Explaining Models or Modelling Explanations

Challenging Existing Paradigms in Trustworthy AI

Patrick Altmeyer Arie van Deursen Cynthia C. S. Liem

Delft University of Technology

2025-05-07

Background

- Economist, now PhD CS
- How can we make opaque AI more trustworthy?
- Explainable AI, Adversarial ML, Probabilistic ML
- Maintainer of Taija (trustworthy AI in Julia)

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

Agenda

- What are counterfactual explanations (CE) and algorithmic recourse (AR) and why are they useful?

Agenda

- What are counterfactual explanations (CE) and algorithmic recourse (AR) and why are they useful?
- What dynamics are generated when off-the-shelf solutions to CE and AR are implemented in practice?

Agenda

- What are counterfactual explanations (CE) and algorithmic recourse (AR) and why are they useful?
- What dynamics are generated when off-the-shelf solutions to CE and AR are implemented in practice?
- Can we generate plausible counterfactuals relying only on the opaque model itself?

Agenda

- What are counterfactual explanations (CE) and algorithmic recourse (AR) and why are they useful?
- What dynamics are generated when off-the-shelf solutions to CE and AR are implemented in practice?
- Can we generate plausible counterfactuals relying only on the opaque model itself?
- How can we leverage counterfactuals during training to build more trustworthy models?

Background

Intuition

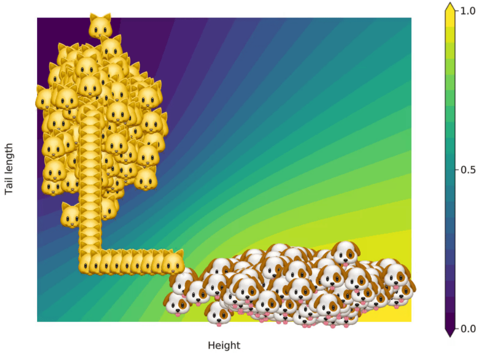

Figure 2: Counterfactual explanation for what it takes to be a dog.

Methodology

Model Training

Objective:

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

Methodology

Model Training

Objective:

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

Solution:

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

Methodology

Counterfactual Search

Objective:

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

Methodology

Counterfactual Search

Objective:

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

Solution:

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

Methodology

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

Counterfactual Explanations explain how inputs into a model need to change for it to produce different outputs[a].

[a] Altmeyer, Deursen, and Liem (2023) @ JuliaCon 2022

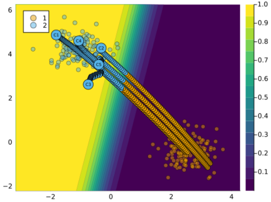

Figure 3: Gradient-based counterfactual search.

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 ..."*

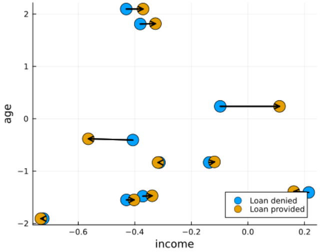

Figure 4: Counterfactuals for random samples from the Give Me Some Credit dataset (Kaggle 2011). Features 'age' and 'income' are shown.

Dynamics of CE and AR

Hidden Cost of Implausibility

AR can introduce costly dynamics[a]

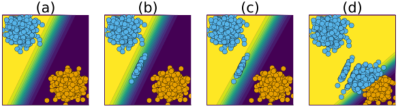

Figure 5: Endogenous Macrodynamics in Algorithmic Recourse.

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

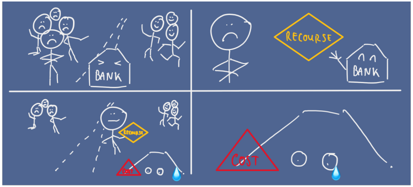

Figure 6: Illustration of external cost of individual recourse.

Insight: individual recourse neglects bigger picture.

Mitigation Strategies

- Incorporate hidden cost in reframed objective (**?@eq-satml**).
- Reducing hidden cost is equivalent to ensuring plausibility.

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

Plausibility at all cost?

Pick your Poison

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

Which one would you pick?

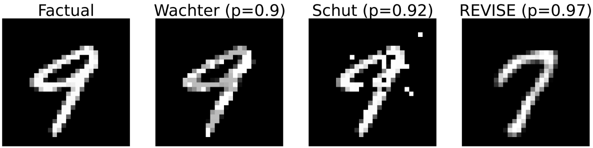

Figure 7: Turning a 9 into a 7: Counterfactual explanations for an image classifier produced using *Wachter* (Wachter, Mittelstadt, and Russell 2017), *Schut* (Schut et al. 2021) and *REVISE* (Joshi et al. 2019).

Faithful First, Plausible Second

Counterfactuals as plausible as the model permits[1].

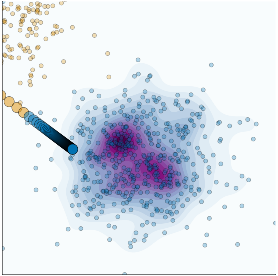

Figure 8: KDE for training data.

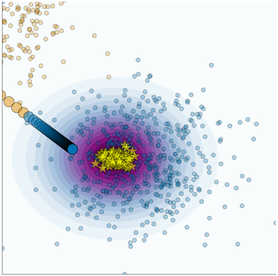

Figure 9: KDE for model posterior.

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

Faithful Counterfactuals

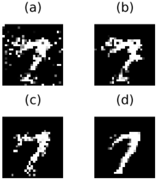

Figure 10: Turning a 9 into a 7. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

Insight: faithfulness facilitates

- model quality checks (Figure 10).
- state-of-the-art plausibility (Figure 11).

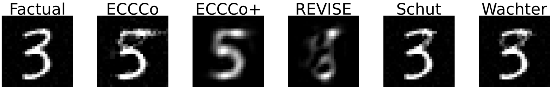

Figure 11: Results for different generators (from 3 to 5).

Teaching models plausible explanations

Counterfactual Training: Method

Idea

Let the model compare its own explanations to plausible ones[2].

1. Contrast faithful counterfactuals with data.
2. Use nascent CE as adversarial examples.

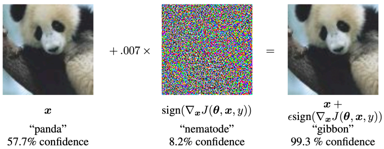

Figure 12: Example of an adversarial attack. Source: Goodfellow, Shlens, and Szegedy (2015)

[2] under review

Counterfactual Training: Results

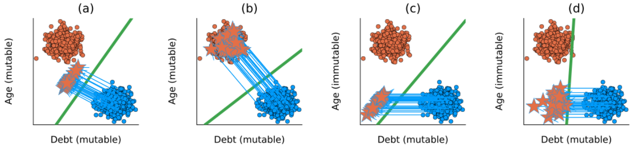

Figure 13: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, *age* immutable; (d) CT, *age* immutable.

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

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[a].

[a] In Altmeyer et al. (2024) @ ICML 2024

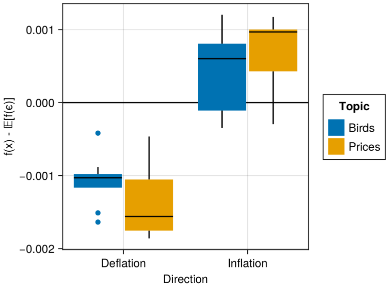

Figure 14: Inflation of prices or birds? It doesn't matter!

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

References

Altmeyer, Patrick, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia CS Liem. 2023. "Endogenous Macrodynamics in Algorithmic Recourse." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 418–31. IEEE.

Altmeyer, Patrick, Andrew M. Demetriou, Antony Bartlett, and Cynthia C. S. Liem. 2024. "Position Paper: Against Spurious Sparks-Dovelating Inflated AI Claims." https://arxiv.org/abs/2402.03962.

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.

Altmeyer, Patrick, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. 2023. "Faithful Model Explanations Through Energy-Constrained Conformal Counterfactuals."