

# ECCCos from the Black Box

## Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals

**Patrick Altmeyer**

Mojtaba Farmanbar  
Cynthia C. S. Liem

Arie van Deursen

Delft University of Technology

2024-10-25

# Pick your Poison

All of these counterfactuals are valid explanations for the model's prediction.

*Which one would you pick?*



Figure 1: Turning a 9 into a 7: Counterfactual explanations for an image classifier produced using *Wachter* (Wachter, Mittelstadt, and Russell 2017), *Schut* (Schut et al. 2021) and *REVISE* (Joshi et al. 2019).

Faithfulness first, plausibility second.

# Faithfulness first, plausibility second.

We propose *ECCCo*: a new way to generate faithful model explanations that are as plausible as the underlying model permits.

# Summary

- **Idea:** generate counterfactuals that are consistent with what the model has learned about the data.

# Summary

- **Idea:** generate counterfactuals that are consistent with what the model has learned about the data.
- **Method:** constrain the model's energy and predictive uncertainty for the counterfactual.

# Summary

- **Idea:** generate counterfactuals that are consistent with what the model has learned about the data.
- **Method:** constrain the model's energy and predictive uncertainty for the counterfactual.
- **Result:** faithful counterfactuals that are as plausible as the model permits.

# Summary

- **Idea:** generate counterfactuals that are consistent with what the model has learned about the data.
- **Method:** constrain the model's energy and predictive uncertainty for the counterfactual.
- **Result:** faithful counterfactuals that are as plausible as the model permits.
- **Benefits:** enable us to distinguish trustworthy from unreliable models.



# Counterfactual Explanations

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

## Counterfactual Explanations (CE)

explain how inputs into a model need to change for it to produce different outputs.



Figure 2: Gradient-based counterfactual search.

# Reconciling Faithfulness and Plausibility

# Plausibility

## Definition (Plausible Counterfactuals)

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

## Why Plausibility?

Plausibility is positively associated with actionability, robustness (Artelt et al. 2021) and causal validity (Mahajan, Tan, and Sharma 2020).



Figure 3: Kernel density estimate (KDE) for the conditional distribution,  $p(\mathbf{x}|\mathbf{y}^+)$ , based on observed data. Counterfactual path as in Figure 2.

# Faithfulness

## Definition (Faithful Counterfactuals)

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

## Trustworthy Models

If the model posterior approximates the true posterior ( $p_\theta(\mathbf{x}|\mathbf{y}^+) \rightarrow p(\mathbf{x}|\mathbf{y}^+)$ ), faithful counterfactuals are also plausible.



Figure 4: KDE for learned conditional distribution,  $p_\theta(\mathbf{x}|\mathbf{y}^+)$ . Yellow stars indicate conditional samples generated through SGLD for a joint energy model (JEM).

# ECCCo

## Key Idea

Use the hybrid objective of joint energy models (JEM) and a model-agnostic penalty for predictive uncertainty: Energy-Constrained ( $\mathcal{E}_\theta$ ) Conformal ( $\Omega$ ) Counterfactuals (ECCCo).

ECCCo objective<sup>a</sup>:

$$\min_{\mathbf{Z}' \in \mathcal{Z}^L} \{L_{\text{clf}}(f(\mathbf{Z}'); M_\theta, \mathbf{y}^+) + \lambda_1 \text{cost}(f(\mathbf{Z}')) + \lambda_2 \mathcal{E}_\theta(f(\mathbf{Z}') | \mathbf{y}^+) + \lambda_3 \Omega(C_\theta(f(\mathbf{Z}'); \alpha))\}$$



Figure 5: Gradient fields and counterfactual paths for different generators.

<sup>a</sup>We leverage ideas from Grathwohl et al. (2020) and Stutz et al. (2022). See the paper and appendix for a derivation of the objective from first principles.

# Results

# Visual Evidence



Figure 6: Turning a 9 into a 7. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

*ECCCo* generates counterfactuals that

- faithfully represent model quality (Figure 6).
- achieve state-of-the-art plausibility (Figure 7).

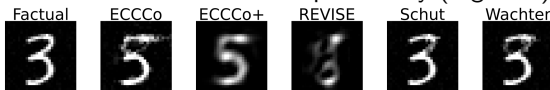


Figure 7: Results for different generators (from 3 to 5).

# The Numbers

- Large benchmarks on a variety of models and datasets from various domains.
- *ECCCo* achieves state-of-the-art faithfulness across models and datasets and approaches state-of-the-art plausibility for more trustworthy models.

| Model        | Generator      | California Housing   |                      |                      | GMSC                 |                      |                      |
|--------------|----------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|              |                | Unfaithfulness ↓     | Implausibility ↓     | Uncertainty ↓        | Unfaithfulness ↓     | Implausibility ↓     | Uncertainty ↓        |
| MLP Ensemble | ECCCo          | <b>3.69 ± 0.08**</b> | 1.94 ± 0.13          | <b>0.09 ± 0.01**</b> | 3.84 ± 0.07**        | 2.13 ± 0.08          | <b>0.23 ± 0.01**</b> |
|              | ECCCo+         | 3.88 ± 0.07**        | 1.20 ± 0.09          | 0.15 ± 0.02          | <b>3.79 ± 0.05**</b> | 1.81 ± 0.05          | 0.30 ± 0.01*         |
|              | ECCCo (no CP)  | 3.70 ± 0.08**        | 1.94 ± 0.13          | 0.10 ± 0.01**        | 3.85 ± 0.07**        | 2.13 ± 0.08          | 0.23 ± 0.01**        |
|              | ECCCo (no EBM) | 4.03 ± 0.07          | 1.12 ± 0.12          | 0.14 ± 0.01**        | 4.08 ± 0.06          | 0.97 ± 0.08          | 0.31 ± 0.01*         |
|              | REVISE         | 3.96 ± 0.07*         | <b>0.58 ± 0.03**</b> | 0.17 ± 0.03          | 4.09 ± 0.07          | <b>0.63 ± 0.02**</b> | 0.33 ± 0.06          |
|              | Schut          | 4.00 ± 0.06          | 1.15 ± 0.12          | 0.10 ± 0.01**        | 4.04 ± 0.08          | 1.21 ± 0.08          | 0.30 ± 0.01*         |
|              | Wachter        | 4.04 ± 0.07          | 1.13 ± 0.12          | 0.16 ± 0.01          | 4.10 ± 0.07          | 0.95 ± 0.08          | 0.32 ± 0.01          |
| JEM Ensemble | ECCCo          | 1.40 ± 0.08**        | 0.69 ± 0.05**        | 0.11 ± 0.00**        | 1.20 ± 0.06*         | 0.78 ± 0.07**        | 0.38 ± 0.01          |
|              | ECCCo+         | <b>1.28 ± 0.08**</b> | 0.60 ± 0.04**        | 0.11 ± 0.00**        | <b>1.01 ± 0.07**</b> | 0.70 ± 0.07**        | 0.37 ± 0.01          |
|              | ECCCo (no CP)  | 1.39 ± 0.08**        | 0.69 ± 0.05**        | 0.11 ± 0.00**        | 1.21 ± 0.07*         | 0.77 ± 0.07**        | 0.39 ± 0.01          |
|              | ECCCo (no EBM) | 1.70 ± 0.09          | 0.99 ± 0.08          | 0.14 ± 0.00*         | 1.31 ± 0.07          | 0.97 ± 0.10          | 0.32 ± 0.01**        |
|              | REVISE         | 1.39 ± 0.15**        | <b>0.59 ± 0.04**</b> | 0.25 ± 0.07          | 1.01 ± 0.07**        | <b>0.63 ± 0.04**</b> | 0.33 ± 0.07          |
|              | Schut          | 1.59 ± 0.10*         | 1.10 ± 0.06          | <b>0.09 ± 0.00**</b> | 1.34 ± 0.07          | 1.21 ± 0.10          | <b>0.26 ± 0.01**</b> |
|              | Wachter        | 1.71 ± 0.09          | 0.99 ± 0.08          | 0.14 ± 0.00          | 1.31 ± 0.08          | 0.95 ± 0.10          | 0.33 ± 0.01          |

Table 1: Results for tabular datasets: sample averages +/- one standard deviation across valid counterfactuals. The 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*).



# Questions?

# Questions?

With thanks to my co-authors Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem.



# Code

The code used to run the analysis for this work is built on top of `CounterfactualExplanations.jl`.

There is also a corresponding paper, *Explaining Black-Box Models through Counterfactuals*, which has been published in JuliaCon Proceedings.



Figure 8: Trustworthy AI in Julia: [github.com/JuliaTrustworthyAI](https://github.com/JuliaTrustworthyAI)

# References

- Artelt, André, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. 2021. "Evaluating Robustness of Counterfactual Explanations." In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 01–09. IEEE.
- Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. "Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One." In *International Conference on Learning Representations*.
- Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. "Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems." <https://arxiv.org/abs/1907.09615>.
- Mahajan, Divyat, Chenhao Tan, and Amit Sharma. 2020. "Preserving Causal Constraints in Counterfactual Explanations for Machine