Explaining Models or Modelling Explanations

Counterfactual Explanations and Algorithmic Recourse for Trustworthy AI

Patrick Altmeyer Arie van Deursen Cynthia C. S. Liem

Delft University of Technology

2026-02-23

Background

- Economist, now 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

A Toy Problem

Figure 2: Cats and dogs in two dimensions.

Traversing the Parameter Space

Model Training

Objective:

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

Traversing the Parameter Space

Model Training

Objective:

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

Solution:

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

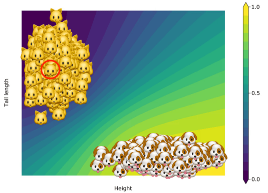

Figure 3: Fitted model. Contour shows predicted probability $y =$.

Traversing the Feature Space

Counterfactual Search

Objective:

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

Traversing the Feature Space

Counterfactual Search

Objective:

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

Solution:

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

$$\mathbf{x}^* = \mathbf{x}_T$$

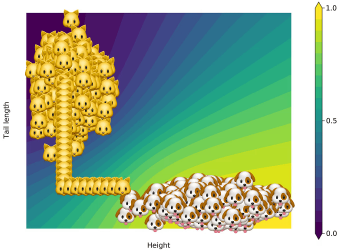

Figure 4: Counterfactual explanation for what it takes to be a dog.

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 ..."*

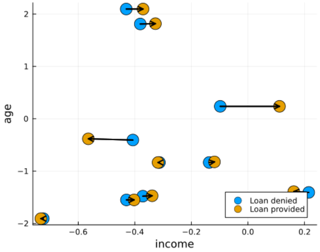

Figure 5: Counterfactuals for random samples from the Give Me Some Credit dataset (Kaggle 2011). Features 'age' and 'income' are shown.

Unexpected Challenges

Hidden Cost of Implausibility

AR can introduce costly dynamics[1]

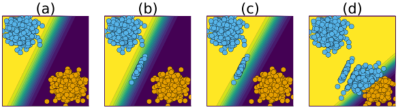

Figure 6: Endogenous Macrodynamics in Algorithmic Recourse.

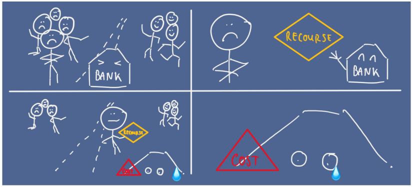

Figure 7: Illustration of external cost of individual recourse.

Insight: Implausible Explanations Are Costly

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

Mitigation Strategies

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

Reframed Objective

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

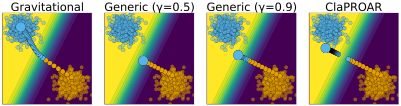

Figure 8: Mitigation strategies to tackle hidden costs of AR.

Paradigm Shift

Plausibility at all cost?

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

Pick your poison ...

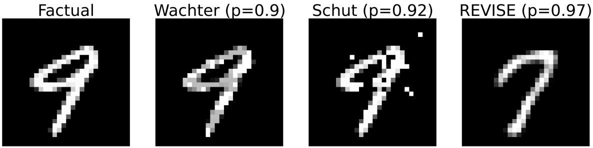

Figure 9: Turning a 9 into a 7: Counterfactual explanations for an image classifier produced using *Wachter* (Wachter, Mittelstadt, and Russell 2017), *Schut* (Schut et al. 2021) and *REVISE* (Joshi et al. 2019).

Faithful First, Plausible Second

Counterfactuals as plausible as the model permits[2].

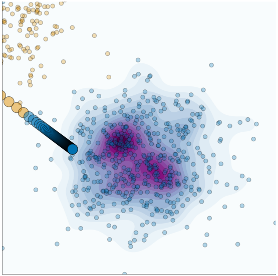

Figure 10: KDE for training data.

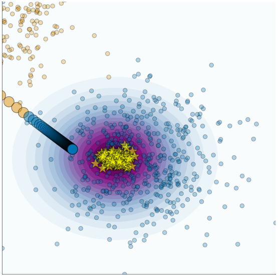

Figure 11: KDE for model posterior.

[2] Altmeyer, Farmanbar, et al. (2023) @ AAAI 2024. [blog]

Faithful Counterfactuals

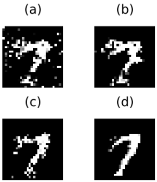

Figure 12: Turning a 9 into a 7. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

Insight: faithfulness facilitates

- model quality checks (Figure 12).
- state-of-the-art plausibility (Figure 13).

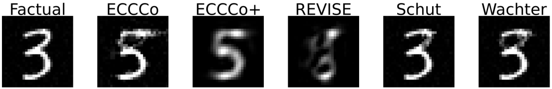

Figure 13: Results for different generators (from 3 to 5).

New Opportunities

Counterfactual Training: Method

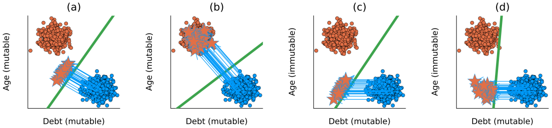

Figure 14: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable.

1. Contrast faithful CE with data.
2. Enforce actionability constraints.
3. Bonus: use nascent CE as AE.

Insight: We can hold models accountable for plausible explanations[3].

Counterfactual Training: Results

- Models trained with CT learn more plausible and (provably) actionable explanations.
- Predictive performance does not suffer, robust performance improves.

Figure 15: *Plausibility*: BL (top row) vs CT using the *ECCCo* generator (bottom row) counterfactuals for a randomly selected factual from class "0" (in blue). CT produces more plausible counterfactuals than BL.

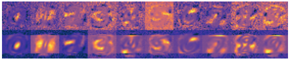

Figure 16: *Actionability*: Sample visual explanations (integrated gradients) for the *MNIST* dataset. Mutability constraints are imposed on the five top and bottom rows of pixels. CT (bottom) is less sensitive to protected features.

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[4].

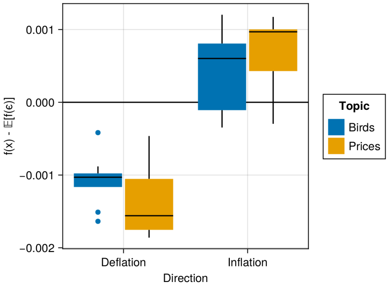

Figure 17: Inflation of prices or birds? It doesn't matter!

4 In Altmeyer et al. (2024) @ ICML 2024

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 18: Trustworthy AI in Julia: github.com/JuliaTrustworthyAI

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, Aleksander Buszydlik, Arie van Deursen, and Cynthia C. S. Liem. 2026. "Counterfactual Training: Teaching Models Plausible and Actionable Explanations." In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE. upcoming.

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, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Faithful Model Explanations Through