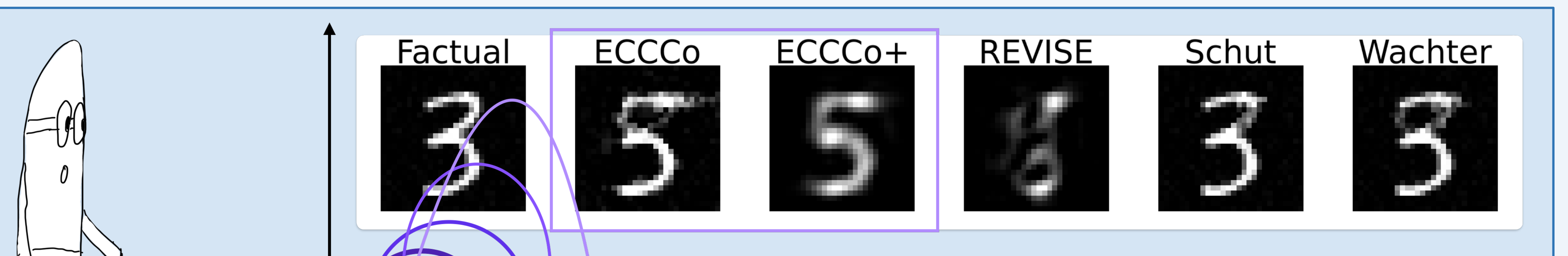
The 38th Annual
AAAI Conference on
Artificial Intelligence

Faithful Model Explanations through *E*nergy-*C*onstrained *C*onformal *Co*unterfactuals

Patrick Altmeyer (p.altmeyer@tudelft.nl),
Mojtaba Farmanbar,
Arie van Deursen,
Cynthia C. S. Liem



ECCCos from the Black Box

BACKGROUND

Counterfactual Explanations (CE) explain

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

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

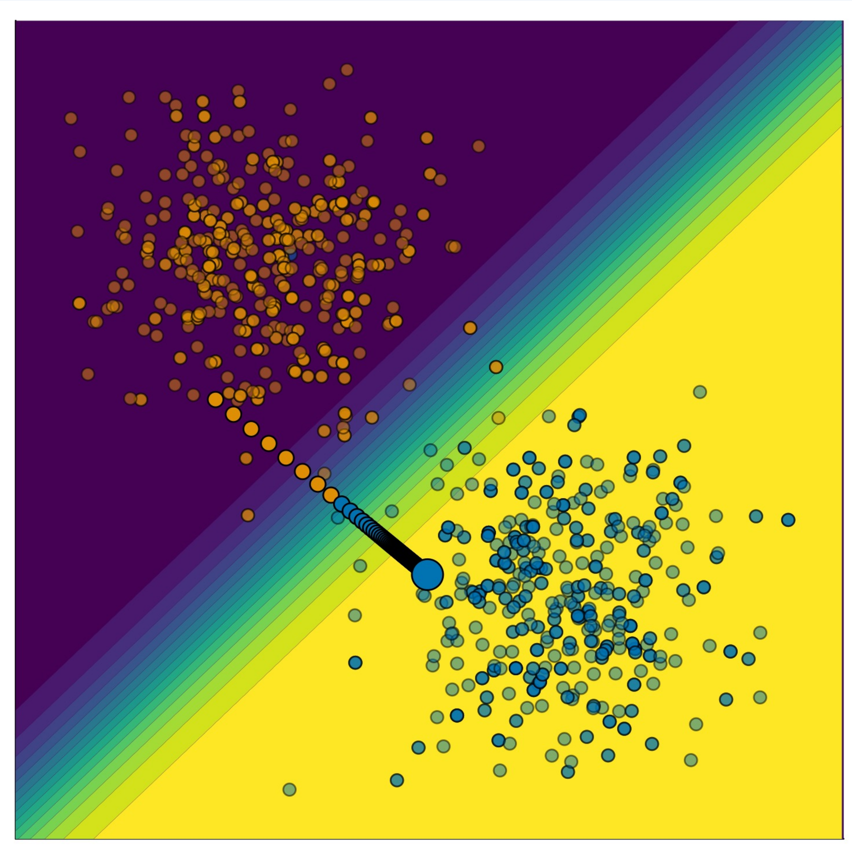


Figure 1: Gradient-based counterfactual search.

MOTIVATION

We argue that counterfactual explanations should only be as plausible as the model permits. In Figure 2,

> which counterfactual provides the most adequate explanation for the classifier?

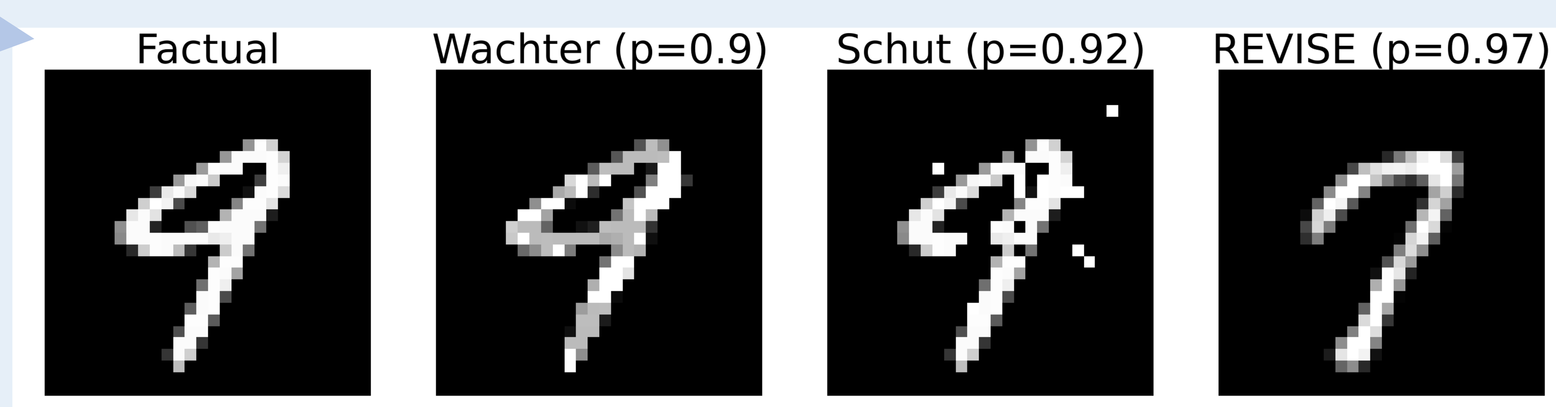


Figure 2: Factual images and counterfactuals for flipping the predicted label of a multi-layer perceptron (MLP) trained on MNIST from 9 to 7.

PLAUSIBILITY

We define plausible counterfactuals as:

> consistent with the true data generating process

Plausibility is positively associated with actionability, robustness and causal validity.

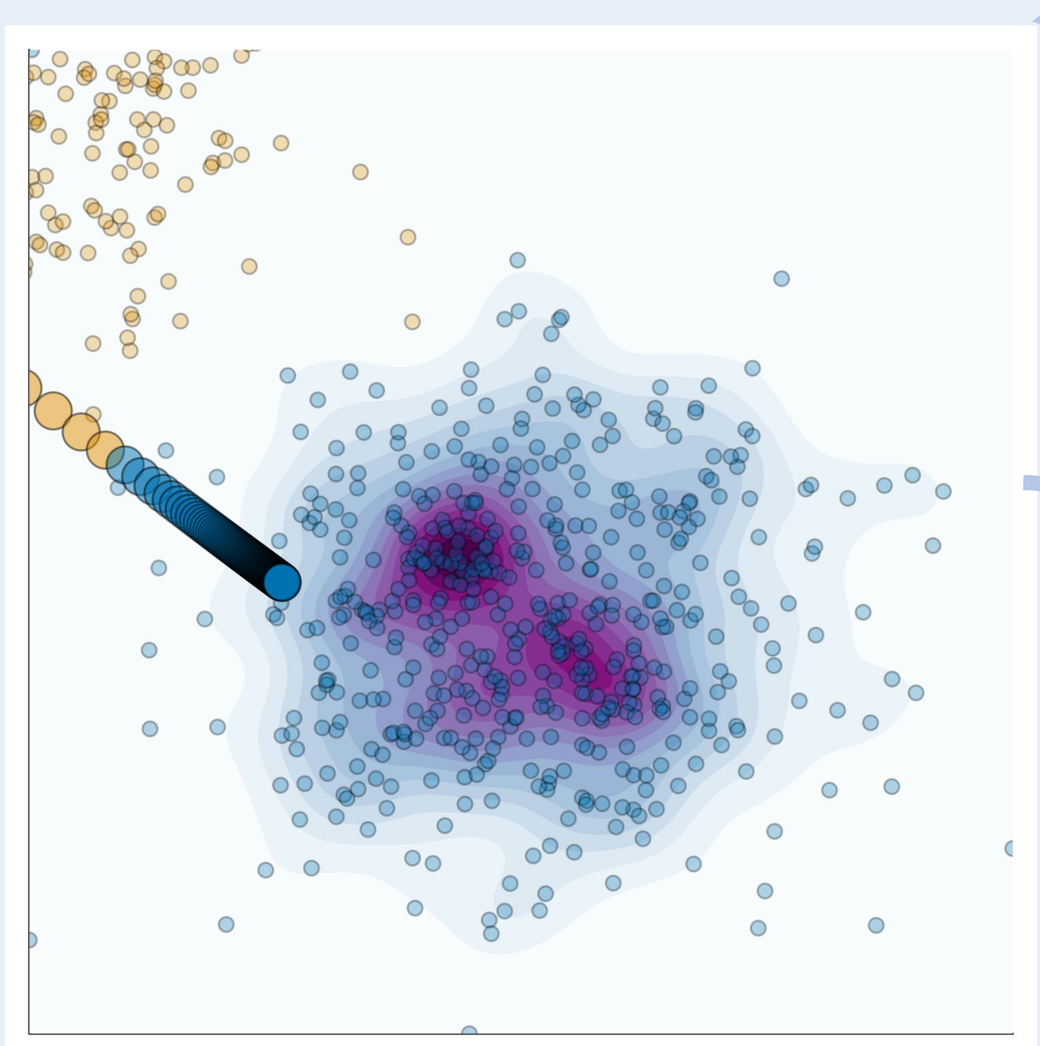


Figure 3: Kernel density estimate (KDE) for the conditional distribution based on observed data.

FAITHFULNESS

We define faithful counterfactuals as:

> consistent with what the model has learned about the data

If the model posterior approximates the true distribution, faithfulness and plausibility coincide.

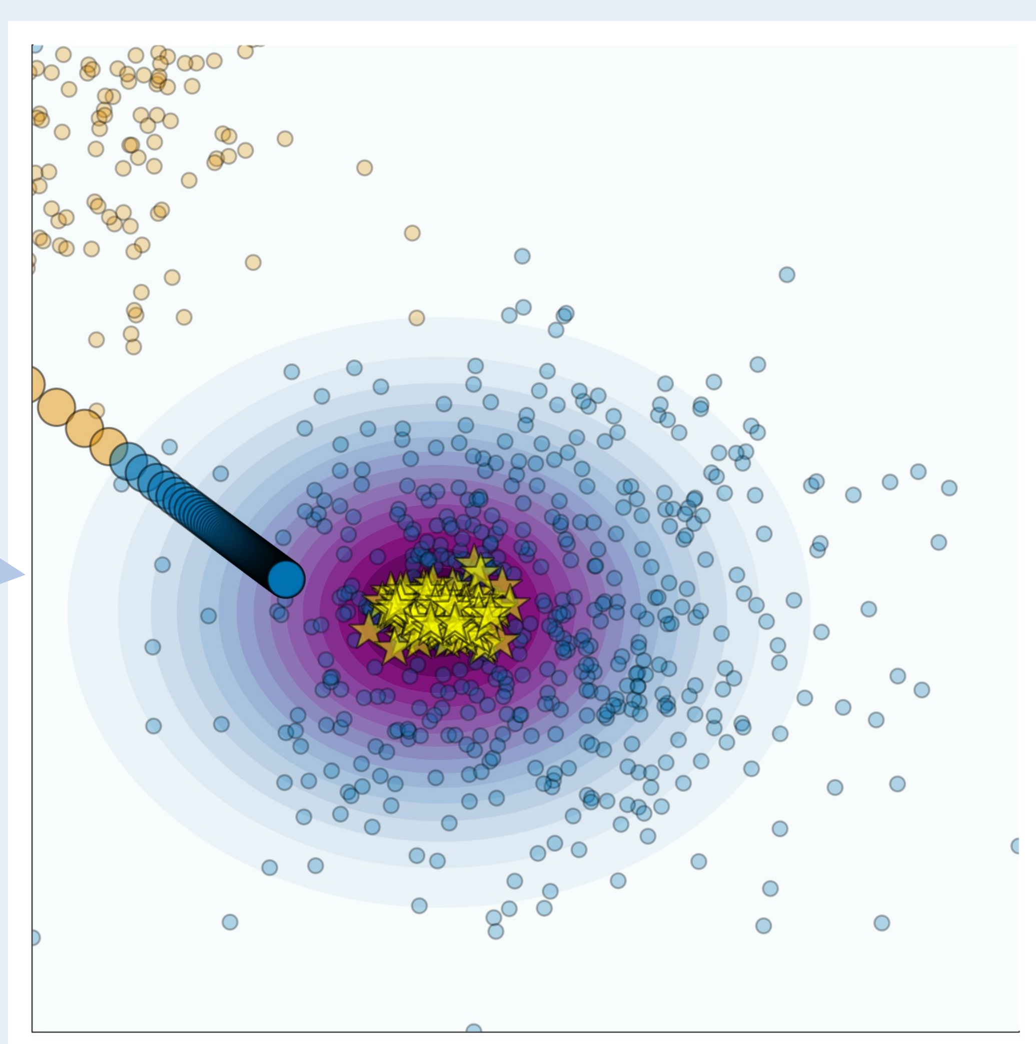


Figure 4: KDE for conditional distribution learned by model. Generated samples in bright yellow.

METHOD

Use the hybrid objective of joint energy models (JEM) and a model-agnostic penalty for predictive uncertainty:

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

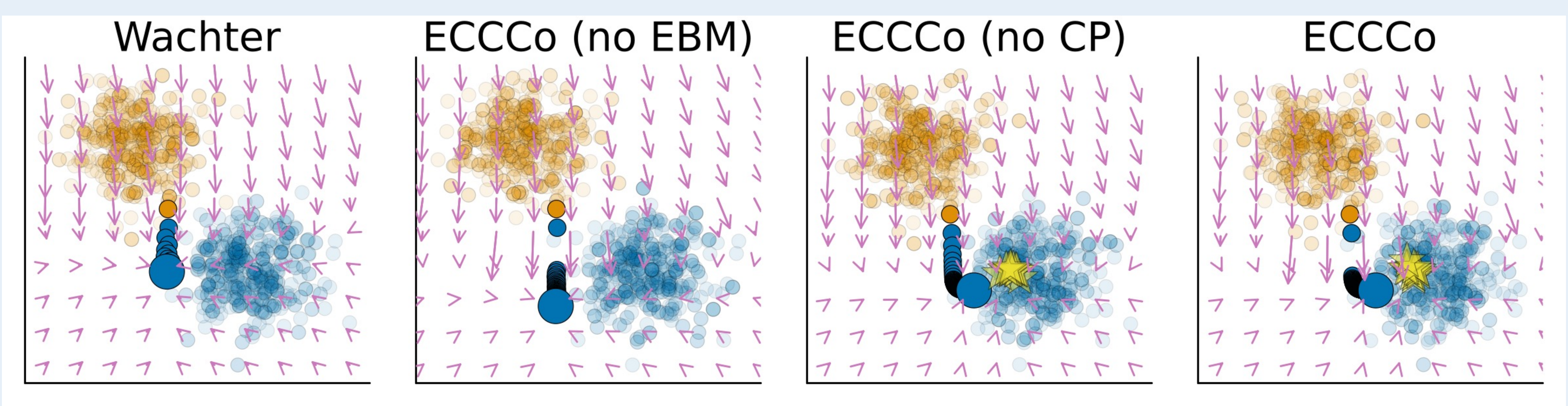


Figure 5: Gradient fields and counterfactual paths for different generators.

RESULTS

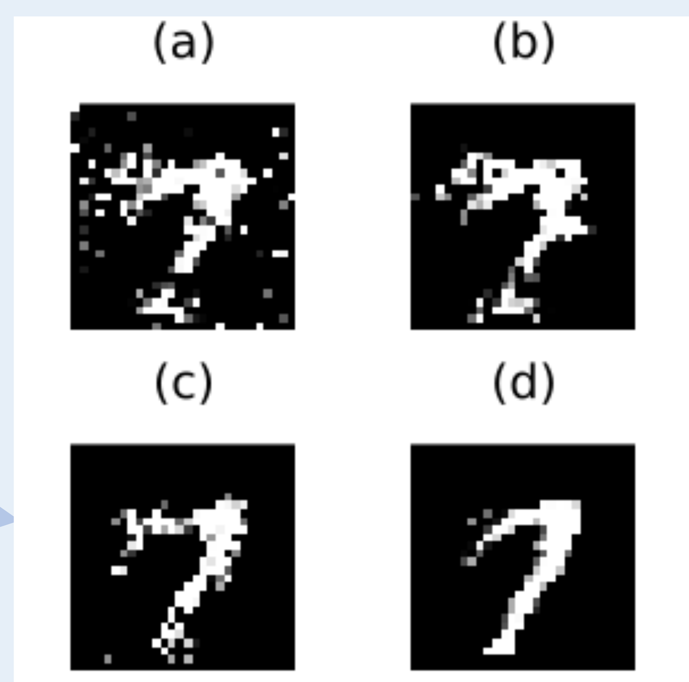


Figure 6: Turning a 'nine' into a 'seven'. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

ECCCo generates counterfactual explanations that

> faithfully represent model quality
> &
> achieve state-of-the-art plausibility

Thus, it can help humans to distinguish trustworthy from unreliable models.

		California Housing			GMSC		
Model	Generator	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓	Unfaithfulness ↓	Implausibility ↓	Uncertainty ↓
MLP Ensemble	ECCCo	**3.69 ± 0.08****	1.94 ± 0.13	**0.09 ± 0.01****	3.84 ± 0.07**	2.13 ± 0.08	**0.23 ± 0.01****
	ECCCo+	3.88 ± 0.07**	1.20 ± 0.09	0.15 ± 0.02	**3.79 ± 0.05****	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*	**0.58 ± 0.03****	0.17 ± 0.03	4.09 ± 0.07	**0.63 ± 0.02****	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+	**1.28 ± 0.08****	0.60 ± 0.04**	0.11 ± 0.00**	**1.01 ± 0.07****	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**	**0.59 ± 0.04****	0.25 ± 0.07	1.01 ± 0.07**	**0.63 ± 0.04****	0.33 ± 0.07
	Schut	1.59 ± 0.10*	1.10 ± 0.06	**0.09 ± 0.00****	1.34 ± 0.07	1.21 ± 0.10	**0.26 ± 0.01****
	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: Subsample of our empirical findings for tabular datasets.

LEARN MORE

Supplementary Appendix

GitHub Repository

Trustworthy AI in Julia - Taija

Personal Website