Endogenous Macrodynamics in Algorithmic Recourse

This supporting document presents a proof-of-concept that mimics the application of Algorithmic Recourse (AR) in practice. It illustrates the main observation that our paper is built on: applying AR in practice to groups of individuals may induce substantial data and model shift. We then highlight some of our key findings. Finally, we showcase our proposed mitigation strategies.

1 Proof-of-Concept

Figure 1 illustrates what we define as Endogenous Macrodynamics in Algorithmic Recourse: (a) we have a simple linear classifier trained for binary classification where samples from the negative class ($y = 0$) are marked in orange and samples of the positive class ($y = 1$) are marked in blue; (b) the implementation of AR for a random subset of individuals in the non-target class leads to a noticeable domain shift; (c) as the classifier is retrained we observe a corresponding model shift; (d) as this process is repeated, the decision boundary moves away from the target class.

Example 1.1 (Consumer Credit)**.** Suppose Figure 1 relates to an automated decision-making system used by a retail bank to evaluate credit applicants with respect to their creditworthiness. Assume that the two features are meaningful in the sense that creditworthiness decreases in the South-East direction. Then we can think of the outcome in panel (d) as representing a situation where the bank supplies credit to more borrowers (blue), but these borrowers are on average less creditworthy and more of them can be expected to default on their loan. This represents a cost to the retail bank.

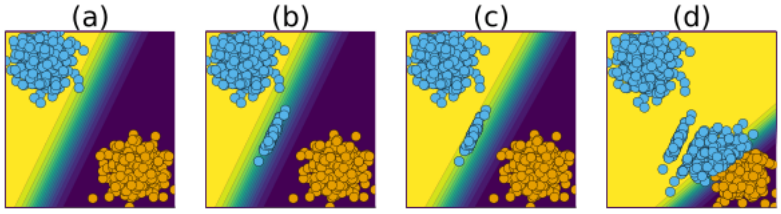

Figure 1: Proof-of-Concept — Applying Algorithmic Recourse in practice. Image by author.

2 Key Findings

Through simulation experiments involving various state-of-the-art counterfactual generators and several benchmark datasets, we generate large numbers of counterfactuals and study the resulting domain and model shifts.

- 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.

As we will touch on briefly below we also find that:

- Various simple strategies based on this notion can be effectively used to mitigate shifts.

3 Proposed Mitigation Strategies

By introducing a second penalty term in the counterfactual search objective, we can explicitly penalize external costs: that is, costs that affect the broad group of stakeholders. Figure 2 illustrates how our mitigation strategies compare to the baseline approach, that is, **Generic** with a decision threshold of 0.5 (Wachter, Mittelstadt, and Russell 2017): choosing a higher decision threshold pushes the counterfactual a little further into the target domain; this effect is even stronger for **ClaPROAR** — our proposed classifier-preserving version of ROAR (Upadhyay, Joshi, and Lakkaraju 2021); finally, using our proposed **Gravitational** generator the counterfactual ends up deep inside the target domain.

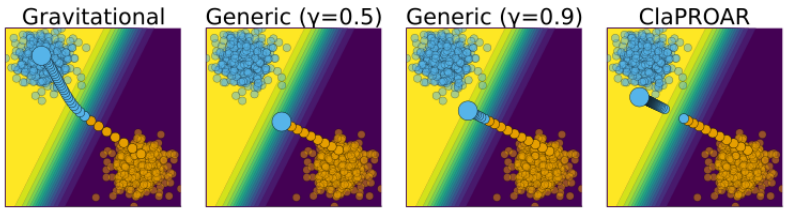

Figure 2: Our proposed mitigation strategies.

4 IEEE SaTML 2023

This work will be presented at the first IEEE Conference on Secure and Trustworthy Machine Learning. You can find out more about our work in this GitHub repository: https://github.com/pat-alt/endogenous-macrodynamics-in-algorithmic-recourse.

References

Upadhyay, Sohini, Shalmali Joshi, and Himabindu Lakkaraju. 2021. "Towards Robust and Reliable Algorithmic Recourse." https://arxiv.org/abs/2102.13620.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." *Harv. JL & Tech.* 31: 841.