

# Dynamics in Algorithmic Recourse

Trustworthy Artificial Intelligence for Finance and Economics

Patrick Altmeyer

Delft University of Technology

Delft, The Netherlands

p.altmeyer@tudelft.nl

## ACM Reference Format:

Patrick Altmeyer. 2022. Dynamics in Algorithmic Recourse: Trustworthy Artificial Intelligence for Finance and Economics. In *Proceedings of Fifth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES '22)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

Recent advances in Artificial Intelligence (AI) have propelled its adoption in domains outside of Computer Science including Healthcare, Bioinformatics and Genetics. In Finance, Economics and other social sciences, applications of AI are still relatively limited. Decision-making in these fields has traditionally been guided by interpretable models that facilitate explanations. Explainability is crucial in this context, since decision-makers are typically held accountable by the public: central banks, for example, are heavily scrutinized for the policies they impose. It is therefore not surprising that practitioners and academics in these fields are reluctant to adopt AI technologies they cannot trust. Deep neural networks, for example, are generally considered as black boxes and therefore not trustworthy in a context that demands explanations. This PhD project is focused on exploring and developing methodologies that improve the trustworthiness of AI and thereby enable its application in Finance and Economics.

The remainder of this extended abstract is structured as follows: Section 2 presents one of the research questions I have investigated during the first months of my PhD: how do counterfactual explanations handle dynamics? Section 3 places this work in the broader context of my research.

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*AIES '22, Oxford, UK,*

© 2022 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . \$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

## 2 DYNAMICS IN ALGORITHMIC RECURSE

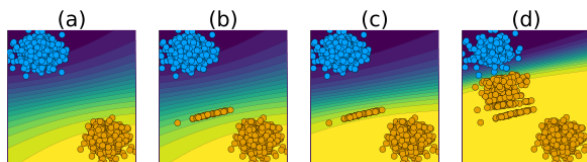
**Counterfactual explanations** (CE) explain how inputs into a model need to change for it to produce different outputs. They are intuitive, simple and intrinsically linked to the potential outcome framework for causal inference, which social scientists are familiar with. Counterfactual explanations that involve realistic and actionable changes can be used for the purpose of **Algorithmic Recourse** (AR) to help individuals who face adverse outcomes. An example relevant to the Finance and Economics domain is consumer credit: in this context AR can be used to guide individuals in improving their creditworthiness, should they have previously been denied access to credit based on an automated decision-making system.

Existing work on CE and AR has largely been limited to the static setting: given some classifier  $M : \mathcal{X} \mapsto \mathcal{Y}$  we are interested in finding close (Wachter, Mittelstadt, and Russell 2017), actionable (Ustun, Spangher, and Liu 2019), realistic Schut et al. (2021), sparse, diverse (Mothilal, Sharma, and Tan 2020) and ideally causally founded counterfactual explanations (Karimi, Schölkopf, and Valera 2021) for some individual  $x$ . The ability of counterfactual explanations to handle dynamics like data and model shifts remains a largely unexplored research challenge at this point (Verma, Dickerson, and Hines 2020). Only one recent work considers the implications of **exogenous** domain and model shifts (Upadhyay, Joshi, and Lakkaraju 2021). The authors propose a simple minimax objective, that minimizes the counterfactual loss function for a maximal model shift. They show that their approach yields more robust counterfactuals in this context than existing approaches.

This project investigates **endogenous** domain and model shifts, that is shifts that occur when AR is actually implemented by a proportion of individuals and the classifier is updated in response. Figure 1 illustrates this idea for a binary problem involving a probabilistic classifier and the counterfactual generator proposed by Wachter, Mittelstadt, and Russell (2017): the implementation of AR for a subset of individuals leads to a domain shift (b), which in turn triggers a model shift (c). As this game of implementing AR and updating the classifier is repeated, the decision boundary moves away from training samples that were originally in the target class (d).

These dynamics may be problematic. As the decision boundary moves in the direction of the non-target class, counterfactual paths become shorter: in the loan example, individuals that previously would have been denied credit based on their input features are suddenly considered as creditworthy. Average default risk across all borrowers can therefore be expected to increase. Conversely, lenders that anticipate such dynamics may choose to deny credit to individuals that have implemented AR, thereby compromising the validity of AR.

To the best of my knowledge this is the first work investigating endogenous dynamics in AR. Through future experiments I want to investigate how this phenomenon plays out across different benchmark datasets including German credit, Boston Housing and COMPAS.<sup>1</sup> Furthermore, I want to assess to what extent the magnitude and direction of domain and model shifts depends on the choice of the counterfactual generator. To this end, I am currently supervising a group of undergraduate students, who are tackling some of these tasks in their final-year research project.



**Figure 1: Dynamics in Algorithmic Recourse:** we have a simple Bayesian model trained for binary classification (a); the implementation of AR for a random subset of individuals leads to a domain shift (b); as the classifier is retrained we observe a model shift (c); as this process is repeated, the decision boundary moves away from the target class (d).

### 3 RELATED AND FUTURE WORK

#### 3.1 Benchmarking CE in Julia

Until recently there existed only one open-source library that provides a unifying approach to generate and benchmark counterfactual explanations for Python models (Pawelczyk et al. 2021). To address this limitation I have developed `CounterfactualExplanations.jl`: a Julia package that can be used to generate counterfactual explanations for models developed and trained not only in Julia, but also in other popular programming languages. The package and companion paper are pending acceptance for a main talk at JuliaCon '22.

#### 3.2 Probabilistic Methods for Realistic CE

To ensure that the generated counterfactuals are realistic it helps to understand which input-output pairs are likely to occur under the data generating process. To this end, previous work has either relied on generative models or restricted the analysis to probabilistic classifiers that incorporate uncertainty in their predictions. While the former approach is more

generally applicable, the latter is computationally more efficient. In future work, I want to explore how recent advances in post-hoc uncertainty quantification, most notably Laplace Redux (Daxberger et al. 2021), can be leveraged to generate realistic and unambiguous counterfactual explanations for any model.<sup>2</sup> With respect to the work-in-progress presented here, I expect that these efforts may help in mitigating endogenous domain and model shifts.

### REFERENCES

- Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. “Laplace Redux-Effortless Bayesian Deep Learning.” *Advances in Neural Information Processing Systems* 34.
- Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjarong, Been Kim, and Joydeep Ghosh. 2019. “Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems.” *arXiv Preprint arXiv:1907.09615*.
- Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. 2021. “Algorithmic Recourse: From Counterfactual Explanations to Interventions.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–62.
- Mothilal, Ramaravind K, Amit Sharma, and Chenhao Tan. 2020. “Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations.” In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–17.
- Pawelczyk, Martin, Sascha Bielawski, Johannes van den Heuvel, Tobias Richter, and Gjergji Kasneci. 2021. “Carla: A Python Library to Benchmark Algorithmic Recourse and Counterfactual Explanation Algorithms.” *arXiv Preprint arXiv:2108.00783*.
- Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. “Generating Interpretable Counterfactual Explanations by Implicit Minimisation of Epistemic and Aleatoric Uncertainties.” In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.
- Upadhyay, Sohini, Shalmali Joshi, and Himabindu Lakkaraju. 2021. “Towards Robust and Reliable Algorithmic Recourse.” *arXiv Preprint arXiv:2102.13620*.
- Ustun, Berk, Alexander Spangher, and Yang Liu. 2019. “Actionable Recourse in Linear Classification.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19.
- Verma, Sahil, John Dickerson, and Keegan Hines. 2020. “Counterfactual Explanations for Machine Learning: A Review.” *arXiv Preprint arXiv:2010.10596*.
- 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.

<sup>1</sup>These benchmark datasets have their issues and controversies, which is one of the challenges I would like to discuss at AIES.

<sup>2</sup>For some initial work on this see my Julia implementation of Laplace Redux: `BayesLaplace.jl`.