

Background
oooooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
oooo

Teaching models plausible explanations
ooo

If we still have time ...
oooo

Explaining Models or Modelling Explanations

Challenging Existing Paradigms in Trustworthy AI

Patrick Altmeyer Arie van Deursen Cynthia C. S. Liem

Delft University of Technology

2026-02-26

Background

👤 Economist, now PhD CS

❓ How can we make opaque AI more trustworthy?

🤖 Explainable AI, Adversarial ML, Probabilistic ML

leftrightarrow Maintainer of Taija (trustworthy AI in Julia)



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

Background
oooooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
oooo

Teaching models plausible explanations
ooo

If we still have time ...
oooo

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
●oooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
oooo

Teaching models plausible explanations
ooo

If we still have time ...
oooo

Background

Background
○●○○○○○○

Dynamics of CE and AR
○○○

Plausibility at all cost?
○○○○

Teaching models plausible explanations
○○○

If we still have time ...
○○○○

CE in Five Slides



Figure 2: Cats and dogs in two dimensions.

Background
○○●○○○○○

Dynamics of CE and AR
○○○

Plausibility at all cost?
○○○○

Teaching models plausible explanations
○○○

If we still have time ...
○○○○

CE in Five Slides

Model Training

Objective:

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

CE in Five Slides

Model Training

Objective:

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

Solution:

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

CE in Five Slides

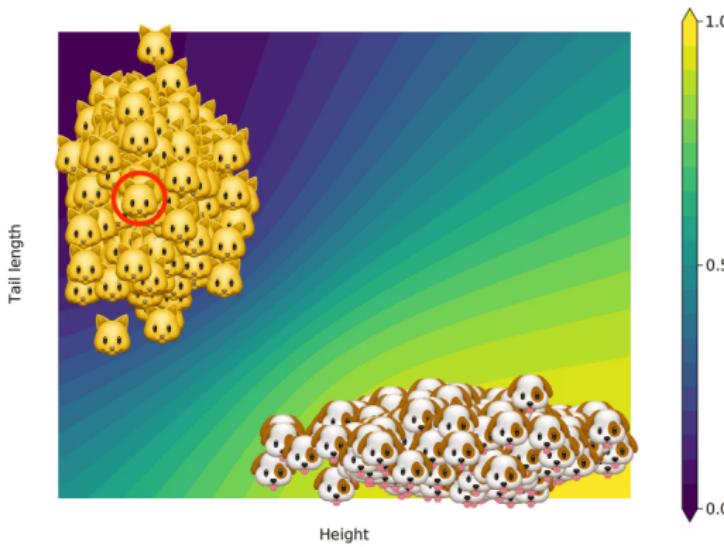


Figure 3: Fitted model. Contour shows predicted probability $y = .$

Background
oooooooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
oooo

Teaching models plausible explanations
ooo

If we still have time ...
oooo

CE in Five Slides

Counterfactual Search

Objective:

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

CE in Five Slides

Counterfactual Search

Objective:

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

Solution:

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

CE in Five Slides

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

Counterfactual Explanations explain how inputs into a model need to change for it to produce different outputs¹.

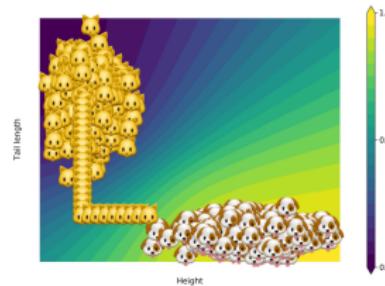


Figure 4: Counterfactual explanation for what it takes to be a dog.

¹ 

Altmeyer, Deursen, and Liem (2023) © JuliaCon 2022

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 ..."*

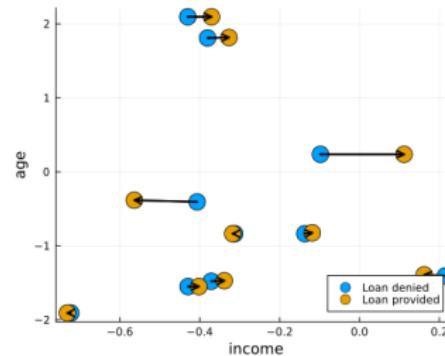


Figure 5: Counterfactuals for random samples from the Give Me Some Credit dataset (Kaggle 2011). Features ‘age’ and ‘income’ are shown.

Background
oooooooooo

Dynamics of CE and AR
●○○

Plausibility at all cost?
○○○○

Teaching models plausible explanations
○○○

If we still have time ...
○○○○

Dynamics of CE and AR

Hidden Cost of Implausibility

AR can introduce costly dynamics²

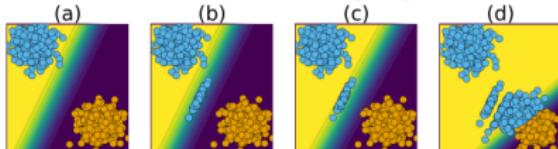


Figure 6: Endogenous Macrodynamics in Algorithmic Recourse.



Figure 7: Illustration of external cost of individual recourse.

🔑 Insight: individual recourse neglects bigger picture.



Altmeyer, Angela, et al. (2023) © SaTML 2023.

Mitigation Strategies

- Incorporate hidden cost in reframed objective.
- Reducing hidden cost is equivalent to ensuring plausibility.

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

Background
oooooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
●ooo

Teaching models plausible explanations
ooo

If we still have time ...
oooo

Plausibility at all cost?

Pick your Poison

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

Which one would you pick?

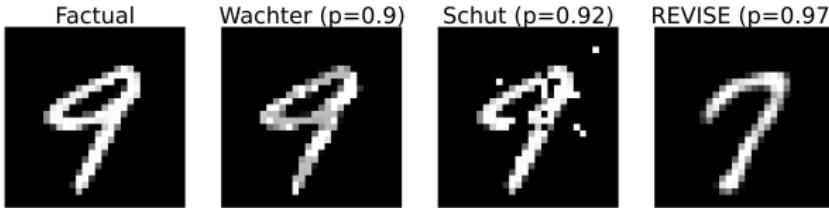


Figure 8: Turning a 9 into a 7: Counterfactual explanations for an image classifier produced using *Wachter* (Wachter, Mittelstadt, and Russell 2017), *Schut* (Schut et al. 2021) and *REVISE* (Joshi et al. 2019).

Faithful First, Plausible Second

Counterfactuals as plausible as the model permits³.

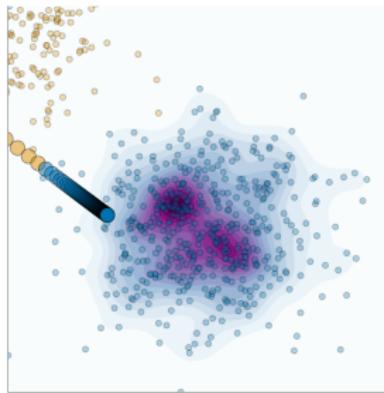


Figure 9: KDE for training data.

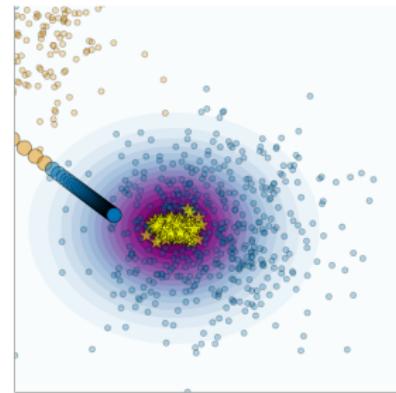


Figure 10: KDE for model posterior.

³■

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

Faithful Counterfactuals

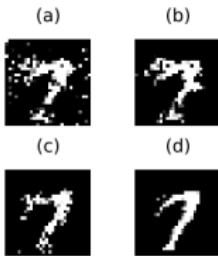


Figure 11: Turning a 9 into a 7. *ECCCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

🔑 **Insight:** faithfulness facilitates

- model quality checks (Figure 11).
- state-of-the-art plausibility (Figure 12).

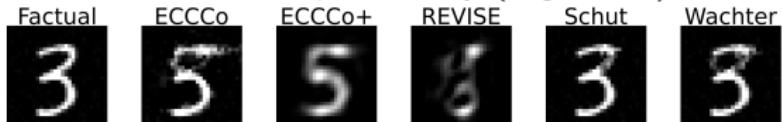


Figure 12: Results for different generators (from 3 to 5).

Background
oooooooooo

Dynamics of CE and AR
ooo

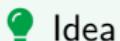
Plausibility at all cost?
oooo

Teaching models plausible explanations
●oo

If we still have time ...
oooo

Teaching models plausible explanations

Counterfactual Training: Method



Let the model compare its own explanations to plausible ones⁴.

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

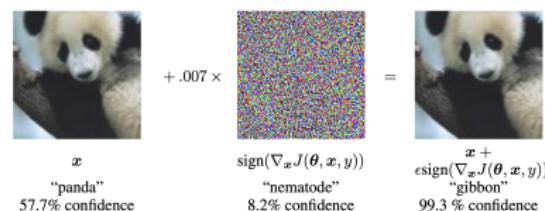


Figure 13: Example of an adversarial attack. Source: Goodfellow, Shlens, and Szegedy (2015)

⁴ under review

Counterfactual Training: Results

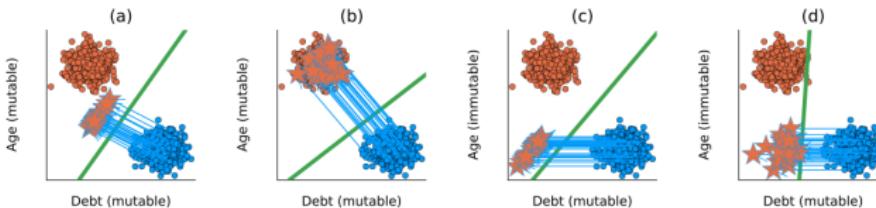


Figure 14: (a) conventional training, all mutable; (b) CT, all mutable; (c) conventional, age immutable; (d) CT, age immutable.

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

Background
oooooooooo

Dynamics of CE and AR
ooo

Plausibility at all cost?
oooo

Teaching models plausible explanations
ooo

If we still have time ...
●ooo

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 AGI⁵.

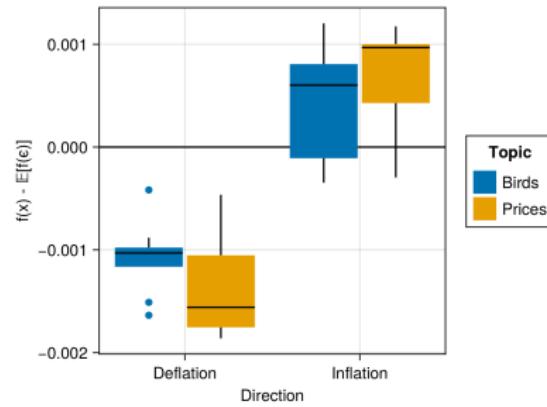


Figure 15: Inflation of prices or birds? It doesn't matter!



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 16: 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, 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. 2023. "Faithful Model Explanations Through