

Figure 13: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is an *MLP*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

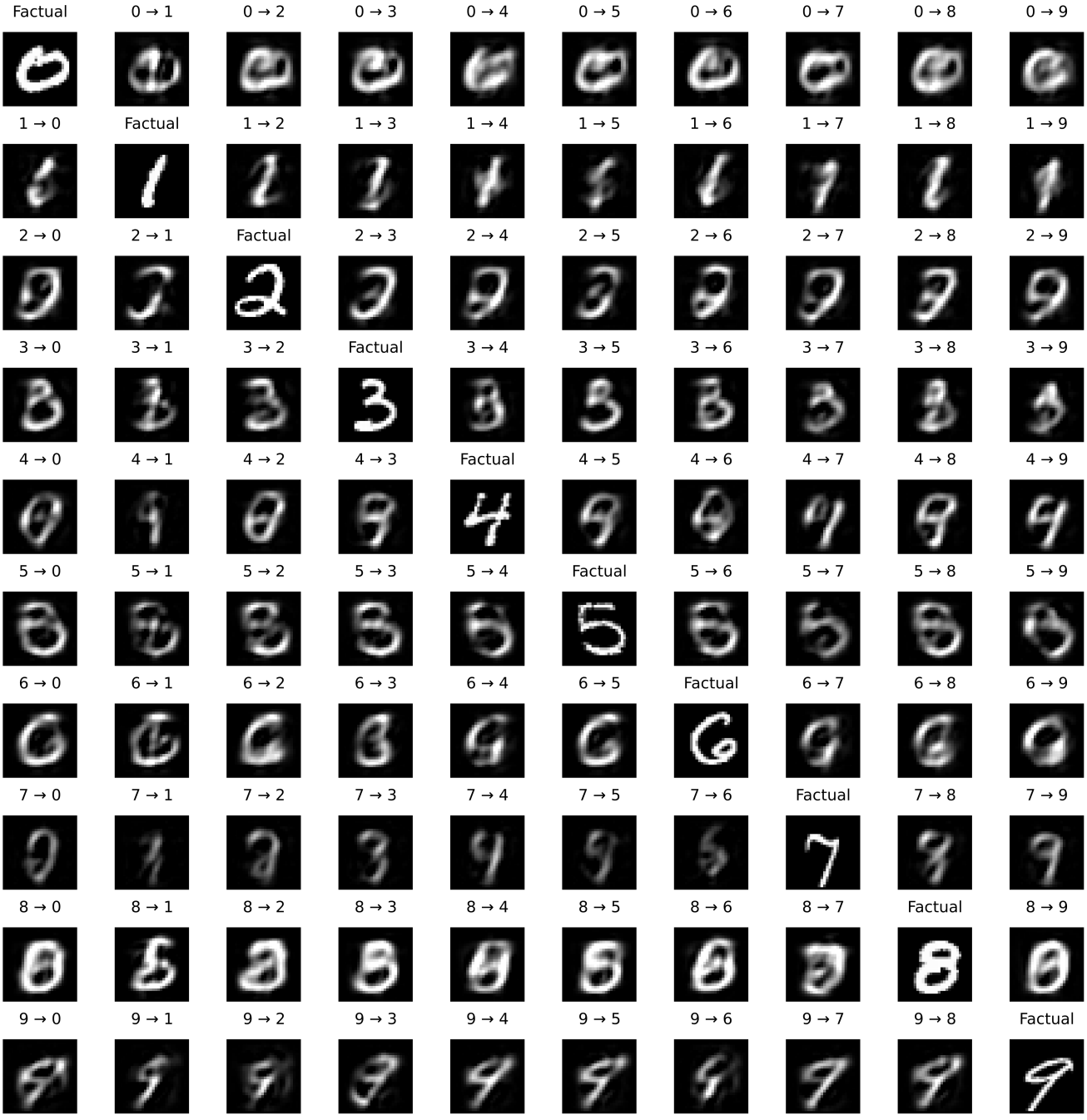

Figure 14: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is an *MLP* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

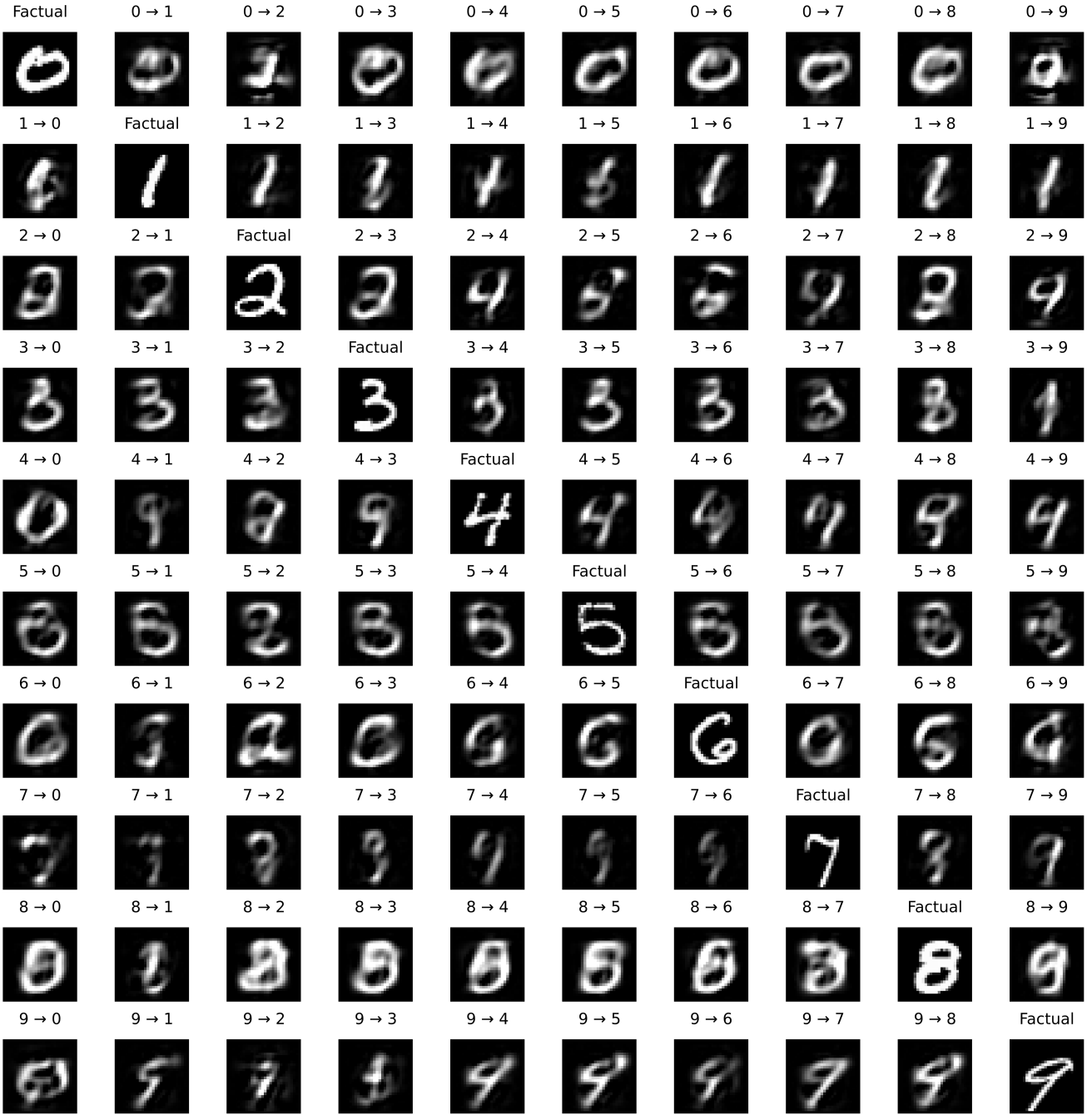

Figure 15: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a *JEM*. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

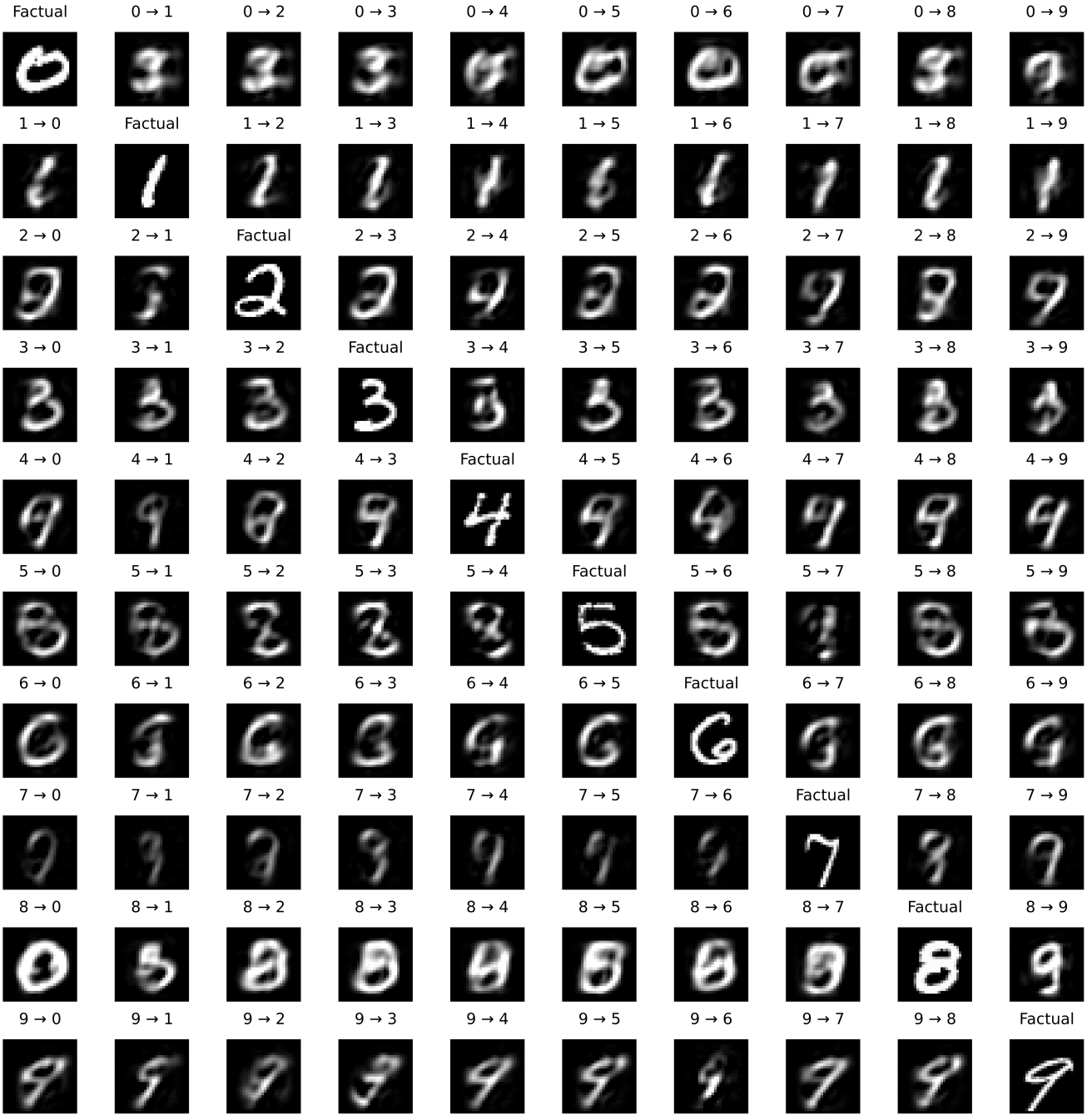

Figure 16: Counterfactuals for *MNIST* data generated by *REVISE*. The underlying model is a *JEM* ensemble. Original images are shown on the diagonal with the corresponding counterfactuals plotted across rows.

Table 7: All results for Linearly Separable dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.06 ± 0.01**	**0.16 ± 0.02****	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.06 ± 0.01****	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.09 ± 0.01**	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	**0.00 ± 0.00**	**1.00 ± 0.00**
	ECCCo (no CP)	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	**0.00 ± 0.00**	**1.00 ± 0.00**
	ECCCo (no EBM)	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.27 ± 0.01	0.18 ± 0.00**	0.00 ± 0.00	**0.42 ± 0.01****	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.29 ± 0.01	0.29 ± 0.00	0.00 ± 0.00	1.40 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	Wachter	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
JEM Ensemble	ECCCo-L1	**0.02 ± 0.00****	0.15 ± 0.01	**0.00 ± 0.00**	0.98 ± 0.02*	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	0.02 ± 0.00**	0.15 ± 0.01	**0.00 ± 0.00**	0.98 ± 0.02*	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.05 ± 0.00**	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.12 ± 0.01**	0.07 ± 0.00**	**0.00 ± 0.00**	0.97 ± 0.02**	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	0.12 ± 0.01**	**0.07 ± 0.00****	**0.00 ± 0.00**	0.97 ± 0.02**	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	0.15 ± 0.01	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.25 ± 0.00	0.18 ± 0.00	0.44 ± 0.05	**0.42 ± 0.01****	0.00 ± 0.00	0.51 ± 0.06
	Schut	0.20 ± 0.01	0.27 ± 0.00	**0.00 ± 0.00**	1.38 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	Wachter	0.14 ± 0.01	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
MLP	ECCCo-L1	0.12 ± 0.00**	**0.06 ± 0.00****	**0.00 ± 0.00**	0.91 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.12 ± 0.00****	0.06 ± 0.00**	**0.00 ± 0.00**	0.90 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	0.46 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.53 ± 0.00	0.18 ± 0.00	0.52 ± 0.01	**0.41 ± 0.02****	0.00 ± 0.00	0.53 ± 0.01
	Schut	0.48 ± 0.00	0.31 ± 0.01	**0.00 ± 0.00**	0.70 ± 0.02	**0.46 ± 0.01****	**1.00 ± 0.00**
	Wachter	0.46 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
MLP Ensemble	ECCCo-L1	**0.12 ± 0.00****	0.06 ± 0.00**	**0.00 ± 0.00**	0.91 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	0.12 ± 0.00**	**0.06 ± 0.00****	**0.00 ± 0.00**	0.91 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	0.45 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.53 ± 0.00	0.18 ± 0.00	0.52 ± 0.01	**0.42 ± 0.01****	0.00 ± 0.00	0.53 ± 0.01
	Schut	0.48 ± 0.00	0.32 ± 0.01	**0.00 ± 0.00**	0.73 ± 0.01	**0.42 ± 0.02****	**1.00 ± 0.00**
	Wachter	0.45 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**

Table 8: All results for Circles dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.28 ± 0.01**	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo-L1 (no CP)	**0.27 ± 0.01****	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo-L1 (no EBM)	0.28 ± 0.01**	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	ECCCo	0.69 ± 0.03	0.69 ± 0.00	0.00 ± 0.00	0.76 ± 0.02	0.00 ± 0.00	0.53 ± 0.00
	ECCCo (no CP)	0.68 ± 0.03	0.69 ± 0.00	0.00 ± 0.00	0.76 ± 0.02	0.00 ± 0.00	0.53 ± 0.00
	ECCCo (no EBM)	0.69 ± 0.02	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
	REVISE	0.32 ± 0.01**	0.68 ± 0.00	0.00 ± 0.00	0.94 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.49 ± 0.02**	**0.29 ± 0.00****	0.00 ± 0.00	**0.48 ± 0.01****	**0.41 ± 0.01****	**1.00 ± 0.00****
	Wachter	0.70 ± 0.02	0.68 ± 0.01	0.00 ± 0.00	0.70 ± 0.01	0.00 ± 0.00	0.53 ± 0.01
JEM Ensemble	ECCCo-L1	0.21 ± 0.01**	0.43 ± 0.03	0.01 ± 0.01**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo-L1 (no CP)	**0.21 ± 0.01****	0.43 ± 0.03	0.02 ± 0.01	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo-L1 (no EBM)	0.22 ± 0.01**	0.43 ± 0.03	0.00 ± 0.00**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	ECCCo	0.52 ± 0.03	0.51 ± 0.02	**0.00 ± 0.00****	0.74 ± 0.03	0.00 ± 0.00	0.72 ± 0.04
	ECCCo (no CP)	0.52 ± 0.03	0.51 ± 0.02	0.00 ± 0.00**	0.74 ± 0.03	0.00 ± 0.00	0.72 ± 0.04
	ECCCo (no EBM)	0.52 ± 0.03	0.43 ± 0.03	0.00 ± 0.00**	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
	REVISE	0.28 ± 0.01**	0.68 ± 0.00	**0.00 ± 0.00****	0.93 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.45 ± 0.02**	**0.26 ± 0.00****	**0.00 ± 0.00****	**0.58 ± 0.01****	**0.38 ± 0.01****	**1.00 ± 0.00****
	Wachter	0.53 ± 0.03	0.43 ± 0.03	0.02 ± 0.01	0.60 ± 0.02	0.00 ± 0.00	0.76 ± 0.04
MLP	ECCCo-L1	0.29 ± 0.01**	0.28 ± 0.01	0.01 ± 0.00**	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.29 ± 0.01****	0.28 ± 0.00	0.02 ± 0.01**	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.71 ± 0.04*	0.56 ± 0.00	**0.00 ± 0.00****	1.38 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	0.71 ± 0.04*	0.56 ± 0.00	**0.00 ± 0.00****	1.38 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.64 ± 0.00**	0.68 ± 0.00	**0.00 ± 0.00****	0.94 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.78 ± 0.02	**0.25 ± 0.00****	**0.00 ± 0.00****	**0.55 ± 0.01****	**0.39 ± 0.01****	**1.00 ± 0.00**
	Wachter	0.77 ± 0.02	0.29 ± 0.00	0.04 ± 0.01	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
MLP Ensemble	ECCCo-L1	0.29 ± 0.01**	0.27 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.29 ± 0.01****	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.64 ± 0.00**	0.68 ± 0.00	0.00 ± 0.00	0.93 ± 0.01	0.00 ± 0.00	0.47 ± 0.00
	Schut	0.78 ± 0.02	**0.25 ± 0.00****	0.00 ± 0.00	**0.58 ± 0.01**	**0.38 ± 0.02****	**1.00 ± 0.00**
	Wachter	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	**1.00 ± 0.00**

Table 9: All results for Moons dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.29 ± 0.02**	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.29 ± 0.02****	0.42 ± 0.04	0.15 ± 0.03	**1.02 ± 0.03**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.02**	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.55 ± 0.06**	0.15 ± 0.01**	**0.01 ± 0.01****	1.28 ± 0.04	0.00 ± 0.00	**1.00 ± 0.00****
	ECCCo (no CP)	0.57 ± 0.05**	**0.15 ± 0.01****	0.01 ± 0.01**	1.28 ± 0.04	0.00 ± 0.00	**1.00 ± 0.00****
	ECCCo (no EBM)	0.69 ± 0.04	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.54 ± 0.05**	0.31 ± 0.02**	0.15 ± 0.06	1.41 ± 0.12	0.00 ± 0.00	0.84 ± 0.07
	Schut	0.72 ± 0.05	0.46 ± 0.04	0.19 ± 0.02	1.12 ± 0.03	**0.06 ± 0.02****	1.00 ± 0.01
	Wachter	0.69 ± 0.05	0.42 ± 0.04	0.14 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.26 ± 0.01**	0.18 ± 0.01	0.02 ± 0.01	1.04 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.26 ± 0.02****	0.18 ± 0.01	0.04 ± 0.01	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.27 ± 0.01**	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.47 ± 0.03**	0.13 ± 0.00**	**0.00 ± 0.00****	1.50 ± 0.04	0.00 ± 0.00	**1.00 ± 0.00****
	ECCCo (no CP)	0.46 ± 0.04**	**0.13 ± 0.00****	**0.00 ± 0.00****	1.50 ± 0.04	0.00 ± 0.00	**1.00 ± 0.00****
	ECCCo (no EBM)	0.68 ± 0.02	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.64 ± 0.02*	0.34 ± 0.01	0.40 ± 0.05	1.39 ± 0.07	0.00 ± 0.00	0.70 ± 0.04
	Schut	0.52 ± 0.03**	0.24 ± 0.01	**0.00 ± 0.00****	1.81 ± 0.03	**0.05 ± 0.01****	1.00 ± 0.00
	Wachter	0.67 ± 0.03	0.18 ± 0.01	0.01 ± 0.01	**1.02 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	**0.45 ± 0.02****	0.34 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	0.45 ± 0.03**	0.35 ± 0.02	0.31 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	1.33 ± 0.06**	0.59 ± 0.03	**0.00 ± 0.00****	2.84 ± 0.09	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	1.32 ± 0.09*	0.59 ± 0.03	**0.00 ± 0.00****	2.84 ± 0.09	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	1.49 ± 0.06	**0.27 ± 0.01****	0.29 ± 0.03	1.22 ± 0.07**	0.00 ± 0.00	0.93 ± 0.03
	Schut	1.58 ± 0.09	0.52 ± 0.04	**0.00 ± 0.00****	**0.72 ± 0.02****	**0.21 ± 0.02****	0.79 ± 0.04
	Wachter	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
MLP Ensemble	ECCCo-L1	0.44 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02	1.43 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.44 ± 0.02****	0.28 ± 0.02	0.24 ± 0.03	1.43 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	1.24 ± 0.07**	0.56 ± 0.02	**0.00 ± 0.00****	2.81 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no CP)	1.23 ± 0.09**	0.56 ± 0.02	**0.00 ± 0.00****	2.81 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	1.50 ± 0.03	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	1.58 ± 0.06	**0.26 ± 0.01****	0.31 ± 0.03	1.25 ± 0.07**	0.00 ± 0.00	0.94 ± 0.03
	Schut	1.64 ± 0.09	0.36 ± 0.02	0.03 ± 0.01**	**0.70 ± 0.02****	**0.24 ± 0.03****	0.89 ± 0.03
	Wachter	1.51 ± 0.04	0.29 ± 0.02	0.24 ± 0.02	1.42 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**

Table 10: All results for California Housing dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.98 ± 0.05**	1.04 ± 0.12	0.24 ± 0.02	1.22 ± 0.07	0.00 ± 0.00	0.98 ± 0.01
	ECCCo-L1 (no CP)	**0.97 ± 0.05****	1.04 ± 0.12	0.25 ± 0.02	1.18 ± 0.07	0.00 ± 0.00	0.98 ± 0.01
	ECCCo-L1 (no EBM)	1.00 ± 0.06**	1.07 ± 0.12	0.24 ± 0.02	1.12 ± 0.07	0.01 ± 0.00	0.98 ± 0.01
	ECCCo	1.94 ± 0.13*	0.78 ± 0.08**	**0.18 ± 0.00****	2.05 ± 0.15	0.01 ± 0.00	**1.00 ± 0.00****
	ECCCo+	1.84 ± 0.12**	0.71 ± 0.07**	0.19 ± 0.00**	3.29 ± 0.18	0.00 ± 0.00	1.00 ± 0.01**
	ECCCo (no CP)	1.95 ± 0.14*	0.78 ± 0.08**	0.18 ± 0.00**	2.03 ± 0.16	0.00 ± 0.00	**1.00 ± 0.00****
	ECCCo (no EBM)	2.13 ± 0.11	1.07 ± 0.12	0.24 ± 0.02	1.12 ± 0.07	0.01 ± 0.00	0.98 ± 0.01
	REVISE	1.89 ± 0.15*	**0.62 ± 0.04****	0.27 ± 0.04	5.40 ± 0.31	0.00 ± 0.00	0.67 ± 0.06
	Schut	2.05 ± 0.13	1.14 ± 0.12	0.21 ± 0.01**	2.82 ± 0.08	**0.29 ± 0.01****	0.99 ± 0.01**
	Wachter	2.14 ± 0.13	1.07 ± 0.12	0.25 ± 0.02	**1.08 ± 0.07**	0.01 ± 0.00	0.98 ± 0.01
JEM Ensemble	ECCCo-L1	0.89 ± 0.05**	1.00 ± 0.17	0.13 ± 0.00**	1.38 ± 0.07	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**0.89 ± 0.06****	1.01 ± 0.17	0.14 ± 0.00*	1.34 ± 0.07	0.01 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	0.92 ± 0.06**	1.02 ± 0.18	0.14 ± 0.00*	1.27 ± 0.08	0.01 ± 0.00	**1.00 ± 0.00**
	ECCCo	1.41 ± 0.08**	0.71 ± 0.09**	0.11 ± 0.00**	2.28 ± 0.16	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo+	1.29 ± 0.08**	**0.62 ± 0.08****	0.11 ± 0.00**	3.41 ± 0.19	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.40 ± 0.08**	0.71 ± 0.09**	0.11 ± 0.00**	2.25 ± 0.16	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	1.71 ± 0.11	1.02 ± 0.18	0.14 ± 0.00*	1.27 ± 0.08	0.01 ± 0.00	**1.00 ± 0.00**
	REVISE	1.42 ± 0.14**	0.63 ± 0.04**	0.21 ± 0.05	5.29 ± 0.38	0.00 ± 0.00	0.62 ± 0.06
	Schut	1.61 ± 0.11	1.13 ± 0.18	**0.09 ± 0.00****	3.08 ± 0.10	**0.31 ± 0.02****	1.00 ± 0.00
	Wachter	1.72 ± 0.11	1.02 ± 0.18	0.14 ± 0.00	**1.23 ± 0.07**	0.01 ± 0.00	**1.00 ± 0.00**
MLP	ECCCo-L1	**1.22 ± 0.04****	1.18 ± 0.09	0.17 ± 0.01*	1.38 ± 0.09	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	1.23 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01	1.35 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.25 ± 0.04**	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	3.81 ± 0.10*	2.31 ± 0.09	**0.14 ± 0.01****	6.19 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	3.95 ± 0.10	1.31 ± 0.08	0.15 ± 0.01**	4.62 ± 0.17	0.00 ± 0.00	0.96 ± 0.02
	ECCCo (no CP)	3.80 ± 0.11*	2.31 ± 0.09	0.14 ± 0.01**	6.17 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	3.98 ± 0.08	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	3.91 ± 0.07*	**0.61 ± 0.04****	0.17 ± 0.04	5.37 ± 0.31	0.00 ± 0.00	0.73 ± 0.06
	Schut	3.99 ± 0.08	1.15 ± 0.09	0.15 ± 0.01**	1.59 ± 0.09	**0.48 ± 0.02****	0.99 ± 0.01
	Wachter	3.98 ± 0.07	1.19 ± 0.09	0.17 ± 0.01	**1.23 ± 0.09**	0.00 ± 0.00	**1.00 ± 0.00**
MLP Ensemble	ECCCo-L1	1.23 ± 0.05**	1.17 ± 0.17	0.13 ± 0.01**	1.21 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**1.22 ± 0.04****	1.17 ± 0.17	0.15 ± 0.01	1.18 ± 0.08	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	1.24 ± 0.04**	1.16 ± 0.17	0.14 ± 0.01**	1.08 ± 0.07	0.01 ± 0.00	**1.00 ± 0.00**
	ECCCo	3.71 ± 0.10**	1.98 ± 0.17	**0.09 ± 0.01****	4.77 ± 0.12	0.01 ± 0.00	**1.00 ± 0.00**
	ECCCo+	3.92 ± 0.09*	1.25 ± 0.14	0.14 ± 0.02	4.22 ± 0.16	0.00 ± 0.00	0.97 ± 0.02
	ECCCo (no CP)	3.72 ± 0.10**	1.98 ± 0.17	0.10 ± 0.01**	4.77 ± 0.12	0.01 ± 0.00	**1.00 ± 0.00**
	ECCCo (no EBM)	4.05 ± 0.08	1.16 ± 0.17	0.14 ± 0.01**	1.08 ± 0.07	0.01 ± 0.00	**1.00 ± 0.00**
	REVISE	3.98 ± 0.06*	**0.61 ± 0.03****	0.15 ± 0.03	5.42 ± 0.34	0.00 ± 0.00	0.74 ± 0.05
	Schut	4.02 ± 0.08	1.20 ± 0.18	0.10 ± 0.01**	1.55 ± 0.08	**0.58 ± 0.02****	1.00 ± 0.00
	Wachter	4.06 ± 0.09	1.16 ± 0.17	0.16 ± 0.01	**1.05 ± 0.07**	0.01 ± 0.00	**1.00 ± 0.00**

Table 11: All results for GMSC dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.94 ± 0.05****	1.11 ± 0.18	0.29 ± 0.01**	1.03 ± 0.04	0.13 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no CP)	0.96 ± 0.05**	1.10 ± 0.18	0.31 ± 0.01*	0.96 ± 0.04	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no EBM)	0.97 ± 0.05**	1.09 ± 0.18	0.31 ± 0.01*	0.89 ± 0.03	0.14 ± 0.01	0.98 ± 0.01
	ECCCo	1.85 ± 0.12	1.01 ± 0.18	0.34 ± 0.01	1.12 ± 0.05	0.13 ± 0.01	**0.99 ± 0.01***
	ECCCo+	1.72 ± 0.13*	0.90 ± 0.18*	0.35 ± 0.01	2.59 ± 0.12	0.09 ± 0.01	0.99 ± 0.01*
	ECCCo (no CP)	1.86 ± 0.12	1.00 ± 0.18	0.36 ± 0.01	1.06 ± 0.05	0.13 ± 0.01	**0.99 ± 0.01***
	ECCCo (no EBM)	1.91 ± 0.11	1.09 ± 0.18	0.31 ± 0.01*	0.89 ± 0.03	0.14 ± 0.01	0.98 ± 0.01
	REVISE	1.67 ± 0.11**	**0.74 ± 0.13****	0.31 ± 0.05	4.46 ± 0.23	0.04 ± 0.01	0.61 ± 0.06
	Schut	1.94 ± 0.12	1.35 ± 0.17	**0.25 ± 0.01****	1.92 ± 0.05	**0.46 ± 0.01****	0.98 ± 0.01
	Wachter	1.93 ± 0.12	1.08 ± 0.18	0.32 ± 0.01	**0.82 ± 0.03**	0.14 ± 0.01	0.98 ± 0.01
JEM Ensemble	ECCCo-L1	**0.88 ± 0.05****	1.15 ± 0.20	0.30 ± 0.01**	1.13 ± 0.04	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no CP)	0.89 ± 0.06**	1.15 ± 0.20	0.31 ± 0.01	1.05 ± 0.05	0.14 ± 0.01	0.98 ± 0.01
	ECCCo-L1 (no EBM)	0.90 ± 0.05**	1.15 ± 0.20	0.31 ± 0.01*	1.00 ± 0.04	0.17 ± 0.01	0.98 ± 0.01
	ECCCo	1.29 ± 0.11*	0.94 ± 0.19	0.37 ± 0.01	1.11 ± 0.06	0.13 ± 0.01	0.99 ± 0.01
	ECCCo+	1.08 ± 0.11**	0.85 ± 0.17*	0.36 ± 0.01	2.64 ± 0.17	0.10 ± 0.01	**0.99 ± 0.01***
	ECCCo (no CP)	1.29 ± 0.12	0.94 ± 0.19*	0.39 ± 0.01	1.04 ± 0.06	0.13 ± 0.01	0.99 ± 0.01
	ECCCo (no EBM)	1.40 ± 0.11	1.15 ± 0.20	0.31 ± 0.01*	1.00 ± 0.04	0.17 ± 0.01	0.98 ± 0.01
	REVISE	1.05 ± 0.09**	**0.74 ± 0.13****	0.30 ± 0.04	4.40 ± 0.28	0.05 ± 0.02	0.60 ± 0.06
	Schut	1.42 ± 0.12	1.38 ± 0.20	**0.25 ± 0.01****	2.08 ± 0.05	**0.44 ± 0.01****	0.98 ± 0.01
	Wachter	1.40 ± 0.11	1.13 ± 0.21	0.33 ± 0.01	**0.92 ± 0.04**	0.17 ± 0.01	0.98 ± 0.01
MLP	ECCCo-L1	**1.21 ± 0.03****	1.06 ± 0.11	0.28 ± 0.01*	1.39 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	ECCCo-L1 (no CP)	1.21 ± 0.03**	1.05 ± 0.11	0.29 ± 0.01	1.34 ± 0.14	0.14 ± 0.01	0.97 ± 0.01
	ECCCo-L1 (no EBM)	1.23 ± 0.04**	1.03 ± 0.11	0.29 ± 0.01	1.25 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	ECCCo	3.84 ± 0.08	1.69 ± 0.10	**0.24 ± 0.01****	3.25 ± 0.21	0.14 ± 0.01	**0.98 ± 0.01**
	ECCCo+	3.76 ± 0.08*	1.14 ± 0.08	0.26 ± 0.01**	3.91 ± 0.20	0.10 ± 0.01	0.97 ± 0.01
	ECCCo (no CP)	3.84 ± 0.09	1.69 ± 0.10	0.24 ± 0.01**	3.23 ± 0.21	0.14 ± 0.01	**0.98 ± 0.01**
	ECCCo (no EBM)	3.83 ± 0.07	1.03 ± 0.11	0.29 ± 0.01	1.25 ± 0.15	0.14 ± 0.01	0.97 ± 0.01
	REVISE	3.78 ± 0.06*	**0.63 ± 0.02****	0.28 ± 0.04	4.52 ± 0.28	0.04 ± 0.01	0.69 ± 0.05
	Schut	3.89 ± 0.09	1.20 ± 0.15	0.28 ± 0.02	1.51 ± 0.07	**0.50 ± 0.02****	0.93 ± 0.02
	Wachter	3.85 ± 0.08	1.02 ± 0.11	0.30 ± 0.01	**1.19 ± 0.15**	0.14 ± 0.01	0.97 ± 0.01
MLP Ensemble	ECCCo-L1	**1.21 ± 0.03****	1.04 ± 0.10	0.29 ± 0.01**	1.46 ± 0.22	0.14 ± 0.01	**0.99 ± 0.01**
	ECCCo-L1 (no CP)	1.21 ± 0.02**	1.03 ± 0.10	0.31 ± 0.01*	1.39 ± 0.22	0.14 ± 0.01	**0.99 ± 0.01**
	ECCCo-L1 (no EBM)	1.24 ± 0.03**	1.00 ± 0.10	0.31 ± 0.01*	1.27 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	ECCCo	3.86 ± 0.08**	2.16 ± 0.09	**0.23 ± 0.01****	4.41 ± 0.25	0.13 ± 0.01	**0.99 ± 0.01**
	ECCCo+	3.81 ± 0.06**	1.84 ± 0.07	0.30 ± 0.01*	5.45 ± 0.24	0.10 ± 0.01	0.99 ± 0.01
	ECCCo (no CP)	3.87 ± 0.07**	2.15 ± 0.10	0.23 ± 0.01**	4.37 ± 0.25	0.13 ± 0.01	**0.99 ± 0.01**
	ECCCo (no EBM)	4.10 ± 0.07	1.00 ± 0.10	0.31 ± 0.01*	1.27 ± 0.22	0.14 ± 0.01	0.99 ± 0.01
	REVISE	4.08 ± 0.06	**0.63 ± 0.02****	0.29 ± 0.04	4.54 ± 0.36	0.04 ± 0.01	0.66 ± 0.04
	Schut	4.10 ± 0.10	1.32 ± 0.17	0.30 ± 0.01*	1.69 ± 0.06	**0.56 ± 0.02****	0.98 ± 0.01
	Wachter	4.11 ± 0.07	0.99 ± 0.10	0.32 ± 0.01	**1.20 ± 0.22**	0.14 ± 0.01	0.99 ± 0.01

Table 12: All results for German Credit dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	1.69 ± 0.03**	5.16 ± 0.08	0.59 ± 0.01**	3.06 ± 0.19	0.29 ± 0.01	1.00 ± 0.01
	ECCCo-L1 (no CP)	**1.69 ± 0.03****	5.16 ± 0.07	0.60 ± 0.01	2.98 ± 0.19	0.29 ± 0.01	1.00 ± 0.01
	ECCCo-L1 (no EBM)	1.70 ± 0.04**	5.18 ± 0.08	0.59 ± 0.01**	3.04 ± 0.19	0.30 ± 0.01	**1.00 ± 0.01**
	ECCCo	4.47 ± 0.09*	5.07 ± 0.07*	0.60 ± 0.01	3.65 ± 0.28	0.27 ± 0.01	1.00 ± 0.01
	ECCCo+	4.08 ± 0.07**	**3.82 ± 0.06****	0.22 ± 0.01**	13.04 ± 0.29	0.09 ± 0.01	0.39 ± 0.01
	ECCCo (no CP)	4.48 ± 0.10	5.06 ± 0.07*	0.61 ± 0.01	3.59 ± 0.28	0.27 ± 0.01	1.00 ± 0.01
	ECCCo (no EBM)	4.55 ± 0.09	5.18 ± 0.08	0.59 ± 0.01**	3.04 ± 0.19	0.30 ± 0.01	**1.00 ± 0.01**
	REVISE	4.11 ± 0.09**	3.88 ± 0.05**	**0.21 ± 0.01****	18.00 ± 0.41	0.04 ± 0.01	0.39 ± 0.00
	Schut	4.58 ± 0.08	5.22 ± 0.08	0.44 ± 0.03**	**2.44 ± 0.09****	**0.80 ± 0.01****	0.75 ± 0.03
	Wachter	4.56 ± 0.08	5.17 ± 0.08	0.60 ± 0.01	2.96 ± 0.19	0.30 ± 0.01	**1.00 ± 0.01**
JEM Ensemble	ECCCo-L1	1.61 ± 0.04**	5.07 ± 0.09	0.88 ± 0.01**	3.51 ± 0.18	0.28 ± 0.01	**1.00 ± 0.00**
	ECCCo-L1 (no CP)	**1.60 ± 0.04****	5.08 ± 0.10	0.90 ± 0.01	3.45 ± 0.19	0.28 ± 0.01	**1.00 ± 0.00**
	ECCCo-L1 (no EBM)	1.62 ± 0.04**	5.09 ± 0.10	0.88 ± 0.01**	3.49 ± 0.19	0.29 ± 0.01	**1.00 ± 0.00**
	ECCCo	3.72 ± 0.11**	4.13 ± 0.06**	0.79 ± 0.02**	7.44 ± 0.31	0.17 ± 0.01	0.95 ± 0.02
	ECCCo+	3.62 ± 0.11**	**3.61 ± 0.04****	0.24 ± 0.00**	13.88 ± 0.38	0.08 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	3.70 ± 0.13**	4.13 ± 0.06**	0.80 ± 0.02**	7.43 ± 0.31	0.17 ± 0.01	0.95 ± 0.02
	ECCCo (no EBM)	4.24 ± 0.12	5.09 ± 0.10	0.88 ± 0.01**	3.49 ± 0.19	0.29 ± 0.01	**1.00 ± 0.00**
	REVISE	3.41 ± 0.13**	3.80 ± 0.07**	**0.20 ± 0.00****	17.46 ± 0.51	0.03 ± 0.01	0.39 ± 0.00
	Schut	4.17 ± 0.11	5.05 ± 0.10	0.58 ± 0.04**	**3.38 ± 0.08**	**0.76 ± 0.01****	0.74 ± 0.04
	Wachter	4.23 ± 0.12	5.10 ± 0.10	0.90 ± 0.01	3.42 ± 0.18	0.29 ± 0.01	**1.00 ± 0.00**
MLP	ECCCo-L1	**1.77 ± 0.03****	4.61 ± 0.09	0.76 ± 0.00*	4.98 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.77 ± 0.03**	4.61 ± 0.09	0.77 ± 0.00	4.93 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.77 ± 0.03**	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	**1.00 ± 0.00****
	ECCCo	5.36 ± 0.07	4.63 ± 0.09	0.81 ± 0.00	4.83 ± 0.21	0.24 ± 0.01**	**1.00 ± 0.00****
	ECCCo+	5.10 ± 0.06**	**3.55 ± 0.02****	0.22 ± 0.00**	13.50 ± 0.29	0.02 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	5.35 ± 0.07	4.62 ± 0.09	0.82 ± 0.00	4.78 ± 0.21	0.23 ± 0.01*	**1.00 ± 0.00****
	ECCCo (no EBM)	5.35 ± 0.06	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	**1.00 ± 0.00****
	REVISE	5.18 ± 0.06**	3.89 ± 0.07**	**0.20 ± 0.00****	17.21 ± 0.53	0.02 ± 0.01	0.39 ± 0.00
	Schut	5.40 ± 0.07	4.74 ± 0.09	0.37 ± 0.03**	**3.29 ± 0.06****	**0.72 ± 0.01****	0.56 ± 0.03
	Wachter	5.34 ± 0.06	4.63 ± 0.09	0.77 ± 0.01	4.90 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**1.64 ± 0.03****	4.16 ± 0.05	0.78 ± 0.02*	7.03 ± 0.28	0.23 ± 0.01	0.95 ± 0.02
	ECCCo-L1 (no CP)	1.64 ± 0.03**	4.16 ± 0.05	0.80 ± 0.02	6.95 ± 0.29	0.24 ± 0.01	0.95 ± 0.02
	ECCCo-L1 (no EBM)	1.65 ± 0.03**	4.17 ± 0.05	0.78 ± 0.02*	6.99 ± 0.28	0.24 ± 0.01	0.95 ± 0.02
	ECCCo	4.48 ± 0.06	4.16 ± 0.05	0.82 ± 0.02	6.95 ± 0.29	0.23 ± 0.01	**0.96 ± 0.02**
	ECCCo+	4.62 ± 0.08	**3.43 ± 0.01****	0.22 ± 0.00**	13.11 ± 0.33	0.02 ± 0.00	0.39 ± 0.00
	ECCCo (no CP)	4.49 ± 0.07	4.16 ± 0.05	0.83 ± 0.02	6.89 ± 0.29	0.24 ± 0.01	**0.96 ± 0.02**
	ECCCo (no EBM)	4.47 ± 0.07	4.17 ± 0.05	0.78 ± 0.02*	6.99 ± 0.28	0.24 ± 0.01	0.95 ± 0.02
	REVISE	4.78 ± 0.09	3.85 ± 0.06**	**0.20 ± 0.00****	16.71 ± 0.46	0.03 ± 0.01	0.39 ± 0.00
	Schut	4.84 ± 0.07	4.52 ± 0.06	0.26 ± 0.02**	**3.76 ± 0.06****	**0.71 ± 0.01****	0.44 ± 0.02
	Wachter	4.49 ± 0.06	4.17 ± 0.05	0.80 ± 0.02	6.91 ± 0.29	0.24 ± 0.01	0.95 ± 0.02

Table 13: All results for MNIST dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.10 ± 0.00****	0.35 ± 0.00	4.59 ± 0.01**	23.75 ± 0.97	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.25 ± 0.00**	0.46 ± 0.00	4.50 ± 0.00**	160.43 ± 1.31	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	4.51 ± 0.00**	196.62 ± 3.62	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.25 ± 0.00	**0.29 ± 0.00****	3.82 ± 0.14**	239.71 ± 6.72	0.00 ± 0.00	0.96 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	**0.73 ± 0.15****	**7.01 ± 0.18****	**0.99 ± 0.00****	0.16 ± 0.03
	Wachter	0.25 ± 0.00	0.35 ± 0.00	4.62 ± 0.01	22.90 ± 0.99	0.00 ± 0.00	**1.00 ± 0.00**
JEM Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	2.44 ± 0.01**	35.89 ± 1.47	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.25 ± 0.00**	0.45 ± 0.00	1.62 ± 0.01**	157.21 ± 1.17	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo+	0.24 ± 0.00**	0.30 ± 0.00**	1.58 ± 0.01**	193.39 ± 2.21	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.25 ± 0.00**	**0.29 ± 0.00****	1.75 ± 0.09**	233.72 ± 3.16	0.00 ± 0.00	0.96 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	**0.14 ± 0.04****	**7.82 ± 0.07****	**0.99 ± 0.00****	0.06 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	2.53 ± 0.01	34.89 ± 1.45	0.00 ± 0.00	**1.00 ± 0.00**
LeNet-5	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	8.62 ± 0.09	46.71 ± 1.28	0.00 ± 0.00	0.99 ± 0.01
	ECCCo	0.25 ± 0.00**	0.39 ± 0.00	**0.01 ± 0.03****	121.18 ± 1.95	0.00 ± 0.00**	**1.00 ± 0.00****
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	0.16 ± 0.10**	173.48 ± 1.80	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	**0.30 ± 0.00****	2.08 ± 0.21**	228.22 ± 3.77	0.00 ± 0.00	0.85 ± 0.04
	Schut	0.25 ± 0.00	0.34 ± 0.00	0.06 ± 0.06**	**8.91 ± 0.08****	**0.98 ± 0.00****	0.01 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	8.69 ± 0.06	45.50 ± 1.26	0.00 ± 0.00	0.99 ± 0.01
MLP	ECCCo-L1	**0.10 ± 0.00****	0.35 ± 0.00	1.08 ± 0.00**	41.48 ± 1.47	0.00 ± 0.00	**1.00 ± 0.00**
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	0.86 ± 0.00**	131.37 ± 1.79	0.00 ± 0.00**	**1.00 ± 0.00**
	ECCCo+	0.25 ± 0.00*	0.31 ± 0.00**	0.86 ± 0.00**	175.89 ± 2.75	0.00 ± 0.00	**1.00 ± 0.00**
	REVISE	0.25 ± 0.00	**0.30 ± 0.00****	0.68 ± 0.04**	224.01 ± 3.66	0.00 ± 0.00	0.86 ± 0.02
	Schut	0.25 ± 0.00	0.34 ± 0.00**	**0.04 ± 0.02****	**8.18 ± 0.16****	**0.99 ± 0.00****	0.04 ± 0.02
	Wachter	0.25 ± 0.00	0.34 ± 0.00	1.16 ± 0.00	39.80 ± 1.44	0.00 ± 0.00	**1.00 ± 0.00**
MLP Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	0.61 ± 0.01**	45.56 ± 1.46	0.00 ± 0.00	0.99 ± 0.01
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	0.33 ± 0.00**	116.67 ± 1.95	0.00 ± 0.00**	**1.00 ± 0.00****
	ECCCo+	0.25 ± 0.00**	0.30 ± 0.00**	0.34 ± 0.00**	174.80 ± 3.42	0.00 ± 0.00	0.99 ± 0.01
	REVISE	0.25 ± 0.00	**0.30 ± 0.00****	0.33 ± 0.03**	224.93 ± 3.99	0.00 ± 0.00	0.87 ± 0.04
	Schut	0.25 ± 0.00	0.34 ± 0.00	**0.01 ± 0.01****	**8.38 ± 0.11****	**0.99 ± 0.00****	0.02 ± 0.01
	Wachter	0.25 ± 0.00	0.34 ± 0.00	0.65 ± 0.01	43.98 ± 1.41	0.00 ± 0.00	0.99 ± 0.01

Table 14: All results for Fashion MNIST dataset: sample averages +/- one standard deviation over all counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.10 ± 0.00****	0.40 ± 0.00	4.15 ± 0.10	31.58 ± 1.34	0.00 ± 0.00	0.84 ± 0.02*
	ECCCo	0.25 ± 0.00**	0.43 ± 0.00	4.56 ± 0.02	79.06 ± 1.35	0.00 ± 0.00**	**1.00 ± 0.00****
	ECCCo+	0.25 ± 0.00**	0.38 ± 0.00**	2.98 ± 0.15**	184.42 ± 4.53	0.00 ± 0.00	0.62 ± 0.03
	REVISE	0.25 ± 0.00	**0.37 ± 0.00****	**0.74 ± 0.17****	196.14 ± 4.92	0.00 ± 0.00	0.16 ± 0.04
	Schut	0.25 ± 0.00	0.40 ± 0.00*	0.88 ± 0.13**	**8.21 ± 0.10****	**0.99 ± 0.00****	0.18 ± 0.03
	Wachter	0.25 ± 0.00	0.40 ± 0.00	4.07 ± 0.11	29.91 ± 1.30	0.00 ± 0.00	0.81 ± 0.02
JEM Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.40 ± 0.00	0.79 ± 0.10	40.74 ± 1.07	0.00 ± 0.00	0.26 ± 0.04*
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.68 ± 0.04	112.32 ± 2.27	0.00 ± 0.00*	**0.91 ± 0.02****
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	1.61 ± 0.13	193.58 ± 3.17	0.00 ± 0.00	0.55 ± 0.04**
	REVISE	0.25 ± 0.00	**0.37 ± 0.00****	0.46 ± 0.12*	219.25 ± 4.89	0.00 ± 0.00	0.15 ± 0.03
	Schut	0.25 ± 0.00	0.39 ± 0.00	**0.23 ± 0.06****	**8.93 ± 0.12****	**0.99 ± 0.00****	0.08 ± 0.02
	Wachter	0.25 ± 0.00	0.39 ± 0.00	0.65 ± 0.09	36.91 ± 1.07	0.00 ± 0.00	0.22 ± 0.03
LeNet-5	ECCCo-L1	**0.10 ± 0.00****	0.40 ± 0.00	0.83 ± 0.08	30.24 ± 0.90	0.00 ± 0.00	0.24 ± 0.02*
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.60 ± 0.07	94.33 ± 1.90	0.00 ± 0.00**	**0.93 ± 0.02****
	ECCCo+	0.25 ± 0.00**	0.38 ± 0.00**	1.79 ± 0.16	173.75 ± 2.75	0.00 ± 0.00	0.59 ± 0.05**
	REVISE	0.25 ± 0.00	**0.38 ± 0.00****	0.32 ± 0.07**	189.62 ± 3.95	0.00 ± 0.00	0.10 ± 0.02
	Schut	0.25 ± 0.00	0.39 ± 0.00*	**0.19 ± 0.07****	**8.59 ± 0.14****	**0.97 ± 0.00****	0.06 ± 0.02
	Wachter	0.25 ± 0.00	0.40 ± 0.00	0.70 ± 0.05	27.66 ± 0.88	0.00 ± 0.00	0.20 ± 0.01
MLP	ECCCo-L1	**0.10 ± 0.00****	0.40 ± 0.00	1.68 ± 0.10	42.02 ± 1.09	0.00 ± 0.00	0.47 ± 0.03**
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	2.89 ± 0.02	112.70 ± 1.67	0.00 ± 0.00**	**0.99 ± 0.01****
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	2.21 ± 0.08	189.26 ± 3.80	0.00 ± 0.00	0.71 ± 0.03**
	REVISE	0.25 ± 0.00	**0.37 ± 0.00****	0.45 ± 0.09**	202.31 ± 5.63	0.00 ± 0.00	0.15 ± 0.03
	Schut	0.25 ± 0.00	0.40 ± 0.00*	**0.29 ± 0.09****	**8.61 ± 0.12****	**0.99 ± 0.00****	0.09 ± 0.03
	Wachter	0.25 ± 0.00	0.40 ± 0.00	1.52 ± 0.11	38.73 ± 1.03	0.00 ± 0.00	0.42 ± 0.03
MLP Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.39 ± 0.00	0.66 ± 0.12	36.10 ± 1.54	0.00 ± 0.00	0.24 ± 0.04*
	ECCCo	0.24 ± 0.00**	0.41 ± 0.00	1.91 ± 0.04	116.23 ± 2.62	0.00 ± 0.00**	**0.92 ± 0.02****
	ECCCo+	0.25 ± 0.00**	**0.37 ± 0.00****	1.61 ± 0.08	188.33 ± 3.71	0.00 ± 0.00	0.71 ± 0.04**
	REVISE	0.25 ± 0.00	0.37 ± 0.00**	0.37 ± 0.05**	198.52 ± 6.17	0.00 ± 0.00	0.14 ± 0.03
	Schut	0.25 ± 0.00	0.39 ± 0.00	**0.11 ± 0.04****	**8.81 ± 0.11****	**0.99 ± 0.00****	0.05 ± 0.02
	Wachter	0.25 ± 0.00	0.39 ± 0.00	0.52 ± 0.04	31.82 ± 1.47	0.00 ± 0.00	0.19 ± 0.01

Table 15: All results for Linearly Separable dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.06 ± 0.01**	**0.16 ± 0.02****	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.06 ± 0.01****	0.16 ± 0.02**	0.00 ± 0.00	0.93 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.09 ± 0.01**	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	**0.00 ± 0.00**	1.00 ± 0.00
	ECCCo (no CP)	0.19 ± 0.01**	0.19 ± 0.01**	0.00 ± 0.00	0.85 ± 0.01**	**0.00 ± 0.00**	1.00 ± 0.00
	ECCCo (no EBM)	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.24 ± 0.02	0.17 ± 0.00**	0.00 ± 0.00	**0.42 ± 0.02****	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.29 ± 0.01	0.29 ± 0.00	0.00 ± 0.00	1.40 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.24 ± 0.01	0.24 ± 0.01	0.00 ± 0.00	0.88 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	**0.02 ± 0.00****	0.15 ± 0.01	**0.00 ± 0.00**	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.02 ± 0.00**	0.15 ± 0.01	**0.00 ± 0.00**	0.98 ± 0.02*	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.05 ± 0.00**	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.12 ± 0.01**	0.07 ± 0.00**	**0.00 ± 0.00**	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.12 ± 0.01**	**0.07 ± 0.00****	**0.00 ± 0.00**	0.97 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.15 ± 0.01	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.01	0.17 ± 0.00	0.45 ± 0.08	**0.42 ± 0.02****	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.20 ± 0.01	0.27 ± 0.00	**0.00 ± 0.00**	1.38 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Wachter	0.14 ± 0.01	0.08 ± 0.00	**0.00 ± 0.00**	1.01 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.12 ± 0.00**	**0.06 ± 0.00****	**0.00 ± 0.00**	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.12 ± 0.00****	0.06 ± 0.00**	**0.00 ± 0.00**	0.90 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.46 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.50 ± 0.00	0.18 ± 0.00	0.97 ± 0.02	**0.41 ± 0.02****	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.48 ± 0.00	0.31 ± 0.01	**0.00 ± 0.00**	0.70 ± 0.02	**0.46 ± 0.01****	1.00 ± 0.00
	Wachter	0.46 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**0.12 ± 0.00****	0.06 ± 0.00**	**0.00 ± 0.00**	0.91 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.12 ± 0.00**	**0.06 ± 0.00****	**0.00 ± 0.00**	0.91 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.16 ± 0.00**	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.17 ± 0.00**	0.31 ± 0.00	**0.00 ± 0.00**	1.44 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.45 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.50 ± 0.00	0.18 ± 0.00	0.98 ± 0.01	**0.42 ± 0.02****	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.48 ± 0.00	0.32 ± 0.01	**0.00 ± 0.00**	0.73 ± 0.01	**0.42 ± 0.02****	1.00 ± 0.00
	Wachter	0.45 ± 0.00	0.07 ± 0.00	**0.00 ± 0.00**	0.65 ± 0.01	0.00 ± 0.00	1.00 ± 0.00

Table 16: All results for Circles dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.24 ± 0.01**	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.24 ± 0.02**	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.25 ± 0.02**	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.61 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.60 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.62 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.62 ± 0.04	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	**0.12 ± 0.02****	0.43 ± 0.00	0.00 ± 0.00	1.27 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.49 ± 0.02**	**0.29 ± 0.00****	0.00 ± 0.00	**0.48 ± 0.01****	**0.41 ± 0.01****	1.00 ± 0.00
	Wachter	0.63 ± 0.05	0.33 ± 0.01	0.00 ± 0.00	0.50 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.19 ± 0.01**	0.30 ± 0.01	0.01 ± 0.01**	**0.56 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.19 ± 0.01**	0.30 ± 0.01	0.02 ± 0.01	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.20 ± 0.01**	0.30 ± 0.01	0.00 ± 0.00**	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.45 ± 0.04*	0.35 ± 0.00	**0.00 ± 0.00****	0.69 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.46 ± 0.03*	0.35 ± 0.00	0.00 ± 0.00**	0.69 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.48 ± 0.03	0.30 ± 0.01	0.00 ± 0.00**	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	**0.05 ± 0.00****	0.43 ± 0.00	**0.00 ± 0.00****	1.26 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.45 ± 0.02*	**0.26 ± 0.00****	**0.00 ± 0.00****	0.58 ± 0.01	**0.38 ± 0.01****	1.00 ± 0.00
	Wachter	0.49 ± 0.04	0.30 ± 0.01	0.02 ± 0.01	0.56 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	0.29 ± 0.01**	0.28 ± 0.01	0.01 ± 0.00**	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.29 ± 0.01**	0.28 ± 0.00	0.02 ± 0.01**	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.71 ± 0.04*	0.56 ± 0.00	**0.00 ± 0.00****	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.71 ± 0.04*	0.56 ± 0.00	**0.00 ± 0.00****	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.01 ± 0.01**	0.57 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	**0.01 ± 0.00****	0.43 ± 0.00	**0.00 ± 0.00****	1.27 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.78 ± 0.02	**0.25 ± 0.00****	**0.00 ± 0.00****	**0.55 ± 0.01****	**0.39 ± 0.01****	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.29 ± 0.00	0.04 ± 0.01	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.29 ± 0.01**	0.27 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.29 ± 0.01**	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.01**	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.71 ± 0.04*	0.57 ± 0.00	0.00 ± 0.00	1.38 ± 0.01	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	**0.01 ± 0.00****	0.43 ± 0.00	0.00 ± 0.00	1.26 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.78 ± 0.02	**0.25 ± 0.00****	0.00 ± 0.00	**0.58 ± 0.01**	**0.38 ± 0.02****	1.00 ± 0.00
	Wachter	0.77 ± 0.02	0.28 ± 0.01	0.00 ± 0.00	0.58 ± 0.02	0.00 ± 0.00	1.00 ± 0.00

Table 17: All results for Moons dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.29 ± 0.02**	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.29 ± 0.02****	0.42 ± 0.04	0.15 ± 0.03	**1.02 ± 0.03**	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.30 ± 0.02**	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.55 ± 0.06**	0.15 ± 0.01**	**0.01 ± 0.01****	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.57 ± 0.05**	**0.15 ± 0.01****	0.01 ± 0.01**	1.28 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.69 ± 0.04	0.42 ± 0.04	0.13 ± 0.03	1.03 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.54 ± 0.05**	0.27 ± 0.01**	0.14 ± 0.05	1.42 ± 0.13	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.72 ± 0.05	0.46 ± 0.04	0.19 ± 0.02	1.12 ± 0.03	**0.06 ± 0.02****	1.00 ± 0.00
	Wachter	0.69 ± 0.05	0.42 ± 0.04	0.14 ± 0.03	1.02 ± 0.03	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.26 ± 0.01**	0.18 ± 0.01	0.02 ± 0.01	1.04 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.26 ± 0.02****	0.18 ± 0.01	0.04 ± 0.01	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.26 ± 0.01**	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.47 ± 0.03**	0.13 ± 0.00**	**0.00 ± 0.00****	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	0.46 ± 0.04**	**0.13 ± 0.00****	**0.00 ± 0.00****	1.50 ± 0.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	0.67 ± 0.02	0.18 ± 0.01	0.00 ± 0.00*	1.03 ± 0.02	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.66 ± 0.03	0.27 ± 0.01	0.44 ± 0.05	1.37 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.51 ± 0.03**	0.24 ± 0.01	**0.00 ± 0.00****	1.81 ± 0.03	**0.05 ± 0.01****	1.00 ± 0.00
	Wachter	0.67 ± 0.03	0.18 ± 0.01	0.01 ± 0.01	**1.03 ± 0.02**	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	**0.45 ± 0.02****	0.34 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.45 ± 0.03**	0.35 ± 0.02	0.31 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	1.33 ± 0.06**	0.59 ± 0.03	**0.00 ± 0.00****	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.32 ± 0.09*	0.59 ± 0.03	**0.00 ± 0.00****	2.84 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.49 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.50 ± 0.06	**0.26 ± 0.01****	0.30 ± 0.03	1.20 ± 0.07**	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.59 ± 0.06	0.38 ± 0.03	**0.00 ± 0.00****	**0.78 ± 0.02****	**0.21 ± 0.03****	1.00 ± 0.00
	Wachter	1.50 ± 0.04	0.35 ± 0.02	0.30 ± 0.02	1.48 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	0.44 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.44 ± 0.02****	0.28 ± 0.02	0.24 ± 0.03	1.43 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.46 ± 0.02**	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	1.24 ± 0.07**	0.56 ± 0.02	**0.00 ± 0.00****	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.23 ± 0.09**	0.56 ± 0.02	**0.00 ± 0.00****	2.81 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.50 ± 0.03	0.28 ± 0.02	0.22 ± 0.02*	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00
	REVISE	1.59 ± 0.06	**0.25 ± 0.01****	0.32 ± 0.03	1.23 ± 0.07**	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.61 ± 0.06	0.28 ± 0.01	0.03 ± 0.01**	**0.75 ± 0.02****	**0.26 ± 0.03****	1.00 ± 0.00
	Wachter	1.51 ± 0.04	0.29 ± 0.02	0.24 ± 0.02	1.42 ± 0.08	0.00 ± 0.00	1.00 ± 0.00

Table 18: All results for California Housing dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	0.98 ± 0.05**	0.99 ± 0.09	0.25 ± 0.02	1.20 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.97 ± 0.05****	0.99 ± 0.09	0.26 ± 0.02	1.15 ± 0.06	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.00 ± 0.05**	1.02 ± 0.09	0.25 ± 0.01	1.10 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	1.93 ± 0.13*	0.77 ± 0.07**	**0.18 ± 0.00****	2.05 ± 0.16	0.01 ± 0.00	1.00 ± 0.00
	ECCCo+	1.83 ± 0.12**	0.69 ± 0.04**	0.19 ± 0.00**	3.27 ± 0.18	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.95 ± 0.15*	0.77 ± 0.07**	0.18 ± 0.00**	2.03 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	2.11 ± 0.11	1.02 ± 0.09	0.25 ± 0.01	1.10 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	REVISE	1.87 ± 0.16*	**0.58 ± 0.04****	0.31 ± 0.06	5.50 ± 0.39	0.00 ± 0.00	1.00 ± 0.00
	Schut	2.03 ± 0.12	1.10 ± 0.08	0.21 ± 0.01**	2.81 ± 0.08	**0.29 ± 0.01****	1.00 ± 0.00
	Wachter	2.12 ± 0.11	1.02 ± 0.09	0.26 ± 0.02	**1.05 ± 0.07**	0.01 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	0.89 ± 0.04**	0.98 ± 0.08	0.13 ± 0.00**	1.37 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**0.89 ± 0.05****	0.98 ± 0.08	0.14 ± 0.00*	1.33 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.92 ± 0.05**	0.99 ± 0.08	0.14 ± 0.00*	1.27 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	1.40 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	2.27 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	1.28 ± 0.08**	0.60 ± 0.04**	0.11 ± 0.00**	3.39 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	1.39 ± 0.08**	0.69 ± 0.05**	0.11 ± 0.00**	2.23 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	1.70 ± 0.09	0.99 ± 0.08	0.14 ± 0.00*	1.27 ± 0.07	0.01 ± 0.00	1.00 ± 0.00
	REVISE	1.39 ± 0.15**	**0.59 ± 0.04****	0.25 ± 0.07	5.37 ± 0.52	0.00 ± 0.00	1.00 ± 0.00
	Schut	1.59 ± 0.10*	1.10 ± 0.06	**0.09 ± 0.00****	3.07 ± 0.10	**0.31 ± 0.02****	1.00 ± 0.00
	Wachter	1.71 ± 0.09	0.99 ± 0.08	0.14 ± 0.00	**1.23 ± 0.07**	0.01 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	**1.22 ± 0.04****	1.18 ± 0.09	0.17 ± 0.01*	1.37 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.23 ± 0.04**	1.18 ± 0.09	0.17 ± 0.01	1.34 ± 0.10	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.25 ± 0.04**	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	3.81 ± 0.10*	2.31 ± 0.09	**0.14 ± 0.01****	6.20 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	3.93 ± 0.10	1.30 ± 0.08	0.16 ± 0.01*	4.65 ± 0.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	3.80 ± 0.11*	2.31 ± 0.09	0.14 ± 0.01**	6.17 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	3.98 ± 0.08	1.19 ± 0.09	0.17 ± 0.01	1.26 ± 0.09	0.00 ± 0.00	1.00 ± 0.00
	REVISE	3.88 ± 0.08*	**0.58 ± 0.03****	0.20 ± 0.04	5.38 ± 0.36	0.00 ± 0.00	1.00 ± 0.00
	Schut	3.99 ± 0.08	1.15 ± 0.09	0.15 ± 0.01**	1.58 ± 0.09	**0.49 ± 0.02****	1.00 ± 0.00
	Wachter	3.98 ± 0.07	1.19 ± 0.09	0.17 ± 0.01	**1.23 ± 0.09**	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	1.22 ± 0.04**	1.13 ± 0.12	0.13 ± 0.01**	1.20 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no CP)	**1.22 ± 0.03****	1.13 ± 0.12	0.15 ± 0.01	1.17 ± 0.07	0.00 ± 0.00	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.24 ± 0.04**	1.12 ± 0.12	0.14 ± 0.01**	1.07 ± 0.06	0.01 ± 0.00	1.00 ± 0.00
	ECCCo	3.69 ± 0.08**	1.94 ± 0.13	**0.09 ± 0.01****	4.76 ± 0.13	0.01 ± 0.00	1.00 ± 0.00
	ECCCo+	3.88 ± 0.07**	1.20 ± 0.09	0.15 ± 0.02	4.23 ± 0.14	0.00 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	3.70 ± 0.08**	1.94 ± 0.13	0.10 ± 0.01**	4.76 ± 0.13	0.01 ± 0.00	1.00 ± 0.00
	ECCCo (no EBM)	4.03 ± 0.07	1.12 ± 0.12	0.14 ± 0.01**	1.07 ± 0.06	0.01 ± 0.00	1.00 ± 0.00
	REVISE	3.96 ± 0.07*	**0.58 ± 0.03****	0.17 ± 0.03	5.45 ± 0.37	0.00 ± 0.00	1.00 ± 0.00
	Schut	4.00 ± 0.06	1.15 ± 0.12	0.10 ± 0.01**	1.55 ± 0.08	**0.58 ± 0.02****	1.00 ± 0.00
	Wachter	4.04 ± 0.07	1.13 ± 0.12	0.16 ± 0.01	**1.04 ± 0.06**	0.01 ± 0.00	1.00 ± 0.00

Table 19: All results for GMSC dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.92 ± 0.04****	0.94 ± 0.09	0.30 ± 0.01**	1.02 ± 0.04	0.13 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.93 ± 0.04**	0.93 ± 0.09	0.31 ± 0.01*	0.94 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.95 ± 0.04**	0.93 ± 0.10	0.31 ± 0.01*	0.88 ± 0.03	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	1.78 ± 0.10	0.86 ± 0.09	0.35 ± 0.01	1.13 ± 0.05	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	1.65 ± 0.10**	0.73 ± 0.05**	0.35 ± 0.01	2.58 ± 0.11	0.09 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	1.79 ± 0.10	0.85 ± 0.09	0.36 ± 0.01	1.06 ± 0.05	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	1.84 ± 0.10	0.93 ± 0.10	0.31 ± 0.01*	0.88 ± 0.03	0.14 ± 0.01	1.00 ± 0.00
	REVISE	1.66 ± 0.13*	**0.63 ± 0.04****	0.34 ± 0.08	4.42 ± 0.30	0.03 ± 0.01	1.00 ± 0.00
	Schut	1.87 ± 0.09	1.19 ± 0.07	**0.26 ± 0.01****	1.90 ± 0.05	**0.46 ± 0.01****	1.00 ± 0.00
	Wachter	1.85 ± 0.11	0.91 ± 0.10	0.33 ± 0.01	**0.80 ± 0.03**	0.14 ± 0.01	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	**0.85 ± 0.04****	0.98 ± 0.10	0.31 ± 0.01**	1.10 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	0.86 ± 0.05**	0.97 ± 0.10	0.32 ± 0.01*	1.03 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	0.87 ± 0.04**	0.97 ± 0.10	0.32 ± 0.01**	0.98 ± 0.03	0.17 ± 0.01	1.00 ± 0.00
	ECCCo	1.20 ± 0.06*	0.78 ± 0.07**	0.38 ± 0.01	1.10 ± 0.06	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	1.01 ± 0.07**	0.70 ± 0.07**	0.37 ± 0.01	2.63 ± 0.16	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	1.21 ± 0.07*	0.77 ± 0.07**	0.39 ± 0.01	1.03 ± 0.06	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	1.31 ± 0.07	0.97 ± 0.10	0.32 ± 0.01**	0.98 ± 0.03	0.17 ± 0.01	1.00 ± 0.00
	REVISE	1.01 ± 0.07**	**0.63 ± 0.04****	0.33 ± 0.07	4.35 ± 0.36	0.04 ± 0.01	1.00 ± 0.00
	Schut	1.34 ± 0.07	1.21 ± 0.10	**0.26 ± 0.01****	2.07 ± 0.05	**0.43 ± 0.01****	1.00 ± 0.00
	Wachter	1.31 ± 0.08	0.95 ± 0.10	0.33 ± 0.01	**0.90 ± 0.03**	0.16 ± 0.01	1.00 ± 0.00
MLP	ECCCo-L1	**1.20 ± 0.03****	0.98 ± 0.08	0.29 ± 0.01*	1.24 ± 0.06	0.15 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.20 ± 0.03**	0.97 ± 0.08	0.30 ± 0.01	1.18 ± 0.06	0.15 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.21 ± 0.04**	0.95 ± 0.08	0.29 ± 0.01	1.09 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	3.80 ± 0.07	1.63 ± 0.07	**0.24 ± 0.01****	3.13 ± 0.13	0.14 ± 0.01	1.00 ± 0.00
	ECCCo+	3.74 ± 0.08	1.10 ± 0.05	0.27 ± 0.01**	3.77 ± 0.11	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.81 ± 0.08	1.62 ± 0.08	0.24 ± 0.01**	3.10 ± 0.13	0.14 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	3.79 ± 0.07	0.95 ± 0.08	0.29 ± 0.01	1.09 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	REVISE	3.81 ± 0.08	**0.63 ± 0.03****	0.32 ± 0.05	4.34 ± 0.34	0.04 ± 0.01	1.00 ± 0.00
	Schut	3.83 ± 0.07	1.07 ± 0.07	0.30 ± 0.02	1.51 ± 0.06	**0.50 ± 0.02****	1.00 ± 0.00
	Wachter	3.82 ± 0.07	0.94 ± 0.08	0.30 ± 0.01	**1.03 ± 0.05**	0.15 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**1.21 ± 0.03****	1.01 ± 0.08	0.30 ± 0.01**	1.30 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.21 ± 0.02**	0.99 ± 0.08	0.31 ± 0.01*	1.23 ± 0.05	0.14 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.23 ± 0.03**	0.97 ± 0.08	0.31 ± 0.01*	1.11 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	ECCCo	3.84 ± 0.07**	2.13 ± 0.08	**0.23 ± 0.01****	4.24 ± 0.08	0.13 ± 0.01	1.00 ± 0.00
	ECCCo+	3.79 ± 0.05**	1.81 ± 0.05	0.30 ± 0.01*	5.29 ± 0.11	0.10 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.85 ± 0.07**	2.13 ± 0.08	0.23 ± 0.01**	4.20 ± 0.08	0.13 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.08 ± 0.06	0.97 ± 0.08	0.31 ± 0.01*	1.11 ± 0.04	0.14 ± 0.01	1.00 ± 0.00
	REVISE	4.09 ± 0.07	**0.63 ± 0.02****	0.33 ± 0.06	4.45 ± 0.44	0.04 ± 0.01	1.00 ± 0.00
	Schut	4.04 ± 0.08	1.21 ± 0.08	0.30 ± 0.01*	1.68 ± 0.06	**0.56 ± 0.02****	1.00 ± 0.00
	Wachter	4.10 ± 0.07	0.95 ± 0.08	0.32 ± 0.01	**1.04 ± 0.04**	0.14 ± 0.01	1.00 ± 0.00

Table 20: All results for German Credit dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	1.69 ± 0.03**	5.16 ± 0.07	0.59 ± 0.00**	3.04 ± 0.20	0.29 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	**1.69 ± 0.03****	5.16 ± 0.07	0.60 ± 0.00	2.96 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.70 ± 0.04**	5.18 ± 0.08	0.59 ± 0.00**	3.02 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
	ECCCo	4.48 ± 0.09*	5.07 ± 0.07*	0.60 ± 0.01	3.63 ± 0.28	0.27 ± 0.01	1.00 ± 0.00
	ECCCo+	4.85 ± 0.10	3.76 ± 0.10**	0.55 ± 0.01**	12.77 ± 0.40	0.08 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	4.48 ± 0.10	5.07 ± 0.07*	0.62 ± 0.01	3.58 ± 0.28	0.27 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.55 ± 0.09	5.18 ± 0.08	0.59 ± 0.00**	3.02 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
	REVISE	4.73 ± 0.16	**3.46 ± 0.07****	**0.53 ± 0.01****	19.21 ± 0.73	0.03 ± 0.01	1.00 ± 0.00
	Schut	4.67 ± 0.10	5.14 ± 0.08	0.59 ± 0.01*	**2.53 ± 0.09****	**0.79 ± 0.01****	1.00 ± 0.00
	Wachter	4.57 ± 0.08	5.17 ± 0.08	0.60 ± 0.00	2.95 ± 0.19	0.30 ± 0.01	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	1.61 ± 0.04**	5.07 ± 0.10	0.88 ± 0.01**	3.51 ± 0.19	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	**1.60 ± 0.04****	5.08 ± 0.10	0.90 ± 0.01	3.44 ± 0.19	0.28 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.62 ± 0.04**	5.09 ± 0.10	0.88 ± 0.01**	3.48 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	ECCCo	3.80 ± 0.12**	4.13 ± 0.06**	0.83 ± 0.01**	7.28 ± 0.30	0.17 ± 0.01	1.00 ± 0.00
	ECCCo+	5.80 ± 0.12	3.55 ± 0.09**	0.60 ± 0.01**	12.12 ± 0.44	0.07 ± 0.01	1.00 ± 0.00
	ECCCo (no CP)	3.77 ± 0.14**	4.13 ± 0.06**	0.84 ± 0.01**	7.26 ± 0.30	0.17 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.24 ± 0.12	5.09 ± 0.10	0.88 ± 0.01**	3.48 ± 0.19	0.29 ± 0.01	1.00 ± 0.00
	REVISE	5.33 ± 0.16	**3.45 ± 0.07****	**0.52 ± 0.01****	17.92 ± 0.89	0.03 ± 0.01	1.00 ± 0.00
	Schut	4.59 ± 0.15	4.90 ± 0.13*	0.78 ± 0.02**	3.46 ± 0.10	**0.76 ± 0.01****	1.00 ± 0.00
	Wachter	4.23 ± 0.12	5.10 ± 0.10	0.90 ± 0.01	**3.42 ± 0.19**	0.29 ± 0.01	1.00 ± 0.00
MLP	ECCCo-L1	**1.77 ± 0.03****	4.61 ± 0.09	0.76 ± 0.00*	4.98 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.77 ± 0.03**	4.61 ± 0.09	0.77 ± 0.00	4.93 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.77 ± 0.03**	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	ECCCo	5.36 ± 0.07	4.63 ± 0.09	0.81 ± 0.00	4.83 ± 0.21	0.24 ± 0.01**	1.00 ± 0.00
	ECCCo+	6.51 ± 0.05	**3.20 ± 0.04****	0.56 ± 0.00**	10.93 ± 0.48	0.04 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	5.35 ± 0.07	4.62 ± 0.09	0.82 ± 0.00	4.78 ± 0.21	0.23 ± 0.01*	1.00 ± 0.00
	ECCCo (no EBM)	5.35 ± 0.06	4.63 ± 0.09	0.76 ± 0.00*	4.95 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
	REVISE	6.53 ± 0.05	3.48 ± 0.07**	**0.52 ± 0.01****	16.89 ± 0.68	0.03 ± 0.01	1.00 ± 0.00
	Schut	6.10 ± 0.13	4.40 ± 0.10**	0.66 ± 0.02**	**2.55 ± 0.10****	**0.78 ± 0.01****	1.00 ± 0.00
	Wachter	5.35 ± 0.06	4.63 ± 0.09	0.77 ± 0.00	4.90 ± 0.21	0.23 ± 0.01	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**1.65 ± 0.03****	4.12 ± 0.05	0.82 ± 0.01**	6.77 ± 0.30	0.24 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no CP)	1.65 ± 0.03**	4.12 ± 0.05	0.84 ± 0.01	6.67 ± 0.31	0.24 ± 0.01	1.00 ± 0.00
	ECCCo-L1 (no EBM)	1.66 ± 0.03**	4.12 ± 0.05	0.82 ± 0.01**	6.72 ± 0.30	0.25 ± 0.01	1.00 ± 0.00
	ECCCo	4.52 ± 0.06	4.13 ± 0.05	0.85 ± 0.01	6.73 ± 0.30	0.23 ± 0.01	1.00 ± 0.00
	ECCCo+	6.29 ± 0.07	**3.17 ± 0.02****	0.57 ± 0.00**	9.64 ± 0.53	0.04 ± 0.00	1.00 ± 0.00
	ECCCo (no CP)	4.52 ± 0.08	4.13 ± 0.05	0.86 ± 0.01	6.67 ± 0.30	0.24 ± 0.01	1.00 ± 0.00
	ECCCo (no EBM)	4.52 ± 0.07	4.12 ± 0.05	0.82 ± 0.01**	6.72 ± 0.30	0.25 ± 0.01	1.00 ± 0.00
	REVISE	6.37 ± 0.06	3.47 ± 0.06**	**0.52 ± 0.01****	15.78 ± 0.82	0.03 ± 0.01	1.00 ± 0.00
	Schut	6.24 ± 0.21	4.10 ± 0.08	0.59 ± 0.02**	**2.90 ± 0.10****	**0.81 ± 0.01****	1.00 ± 0.00
	Wachter	4.54 ± 0.07	4.12 ± 0.05	0.84 ± 0.01	6.63 ± 0.30	0.25 ± 0.01	1.00 ± 0.00

Table 21: All results for MNIST dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.10 ± 0.00****	0.35 ± 0.00	4.59 ± 0.01**	23.75 ± 0.97	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.46 ± 0.00	4.50 ± 0.00**	160.43 ± 1.31	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	4.51 ± 0.00**	196.62 ± 3.62	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	**0.29 ± 0.00****	**3.90 ± 0.14****	240.16 ± 6.54	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.32 ± 0.01**	4.61 ± 0.05	**6.02 ± 0.60****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.35 ± 0.00	4.62 ± 0.01	22.90 ± 0.99	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	2.44 ± 0.01**	35.89 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.45 ± 0.00	1.62 ± 0.01**	157.21 ± 1.17	0.00 ± 0.00	1.00 ± 0.00
	ECCCo+	0.24 ± 0.00**	0.30 ± 0.00**	**1.58 ± 0.01****	193.39 ± 2.21	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	**0.29 ± 0.00****	1.78 ± 0.09**	234.28 ± 3.19	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.31 ± 0.01**	2.46 ± 0.11	**7.20 ± 1.04****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	2.53 ± 0.01	34.89 ± 1.45	0.00 ± 0.00	1.00 ± 0.00
LeNet-5	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	8.67 ± 0.04*	46.52 ± 1.16	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.39 ± 0.00	**0.01 ± 0.03****	121.18 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.31 ± 0.00**	0.17 ± 0.10**	173.43 ± 1.78	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	0.30 ± 0.00**	2.25 ± 0.20**	230.65 ± 3.73	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	**0.29 ± 0.02***	4.26 ± 3.87*	**7.28 ± 1.04****	**0.98 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.33 ± 0.00	8.74 ± 0.00	45.32 ± 1.14	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	**0.10 ± 0.00****	0.35 ± 0.00	1.08 ± 0.00**	41.48 ± 1.47	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	0.86 ± 0.00**	131.37 ± 1.79	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00*	0.31 ± 0.00**	0.86 ± 0.00**	175.89 ± 2.75	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	**0.30 ± 0.00****	**0.74 ± 0.04****	224.80 ± 4.65	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.30 ± 0.01**	1.12 ± 0.09	**7.88 ± 0.67****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	1.16 ± 0.00	39.80 ± 1.44	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.34 ± 0.00	0.61 ± 0.01**	45.21 ± 1.45	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	**0.33 ± 0.00****	116.67 ± 1.95	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.30 ± 0.00**	0.34 ± 0.00**	174.29 ± 3.50	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.00**	0.36 ± 0.03**	226.43 ± 4.59	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	**0.29 ± 0.02****	0.68 ± 0.21	**7.83 ± 1.24****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.34 ± 0.00	0.66 ± 0.01	43.42 ± 1.44	0.00 ± 0.00	1.00 ± 0.00

Table 22: All results for Fashion MNIST dataset: sample averages +/- one standard deviation over all valid counterfactuals. Best outcomes are highlighted in bold. Asterisks indicate that the given value is more than one (*) or two (**) standard deviations away from the baseline (*Wachter*).

Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Cost ↓	Redundancy ↑	Validity ↑
JEM	ECCCo-L1	**0.10 ± 0.00****	0.39 ± 0.01	4.95 ± 0.01**	29.93 ± 1.60	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.43 ± 0.00	4.57 ± 0.01**	78.99 ± 1.28	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.37 ± 0.00**	4.79 ± 0.04**	175.11 ± 4.20	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00*	**0.30 ± 0.02****	**3.82 ± 0.49****	208.22 ± 16.18	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.34 ± 0.01**	4.78 ± 0.04**	**8.29 ± 0.23****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.39 ± 0.01	5.01 ± 0.02	28.33 ± 1.56	0.00 ± 0.00	1.00 ± 0.00
JEM Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.37 ± 0.01	2.98 ± 0.04	37.85 ± 3.04	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00**	0.42 ± 0.00	2.95 ± 0.03	109.67 ± 2.72	0.00 ± 0.00*	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.35 ± 0.00**	2.94 ± 0.02*	183.89 ± 5.64	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00*	0.31 ± 0.01**	**2.04 ± 0.34****	241.12 ± 12.81	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00*	**0.30 ± 0.01****	2.90 ± 0.04*	**8.44 ± 0.36****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.37 ± 0.01	2.97 ± 0.05	35.32 ± 4.17	0.00 ± 0.00	1.00 ± 0.00
LeNet-5	ECCCo-L1	**0.10 ± 0.00****	0.37 ± 0.01	3.39 ± 0.02*	23.45 ± 2.28	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.25 ± 0.00*	0.41 ± 0.00	2.80 ± 0.02**	92.19 ± 2.20	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.36 ± 0.01	3.04 ± 0.04**	167.29 ± 5.17	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00**	**0.29 ± 0.02****	**2.43 ± 0.28****	201.66 ± 17.81	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	0.30 ± 0.02**	3.26 ± 0.14*	**7.82 ± 0.36****	**0.98 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.36 ± 0.01	3.42 ± 0.02	20.30 ± 1.82	0.00 ± 0.00	1.00 ± 0.00
MLP	ECCCo-L1	**0.10 ± 0.00****	0.37 ± 0.01	3.56 ± 0.02**	34.94 ± 2.24	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.42 ± 0.00	2.91 ± 0.01**	112.25 ± 1.61	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00**	0.36 ± 0.00	3.13 ± 0.02**	188.22 ± 6.05	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.30 ± 0.01**	**2.40 ± 0.38****	224.88 ± 16.50	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	**0.28 ± 0.02****	3.30 ± 0.09**	**8.27 ± 0.34****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.36 ± 0.01	3.63 ± 0.03	31.09 ± 2.89	0.00 ± 0.00	1.00 ± 0.00
MLP Ensemble	ECCCo-L1	**0.10 ± 0.00****	0.33 ± 0.02	2.71 ± 0.04	27.86 ± 1.82	0.00 ± 0.00	1.00 ± 0.00
	ECCCo	0.24 ± 0.00**	0.40 ± 0.00	2.07 ± 0.02**	112.98 ± 3.11	0.00 ± 0.00**	1.00 ± 0.00
	ECCCo+	0.25 ± 0.00	0.35 ± 0.00	2.26 ± 0.02**	181.76 ± 4.91	0.00 ± 0.00	1.00 ± 0.00
	REVISE	0.25 ± 0.00	0.28 ± 0.02**	**1.81 ± 0.14****	207.62 ± 16.42	0.00 ± 0.00	1.00 ± 0.00
	Schut	0.25 ± 0.00	**0.27 ± 0.02***	2.45 ± 0.15*	**8.08 ± 0.43****	**0.99 ± 0.00****	1.00 ± 0.00
	Wachter	0.25 ± 0.00	0.32 ± 0.01	2.71 ± 0.04	23.84 ± 1.88	0.00 ± 0.00	1.00 ± 0.00

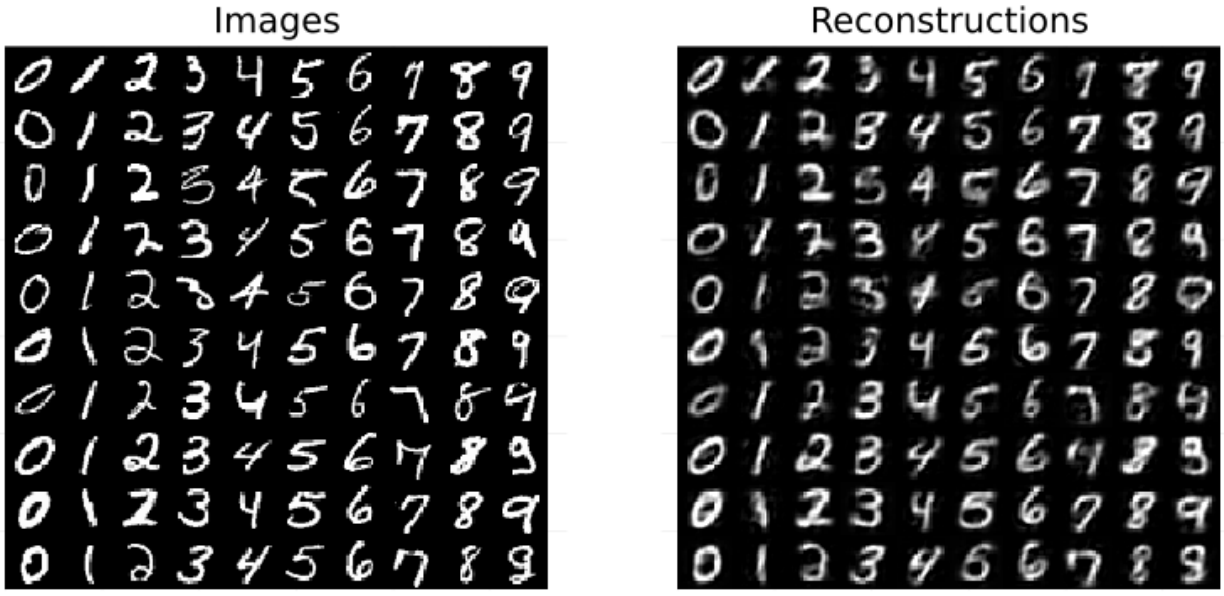

Figure 17: Randomly drawn *MNIST* images and their reconstructions generated by the VAE used by *REVISE*.