Explaining Models or Modelling Explanations

Counterfactual Explanations and Algorithmic Recourse for Trustworthy AI

Patrick Altmeyer

Delft University of Technology

2026-04-16

Background

- Economist, then PhD CS
- How can we make opaque AI more trustworthy?
- Explainable AI, Adversarial ML, Probabilistic ML
- </> Core developer and maintainer of Taija (Trustworthy AI in Julia)

Figure 1: Scan for slides.
Links to www.patalt.org.

Agenda

- **Intro**: counterfactual explanations (CE) and algorithmic recourse (AR)

Agenda

- **Intro**: counterfactual explanations (CE) and algorithmic recourse (AR)
- **Unexpected Challenges**: endogenous dynamics of AR

Agenda

- **Intro**: counterfactual explanations (CE) and algorithmic recourse (AR)
- **Unexpected Challenges**: endogenous dynamics of AR
- **Paradigm Shift**: explanations should be faithful first, plausible second

Agenda

- **Intro**: counterfactual explanations (CE) and algorithmic recourse (AR)
- **Unexpected Challenges**: endogenous dynamics of AR
- **Paradigm Shift**: explanations should be faithful first, plausible second
- **New Opportunities**: teaching models plausible explanations through CE

Intro

Training Opaque Models

Tweaking Parameters

Objective:

$$\min_{\theta}\{\mathsf{yloss}(M_\theta(\mathbf{x}), \mathbf{y})\}$$

Training Opaque Models

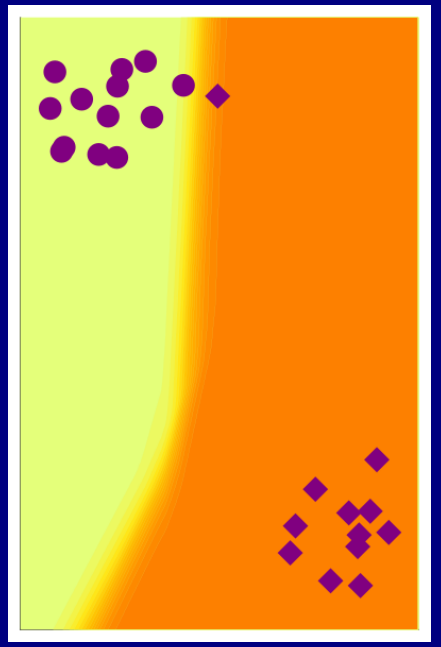
Tweaking Parameters

Objective:

$$\min_{\theta}\{\mathsf{yloss}(M_\theta(\mathbf{x}), \mathbf{y})\}$$

Solution:

$$\theta_{t+1} = \theta_t - \nabla_\theta\{\mathsf{yloss}(M_\theta(\mathbf{x}), \mathbf{y})\}$$
$$\theta^* = \theta_T$$

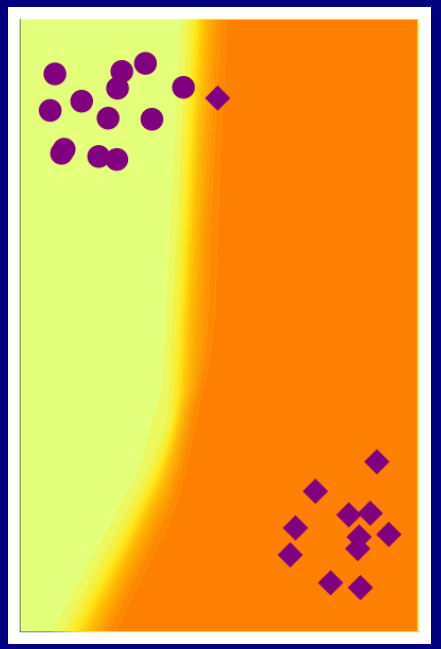


Explaining Opaque Models

Tweaking Inputs

Objective:

$$\min_{\mathbf{x}} \{\mathsf{yloss}(M_{\theta^*}(\mathbf{x}), \mathbf{y}^+) + \lambda \mathsf{reg}(\mathbf{x}; \cdot)\}$$



Explaining Opaque Models

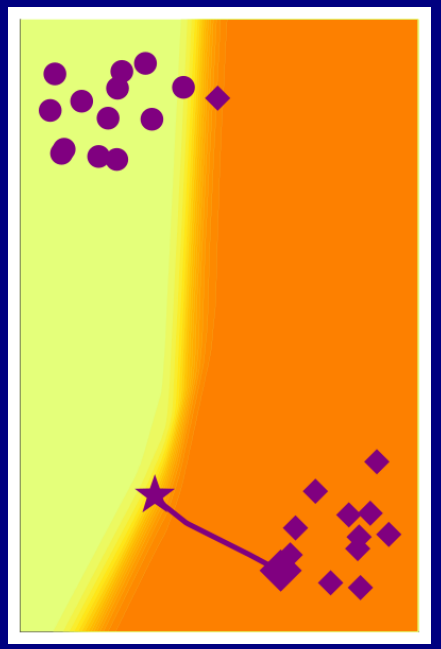
Tweaking Inputs

Objective:

$$\min_{\mathbf{x}} \{\text{yloss}(M_{\theta^*}(\mathbf{x}), \mathbf{y}^+) + \lambda \text{reg}(\mathbf{x}; \cdot)\}$$

Solution:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{x}_t - \nabla_{\mathbf{x}} \{\text{yloss}(M_{\theta^*}(\mathbf{x}), \mathbf{y}^+) \\ &\quad + \lambda \text{reg}(\mathbf{x}; \cdot)\} \\ \mathbf{x}^* &= \mathbf{x}_T \end{aligned}$$



Algorithmic Recourse

Provided CE is valid, plausible and actionable, it can be used to provide recourse to individuals negatively affected by models.

"If your income had been x, then ..."

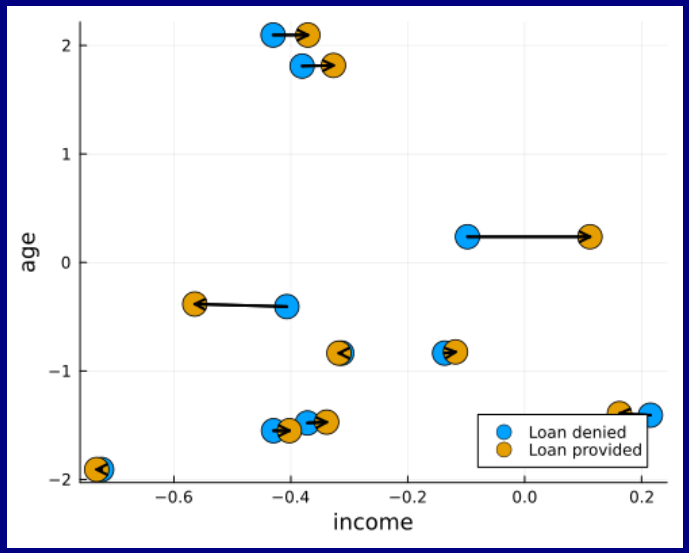


Figure 2: Counterfactuals for random samples from the Give Me Some Credit dataset (Kaggle 2011). Features 'age' and 'income' are shown.

Unexpected Challenges

Hidden Cost of Implausibility

AR can introduce costly dynamics[1]

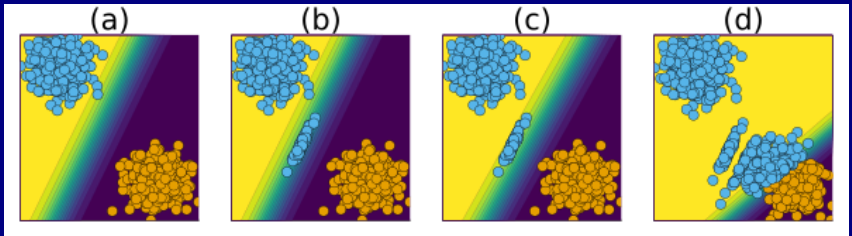


Figure 3: Endogenous Macrodynamics in Algorithmic Recourse.

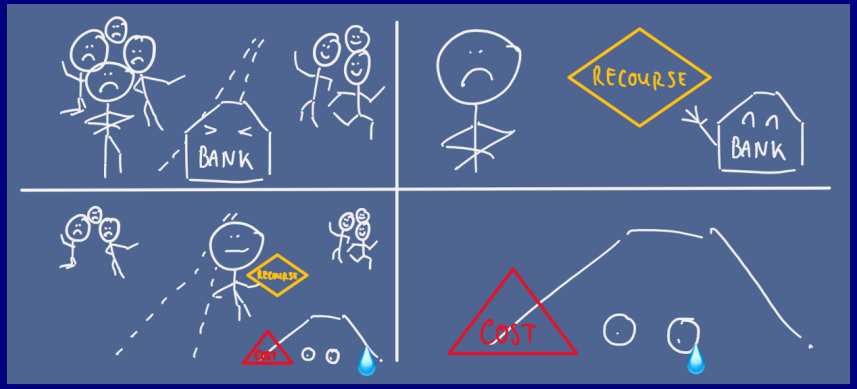


Figure 4: Illustration of external cost of individual recourse.

🔑 **Insight**: Implausible Explanations Are Costly

[1] Altmeyer et al. (2023) @ SaTML 2023.

Mitigation Strategies

Reframed Objective

$$\mathbf{s}' = \arg\min_{\mathbf{s}' \in \mathcal{S}} \{\text{yloss}(M(f(\mathbf{s}')), y^*) + \lambda_1 \text{cost}(f(\mathbf{s}')) + \lambda_2 \text{extcost}(f(\mathbf{s}'))\}$$

- Even simple mitigation strategies can help.
- Reducing hidden cost is (roughly) equivalent to ensuring plausibility.

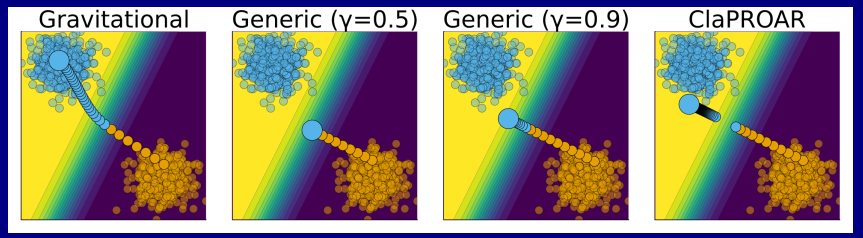


Figure 5: Mitigation strategies to tackle hidden costs of AR.

Explanation or Adversarial Example?

Plausibility at all cost?

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

Pick your poison ...

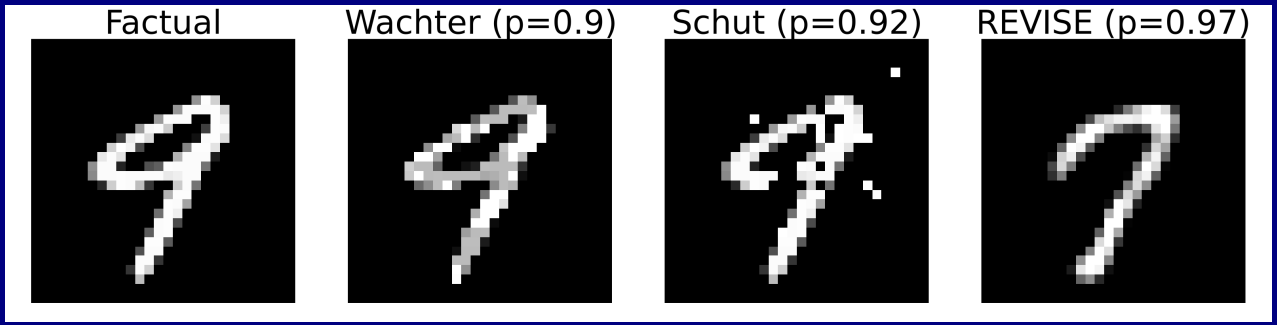


Figure 6: Turning a 9 into a 7: Counterfactual explanations for an image classifier using different approaches (Altmeyer, Farmanbar, et al. 2024).

Faithful First, Plausible Second

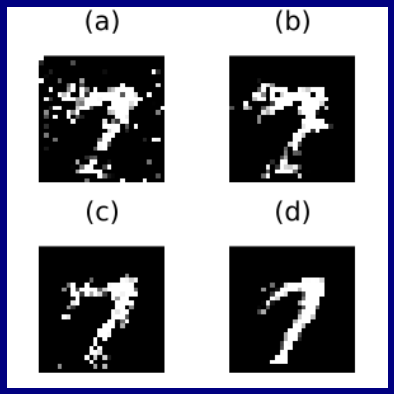


Figure 7: Turning a 9 into a 7. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

🔑 **Insight**: faithfulness facilitates[2]

- model quality checks (Figure 7).
- state-of-the-art plausibility (Figure 8).

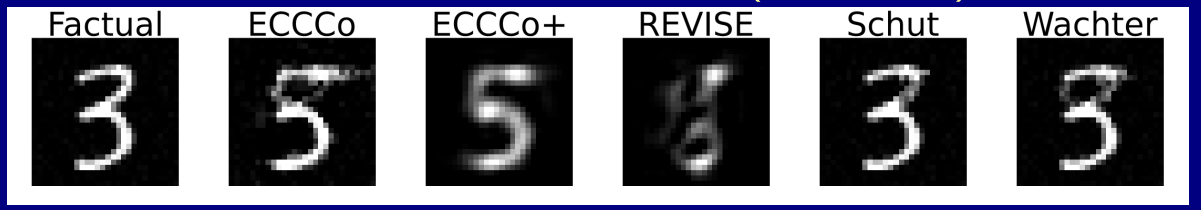


Figure 8: Results for different generators (from 3 to 5).

[2] Altmeyer, Farmanbar, et al. (2024) @ AAAI 2024. [blog]

Putting it all together

Counterfactual Training

First, *Tweaking Inputs*[3]

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \nabla_{\mathbf{x}}\{ECCCo(M_{\theta^*}(\mathbf{x}), \mathbf{y}^+)\}$$
$$\mathbf{x}^* = \mathbf{x}_T$$

Then, *Tweaking Parameters*

$$\theta_{t+1} = \theta_t - \nabla_{\theta}\{\text{yloss}(M_\theta(\mathbf{x}), \mathbf{y}) + \text{div}(\mathbf{x}^*, \mathbf{x}^+, y^+; \theta)\}$$
$$\theta^* = \theta_T$$

[3] Generate faithful explanations using *ECCCo* objective (Altmeyer, Farmanbar, et al. 2024).

Counterfactual Training

1. Contrast faithful CE with data $\rightarrow$ **Explainability** $\uparrow$

Counterfactual Training

1. Contrast faithful CE with data $\rightarrow$ **Explainability** $\uparrow$
2. Feature mutability constraints $\rightarrow$ **Actionability** $\uparrow$(holds provably under certain assumptions)

Counterfactual Training

1. Contrast faithful CE with data $\rightarrow$ **Explainability** $\uparrow$
2. Feature mutability constraints $\rightarrow$ **Actionability** $\uparrow$(holds provably under certain assumptions)
3. Bonus: use nascent CE as AE $\rightarrow$ **Robustness** $\uparrow$

Counterfactual Training

1. Contrast faithful CE with data $\rightarrow$ **Explainability** $\uparrow$
2. Feature mutability constraints $\rightarrow$ **Actionability** $\uparrow$(holds provably under certain assumptions)
3. Bonus: use nascent CE as AE $\rightarrow$ **Robustness** $\uparrow$

Counterfactual Training

1. Contrast faithful CE with data → **Explainability** ↑
2. Feature mutability constraints → **Actionability** ↑(holds provably under certain assumptions)
3. Bonus: use nascent CE as AE → **Robustness** ↑

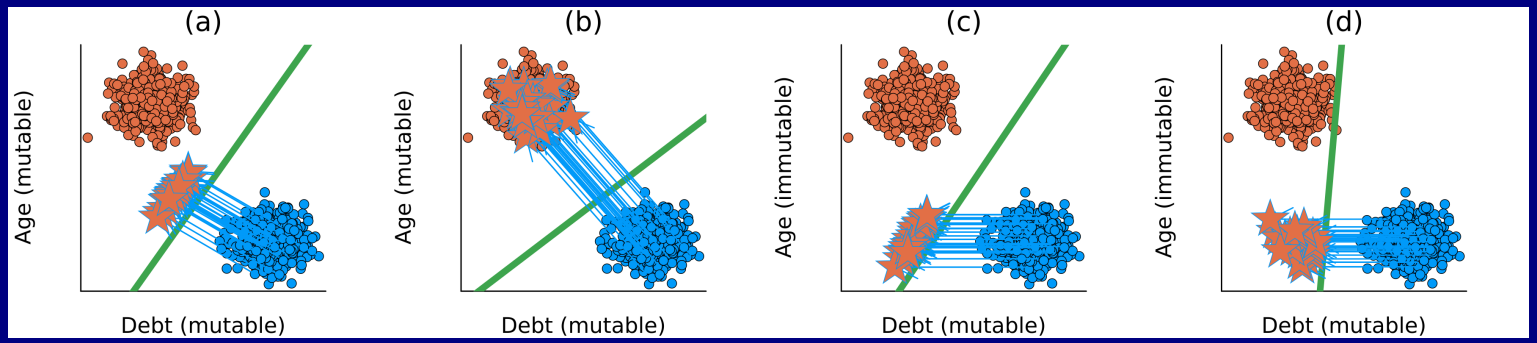


Figure 9: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable.

Counterfactual Training: Results

Figure 10: **Plausibility**: Visual explanations (counterfactuals) for baseline (top row) vs CT (bottom).

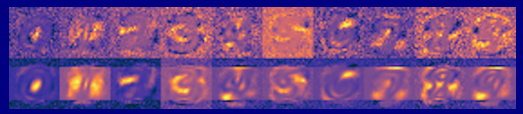


Figure 11: **Actionability**: Visual explanations (integrated gradients) as before. Five top and bottom rows immutable.

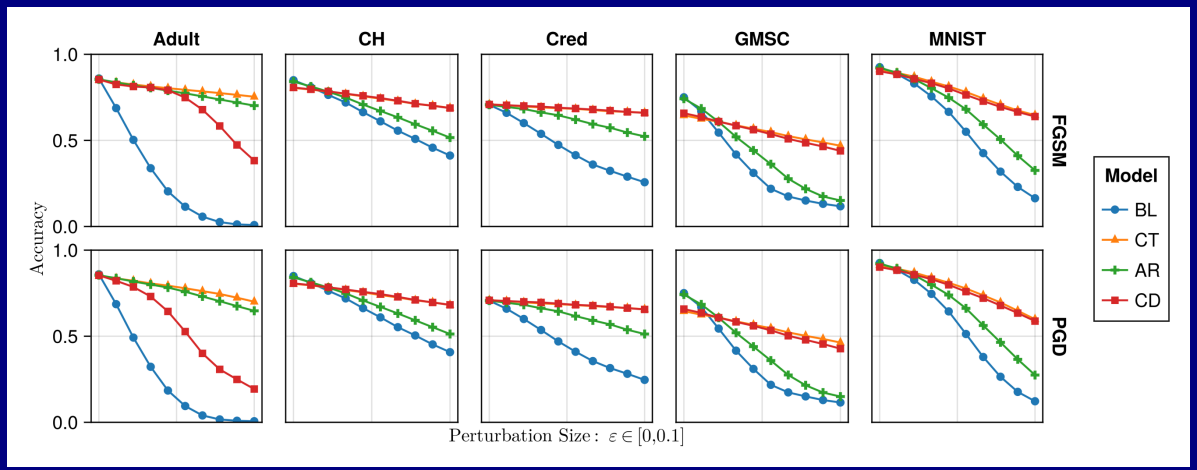


Figure 12: Test accuracies on adversarially perturbed data with varying perturbation sizes.

The Hard Numbers

Extensive experiments and ablation studies on nine datasets–synthetic, tabular and vision–generating millions of counterfactuals:[4]

1. **Plausibility** of CEs increases by up to 90%.
2. **Actionability**: cost of reaching valid counterfactuals with protected features decreases by 19% on average.
3. Models' adversarial **robustness** improves consistently.

[4] Facilitated by our CounterfactualExplanations.jl (Altmeyer, Deursen, and Liem 2023) with multi-processing support and (DHPC) (2022).

Check it out!

Figure 13: Preprint

Figure 14: Software

Figure 15: Homepage

Taija

- Model Explainability (CounterfactualExplanations.jl)
- Predictive Uncertainty Quantification (ConformalPrediction.jl)
- Effortless Bayesian Deep Learning (LaplaceRedux.jl)
- ... and more!

- Work presented @ JuliaCon 2022, 2023, 2024.
- Google Summer of Code and Julia Season of Contributions 2024.
- Total of three software projects @ TU Delft.

Figure 16: Trustworthy AI in Julia: github.com/JuliaTrustworthyAI

If we still have time ...

Spurious Sparks of AGI

We challenge the idea that the finding of meaningful patterns in latent spaces of large models is indicative of AGI[5].

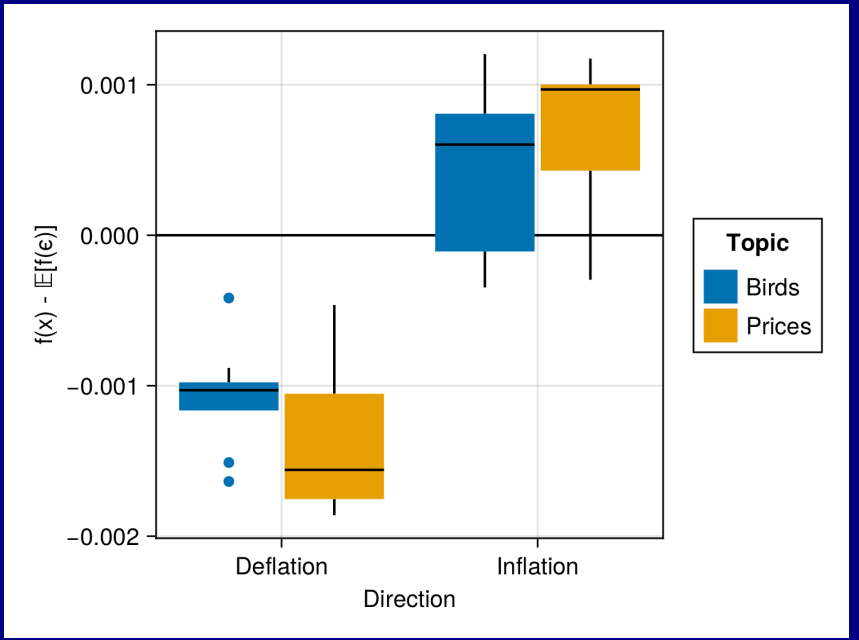


Figure 17: Inflation of prices or birds? It doesn't matter!

[5] In Altmeyer, Demetriou, et al. (2024) @ ICML 2024

References

Altmeyer, Patrick, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Endogenous Macrodynamics in Algorithmic Recourse." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 418–31. IEEE. https://doi.org/10.1109/satml54575.2023.00036.

Altmeyer, Patrick, Andrew M Demetriou, Antony Bartlett, and Cynthia C. S. Liem. 2024. "Position: Stop Making Unscientific AGI Performance Claims." In *International Conference on Machine Learning*, 1222–42. PMLR. https://proceedings.mlr.press/v235/altmeyer24a.html.

Altmeyer, Patrick, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Explaining Black-Box Models through Counterfactuals." In *Proceedings of the JuliaCon Conferences*, 1:130. https://doi.org/10.21105/jcon.00130.

Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2024. "Faithful Model Explanations through