Endogenous Macrodynamics in Algorithmic Recourse

1st IEEE Conference on Secure and Trustworthy Machine Learning

Patrick Altmeyer (p.altmeyer@tudelft.nl)
Giovan Angela (g.j.a.angela@student.tudelft.nl)
Aleksander Buszydlik (a.j.buszydlik@student.tudelft.nl)
Karol Dobiczek (k.t.dobiczek@student.tudelft.nl)
Arie van Deursen (arie.vandeursen@tudelft.nl)
Cynthia C. S. Liem (c.c.s.liem@tudelft.nl)

In a nutshell …

"[…] we run experiments that simulate the application of recourse in practice using various state-of-the-art counterfactual generators and find that [they] induce substantial domain and model shifts."

— Altmeyer et. al (2023)

Proof-of-Concept

Figure 1 illustrates the what we understand as Endogenous Macrodynamics in Algorithmic Recourse:

a) Simple linear classifier trained for binary classification.
b) Implementation of AR leads to a domain shift.
c) Classifier retraining leads to corresponding model shift.
d) Over time decision boundary moves away from target class (blue).

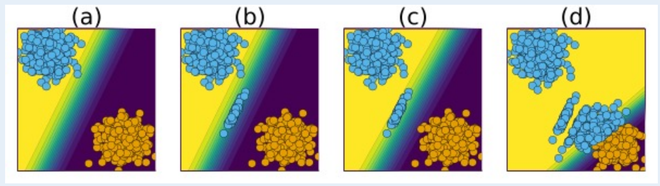


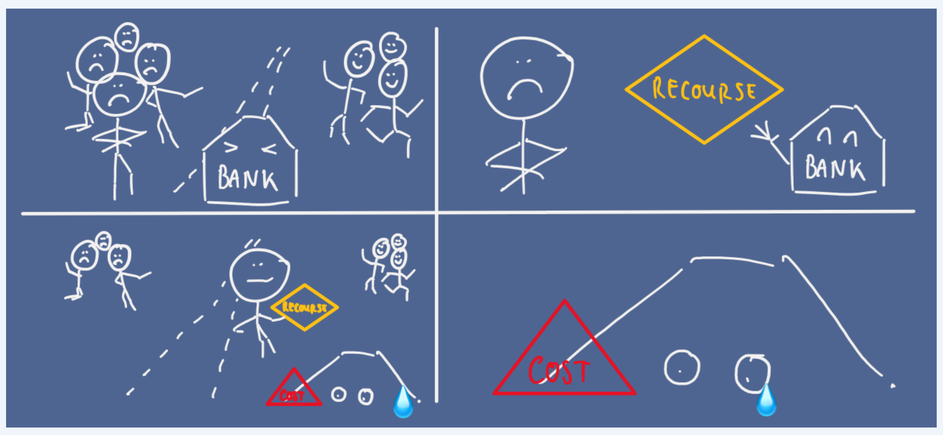
Figure 1: Dynamics in Algorithmic Recourse. Individuals in target class marked in blue.

Key Takeaways

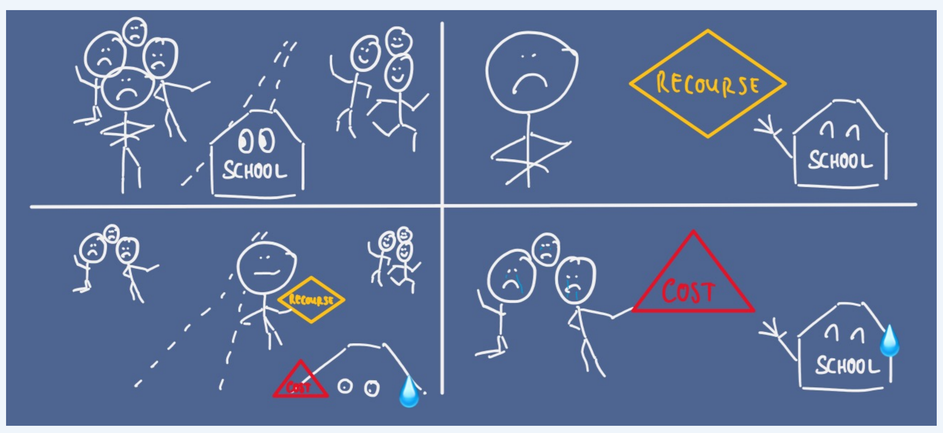
- 🔑 Our findings indicate that state-of-the-art approaches to Algorithmic Recourse induce substantial domain and model shifts.
- 🔑 We would argue that the expected external costs of individual recourse should be shared by all stakeholders.
- 🔑 A straightforward way to achieve this is to penalize external costs in the counterfactual search objective function.
- 🔑 Various simple strategies based on this notion can be effectively used to mitigate shifts.

MOTIVATION

Example 1 (Consumer Credit)



Example 2 (Student Admission)



BACKGROUND

- **Counterfactual Explanation** (CE) explain how inputs into a model need to change for it to produce different outputs.
- Counterfactual Explanations that involve realistic and actionable changes can be used for the purpose of **Algorithmic Recourse** (AR).

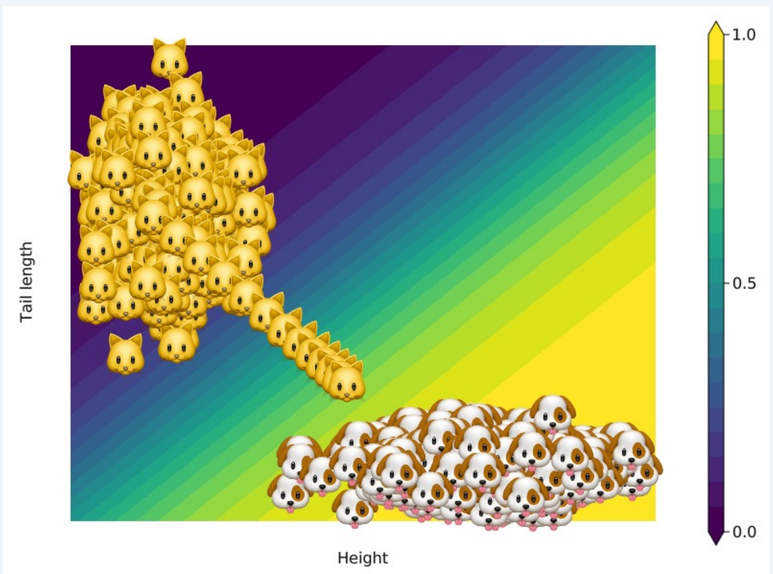


Figure 2: Generating a counterfactual for 🐱 following Wachter et al. (2018). The contour shows the predictions of a simple multi-layer perceptron (MLP).

TOWARDS COLLECTIVE RECOURSE

By introducing a second penalty term, we can explicitly penalize external costs:

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

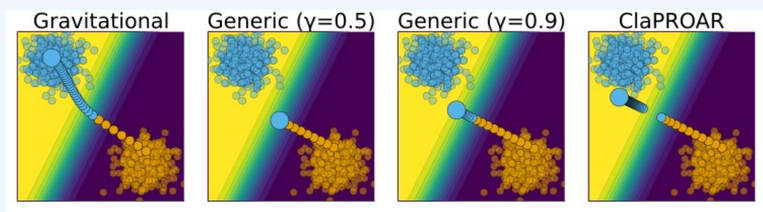


Figure 3: Mitigation strategies compared to the baseline approach, that is, Wachter (Generic) with γ = 0.5; choosing a higher decision threshold pushes the counterfactual a little further into the target domain; this effect is even stronger for ClaPROAR; finally, using the Gravitational generator the counterfactual ends up all the way inside the target domain

MODELLING RECOURSE DYNAMICS

Research Questions

Endogenous Shifts
Does the repeated implementation of recourse provided by state-of-the-art generators lead to shifts in the domain and model?

Costs
If so, are these dynamics substantial enough to be considered costly to stakeholders involved in real-world automated decision-making processes?

Heterogeneity
Do different counterfactual generators yield significantly different outcomes in this context? Furthermore, is there any heterogeneity concerning the chosen classifier and dataset?

Drivers
What are the drivers of endogenous dynamics in Algorithmic Recourse?

Mitigation Strategies
What are potential mitigation strategies with respect to endogenous macrodynamics in AR?

Principal Findings

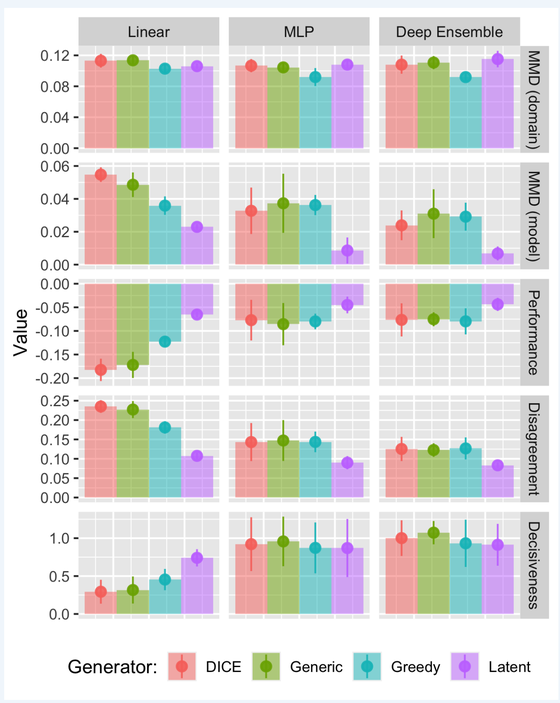


Figure 4: Results for synthetic data.

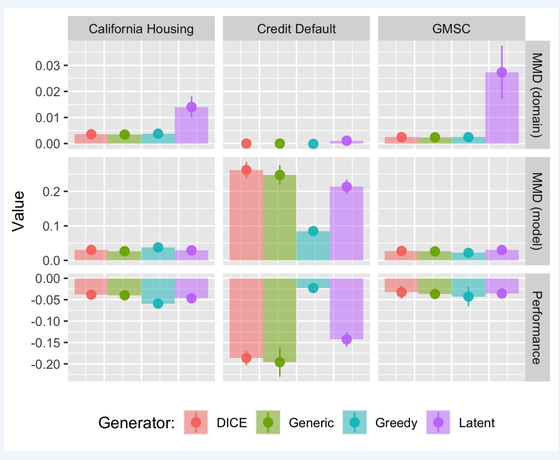


Figure 5: Results for real-world data.

Endogenous Shifts
Costs
Heterogeneity
Drivers
Minimizing private costs vs. complying with data generating process

Secondary Findings

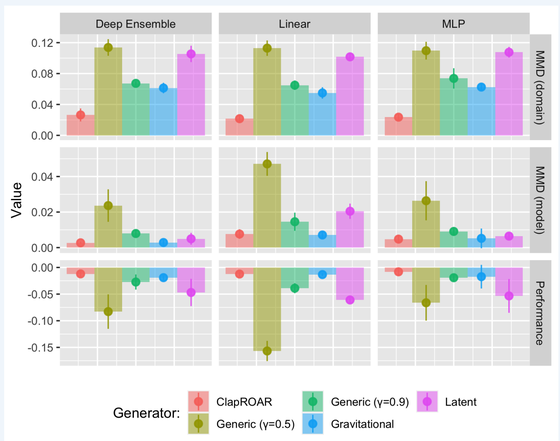


Figure 6: Results for synthetic data using mitigation strategies.

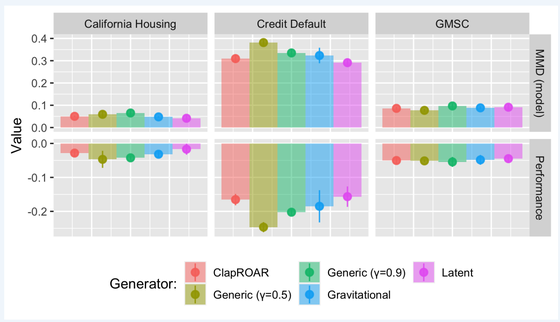


Figure 7: Results for real-world data using mitigation strategies.

Mitigation Strategies
Can effectively mitigate shifts by penalizing external costs

LIMITATIONS & FUTURE WORK

- 🎯 Ad-hoc solution to tradeoff between private vs. external costs.
- 🎯 Experimental design is vast over-simplification of potential real-world scenarios.
- 🎯 We have omitted recourse generators that incorporate causal knowledge.
- 🎯 Analysis limited to differentiable linear and non-linear classifiers, no trees.

RESOURCES