# The dynamics induced by Algorithmic Recourse

## 1. Background

Machine learning classifiers have become wildly used by banks, governments, and healthcare institutes [1]. Counterfactual explanation (CFE) was introduced to help explain the decision-making process of the classifier. They provide 'what if' scenarios, counterfactuals (CFs), to achieve a favourable outcome [2]. Algorithmic recourse (AR) provides an actionable set of changes, a CF, that a person can perform to attain the desired outcome.

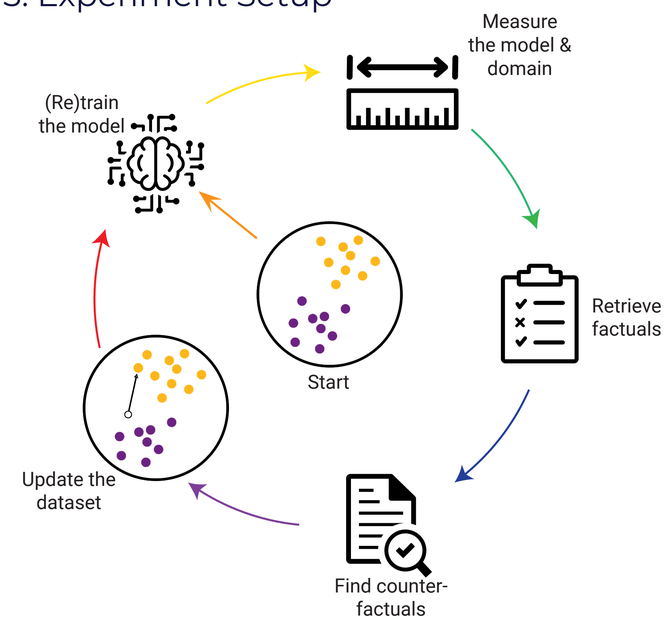
One side effect of AR is the shifts that may occur in the domain and model, called dynamics, when the model is retrained.

The two recourse generators we tested are Wachter et al. and REVISE. Wachter et al. uses an 'optimization approach' that finds counterfactuals that are the closest to the given factuals [3]. REVISE finds counterfactuals that are 'likely to occur under the data distribution' [3]. It does so by using a variational autoencoder (VAE).

## 2. Research Question

- How can we quantify the dynamics of REVISE?
  - Does the magnitude of induced dynamics differ compared to the baseline generator?
  - What factors might be playing a role here?
  - What appear to be good ways to mitigate the dynamics?
  - How can we quantify the dynamics?

## 3. Experiment Setup




### What is the maximum mean discrepancy?

The maximum mean discrepancy (MMD) is a multivariate two-sample test proposed by Gretton et al. [4] to determine if two samples originate from different distributions. An unbiased estimator is given by:

$$MMD^2(X, Y) = E[\kappa(X, X)] + E[\kappa(Y, Y)] - 2E[\kappa(X, Y)]$$

### What is the disagreement?

Sum the number of times two models differ in classification and divide this sum by the size of the dataset.

## 4. Results

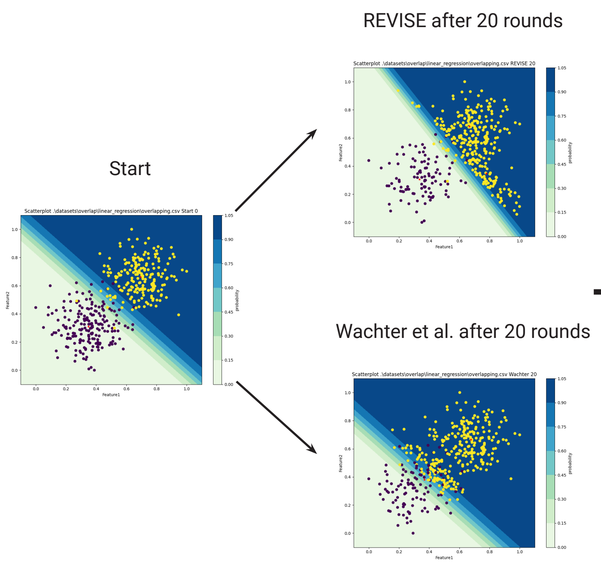



Figure 1: The scatterplot of the 'overlapping' dataset (n=400) at the start and at the end of Wachter et al. and REVISE.

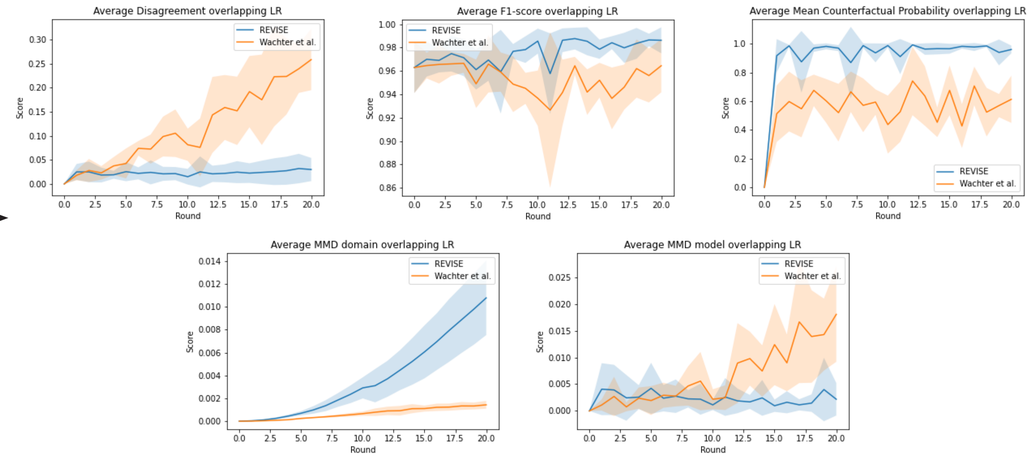



Figure 2: The average F1-score, disagreement, MMD domain, and MMD model of Wachter et al. and REVISE on the 'overlapping' dataset using a logistic regression model.

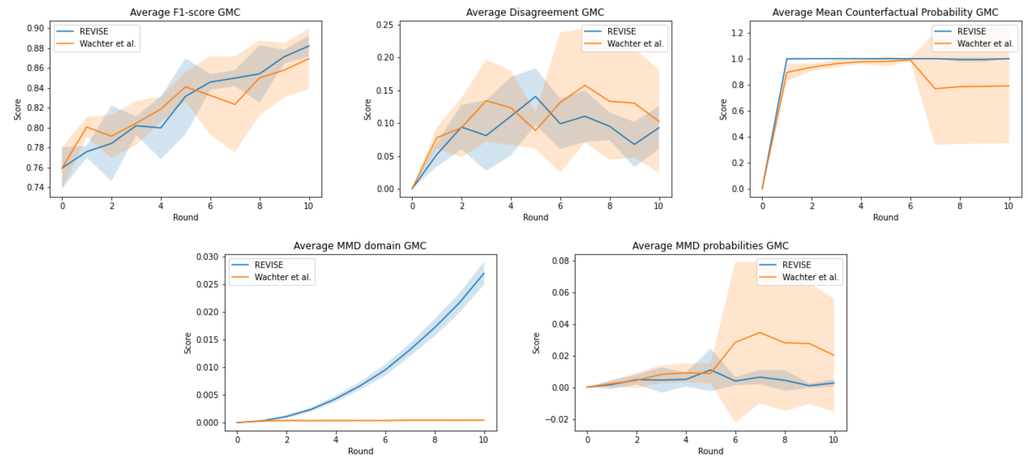



Figure 3: The average F1-score, disagreement, mean counterfactual probability, MMD of the domain, and MMD of the probabilities for Wachter et al. and REVISE on a sub-sampled real-life dataset 'give me some credit' (n=3000).

## 5. Conclusion

REVISE induces increasing shifts in the domain, while the shift in the model remains small. Wachter et al. performs better in reducing the domain shift, while performing worse with model shifts. The difference in implementation of both generators may explain why this is the case. The MMD seems an adequate metric for quantifying the dynamics.

## 6. Improvements

First, we suggest doing more testing with larger real-life datasets.

Secondly, we would like to find out if and how the VAE influences the dynamics induced by REVISE.

Moreover, see if there is a difference in the results when using predicted factuals vs actual factuals.

## References:


[1] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," 2019.
[2] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," Harvard Journal of Law & Technology, 2018, 11 2017.
[3] M. Pawelczyk, S. Bielawski, J. v. d. Heuvel, T. Richter, and G. Kasneci, "Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms," 2021.
[4] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," Journal of Machine Learning Research, vol. 13, no. 25, pp. 723–773, 2012.


Icons made by Creaticca Creative Agency from flaticon.

## Contact information:


Giovan Angela
g.j.a.angela@student.tudelft.nl

Responsible Professor:
Cynthia Liem
C.C.S.Liem@tudelft.nl

Supervisor:
Patrick Altmeyer
P.Altmeyer@tudelft.nl


TU Delft