Dynamics in Algorithmic Recourse

Trustworthy Artificial Intelligence for Finance and Economics

Personal Details

	Author	Research Advisor
Name	Patrick Altmeyer	Cynthia C. S. Liem
Email	p.altmeyer@tudelft.nl	c.c.s.liem@tudelft.nl
Phone	+4917648726927	+31152782188
Home page	www.paltmeyer.com	www.cynthialiem.com
Affiliation	Delft University of Technology	Delft University of Technology
Affiliation URL	www.tudelft.nl	www.tudelft.nl
Citizenship	German	Dutch
Expected graduation date	2025-09-15	—
Years in program	<1	—

Key words: Counterfactual Explanations, Algorithmic Recourse, Recourse Dynamics, Endogenous Shifts, Explainable AI

Letter of Support

> As substantiated in more detail in my full letter of support (attached with this application), as Patrick's advisor, I am very happy to support his application to the AIES student track. Patrick is an eager student, with a pro-active enthusiasm to learn from others, and contribute himself to the learning progress of relevant communities. Getting him to present at AIES will be very valuable for receiving feedback on his current work, while I would more broadly be happy to receive the community's guidance on how to best navigate questions of ethics and responsibility in a project like his. In case of a positive outcome of this application, we are in a position to fully cover Patrick's costs, and thus will not lay any claim on the financial waivers. — Cynthia C. S. Liem

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.

AIES '22, Oxford, UK,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 DYNAMICS IN ALGORITHMIC RECOURSE

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.[1] 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.

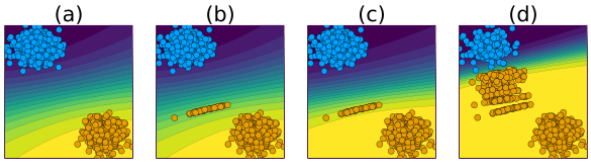

Figure 1: Dynamics in Algorithmic Recourse: we have a simple Bayesian model trained for binary classification (a); the implementation of AR for a random subset of individuals leads to a domain shift (b); as the classifier is retrained we observe a model shift (c); as this process is repeated, the decision 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.[2] With respect to the work-in-progress presented here, I expect that these efforts may help in mitigating endogenous domain and model shifts.

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

[2] For some initial work on this see my Julia implementation of Laplace Redux: BayesLaplace.jl.

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 Vijitbenjaronk, 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.

PATRICK ALTMEYER

Researching Trustworthy Artificial Intelligence (AI) for Finance and Economics

I am an economist by background with an interest in cross-disciplinary research on the intersection of Trustworthy AI and Financial Economics. For my PhD in Trustworthy AI, I currently focus on Counterfactual Explanations and Probabilisitic Machine Learning under supervision of Cynthia Liem at Delft University of Technology.

EDUCATION

2025 | 2021 **PhD in Computer Science**
Delft University of Technology — Delft, Netherlands

Thesis topic: Trustworthy Artificial Intelligence

2021 **Master in Data Science**
Barcelona School of Economics — Barcelona, Spain

Thesis: Deep Vector Autoregression for Macroeconomic Data

2018 **Master in Finance**
Barcelona School of Economics — Barcelona, Spain

Thesis: Option Pricing in the Heston Stochastic Volatility Model

2017 **Master of Arts with Honours in Economics**
University of Edinburgh — Edinburgh, United Kingdom

Thesis: Can misguided monetary policy explain the European housing bubble?

PROFESSIONAL EXPERIENCE

2021 | 2018 **Economist**
Bank of England — London, United Kingdom

- Co-author of two staff working papers (upcoming).
- Co-initiated and led app development.
- Briefing work for policy committees.

2017 **Postgraduate Intern**
Bank of England — London, United Kingdom

- Econometric analysis of transaction data set in R.
- Internal presentation of project results.

CONTACT INFO

p.altmeyer@tudelft.nl
www.paltmeyer.com
github.com/pat-alt
+49 176 48726927

For more information, please contact me via email.

SKILLS

Experienced in Machine Learning, Finance, Economics and Monetary Policy.

Highly skilled in Julia, R, Python and Markdown.

Solid knowledge of C++ (Rcpp), SQL, MATLAB, Stata, HTML, CSS, git, docker, AWS and LaTeX.

Last updated on 2022-04-30.

TEACHING EXPERIENCE

2022 **Research project**
Research supervisor for group of students — Delft, Netherlands
- Proposal of final-year research project on Endogenous Dynamics in Algorithmic Recourse.
- Supervision of group of three undergraduate students working on the project.

2021 **Foundations of Data Science Summer School**
Teaching Assistant at Barcelona School of Economics — Barcelona, Spain

2020 | 2019 **Introduction course to R and Git**
Lead Trainer at Analytics Enablement Hub, Bank of England. — London, United Kingdom

2017 | 2016 **Honours Modules in Econometrics**
Teaching assistant at School of Economics, University of Edinburgh — Edinburgh, United Kingdom

SELECTED PUBLICATIONS AND POSTERS

2022 **Yield Curve Sensitivity to Investor Positioning Around Economic Shocks**
Bank of England Staff Working Paper (upcoming) — London, United Kingdom
Altmeyer P., Boneva L., Kinston R., Saha S., Stoja E.

2021 **Deep Vector Autoregression for Macroeconomic Data**
Masters Thesis (selected for publication): [PDF], [GitHub] — Barcelona, Spain
Agustí M., **Altmeyer P.**, Vidal-Quadras Costa I.

2018 **Option Pricing in the Heston Stochastic Volatility Model: an Empirical Evaluation**
Masters Thesis (selected for publication): [PDF] — Barcelona, Spain
Altmeyer P., Grapendal J., Pravosud M., Quintana G.

CONFERENCES AND WORKSHOPS

2022 **IFC and Bank of Italy workshop on "Data science in central banking"**
Presentation of Altmeyer, Agusti, and Vidal-Quadras Costa (2021): [url], [YouTube] — Virtual

2021 **NeurIPS 2021 MLECON Workshop**
Poster presentation of Altmeyer, Agusti, and Vidal-Quadras Costa (2021): [url] — Virtual

2021 **IFABS 2021 Oxford**
Presented our BoE Staff Working Paper on yield curve pricing — Virtual

2019 **Money markets and Central Bank Balance Sheets**
Presented research on demand for central bank reserves at ECB: [url] — Frankfurt, Germany

SELECTED OPEN-SOURCE SOFTWARE

2021-2022 **`CounterfactualExplanations.jl`**
Julia package for Counterfactual Explanations: [stable docs], [GitHub]

2021-2022 **`BayesLaplace.jl`**
Pure-play Julia package for effortless Bayesian Deep Learning: [dev docs], [GitHub]

2021-2022

`deepvars`
R package implementing Deep Vector Autoregression (Altmeyer, Agusti, and Vidal-Quadras Costa 2021): [GitHub]

OUTREACH AND VOLUNTEERING

2022 | 2021

Personal blog
Presenting AI in an accessible manner: [url]

2020

Class representative
Masters in Data Science — Barcelona, Spain

2016

TEDx talk
Held a TEDx talk about European Integration: [YouTube] — Edinburgh, United Kingdom

SCHOLARSHIPS AND AWARDS

2020

Novartis Datathon
3rd Price Winner of Datathon — Barcelona, Spain

2020

Fee Waiver and Funding for Masters
Full funding for Masters in Data Science through BSE and Bank of England — Barcelona, Spain

2017

Fee waiver for Masters
Total tuition fee waiver for Master in Finance through BSE — Barcelona, Spain

2017

School of Economics Prize
Edinburgh University School of Economics Joint Prize for the best performance in Economics
Edinburgh, United Kingdom

2015

School of Economics Prize
School of Economics Prize for academic excellence in Economics — Edinburgh, United Kingdom

Personal Statement

As a first-year PhD student in Computer Science with a background in Finance and Economics, I was very happy to see that AIES offers a student track. I have worked on several projects during the first months of my PhD and while the one I hope to present at AIES is the most mature research project at this point, it is still very much a work-in-progress. I think that the question I pose about the dynamics in Algorithmic Recourse is an interesting, albeit still fairly broad and open one. Through my participation in AIES, I would hope to get feedback that helps me refine the research question. As experiments related to the question are now being carried out by a group of students under my supervision, the work can be expected to progress further until the actual AIES event. I therefore expect that I can share more experimental results that may be of interest to the broader community. AIES would also be an opportunity to discuss ideas for potential remedies to the issues this work highlights.

I also think that my broader PhD research agenda on the intersection of trustworthy AI and the social sciences should be of interest to the AIES community. Thanks to my academic background in Economics, Finance and Data Science as well my previous professional experience in Monetary Policy, I believe that I can contribute insights to a wide range of discussions revolving around the diverse set of topics relevant to AIES. It would be immensely helpful to learn from more experienced colleagues as well as fellow early-stage researchers, who are working on related research questions. At the main conference and the student event I expect to be exposed to new research ideas that I have not thought of myself, but that may well be highly relevant to my own research. I would even hope to potentially find opportunities for future collaborations.

This would be the first student program I have attended, although it is worth mentioning that in the past I have presented research at the European Central Bank (2019), Bank of England, IFABS (2021), NeurIPS (2021) and the Irving Fisher Committee on Central Bank Statistics (2022). More informations on this can be found in my resume. Since the main PhD research project that I would like to dicuss is still in its early stages, I have not submitted anything to the main track of AIES. I would hope that following my participation in the student track I can use the feedback to carry the work forward and submit it to the main track next year.

Date April 28, 2022

Contact person Dr C.C.S Liem MMus
Phone +31 15 275 2188
E-mail c.c.s.liem@tudelft.nl
Subject Advisor support letter for Patrick Altmeyer

Delft University of Technology

Faculty of Electrical Engineering,
Mathematics and Computer Science
Intelligent Systems / Multimedia Computing

Visit address
Van Mourik Broekmanweg 6
2628 XE Delft
Post address
P.O. Box 5031
2600 GA Delft

To whom it may concern,

With this letter, I am very happy to support the intended participation of Patrick Altmeyer to the AIES Conference. Patrick is a PhD candidate under my supervision, working on Trustworthy AI in the context of the AI for Fintech lab, which is sponsored by the ING Bank in The Netherlands.

A position like this requires appropriate sensitivity to the impact of AI technology on humans, and an openness to learning from insights from neighboring fields. I hired Patrick because of his multidisciplinary background, and great eagerness in making impact across disciplines. While he only started his position in Fall 2021, I have been really proud to already witness this eagerness in many forms, where I really see Patrick consciously investing in serving the broader community with his work (e.g. accessibly writing about his progress in the blog on his website, explicitly open-sourcing his own work in the Julia community, and currently taking multiple undergraduate students under his wing in their first research project).

I would love for Patrick to engage in a deeper discussion about his work with the AIES audience; both with regard to our suspicions on potentially problematic dynamic effects of algorithmic recourse, but also at a more general level. We are aware that applying AI to data involving financial decisions and human information needs great care. We also know that the common benchmark datasets used in machine learning approaches to fairness questions come with controversies of their own. Finally, while this project is sponsored by a bank, we retain our independence as researchers at a public institute, and seek to navigate how to best position ourselves in this space. In all this, I am sure the AIES colleagues can give valuable practical guidance on how to push our work forward responsibly.

In case of an accepted presentation, I am allowed to cover Patrick's costs for travel, registration and housing, and thus would be happy to not claim any waivers, so they can benefit those who need it more.

Thank you for your consideration, sincerely yours,

Dr Cynthia C. S. Liem MMus
Associate Professor
Multimedia Computing Group