Intended for submission to the *Fifth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*

Dynamics in Algorithmic Recourse

Patrick Altmeyer

Recent advances in artificial intelligence (AI) have propelled its adoption in domains outside of computer science including health care, 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 Generalized Linear Models (GLM), which are theoretically founded, interpretable and often sufficient to model relationships between variables. Model interpretability is crucial in the social sciences context, because inference is typically at least as important as predictive performance. Decision-makers in the social sciences are also typically required to explain their decisions to human stakeholders: central bankers, for example, are held accountable by the public for the policies they decide on. It is therefore not surprising that practitioners and academics in these fields are reluctant to adopt AI technologies that ultimately cannot be trusted. Deep learning models, for example, are generally considered as black boxes and therefore difficult to apply in a context that demands explanations.

In my research I explore and develop methodologies that improve the trustworthiness of AI. I would like to understand how we can unlock the enormous potential of AI without sacrificing the human aspect of decision-making in finance and economics. My work so far has focused primarily on counterfactual explanations, algorithmic recourse and probabilistic machine learning. Counterfactual explanations are intuitive, largely model-agnostic and straight-forward to implement. They are also intrinsically linked to the potential outcome framework for causal inference and therefore should be somewhat familiar to social scientists. Counterfactual explanations that involve realistic and actionable changes can be used for the purpose of algorithmic recourse to help individuals facing adverse decisions. Probabilistic machine learning can be leveraged in this context and more generally facilitates inference and interpretability. It is also closely related to Bayesian statistics, which has played an important role in both finance and economics for many years.

In the following (Section 1), I will first briefly present one particular research question I have explored during the first months of my PhD: how do counterfactual explanations handle dynamics? I will also briefly present related projects I have worked on (Section 2) and ideas for future projects (Section 3).

1 DYNAMICS IN ALGORITHMIC RECOURSE

Existing work on counterfactual explanations and algorithmic recourse has largely been limited to the following 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), plausible Schut et al. (2021), sparse (Schut et al. 2021), 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 shifts on the validity of recourse (Upadhyay, Joshi, and Lakkaraju 2021). The authors propose a simple minimax objective, that minimizes the counterfactual loss function for a maximal domain and model shift. They show that their approach yields more robust counterfactuals than existing approaches. In my project I investigate **endogenous** domain and model shifts, i.e. shifts that occur as algorithmic recourse is actually impemented by a proportion of individuals. Preliminary findings indicate that individuals who receive and implement algorithmic recourse end up forming a distinct subgroup inside the target class, which may leave them vulnerable to discrimination (Figure 1). This is a work-in-progress that I would like to present and discuss at AIES.

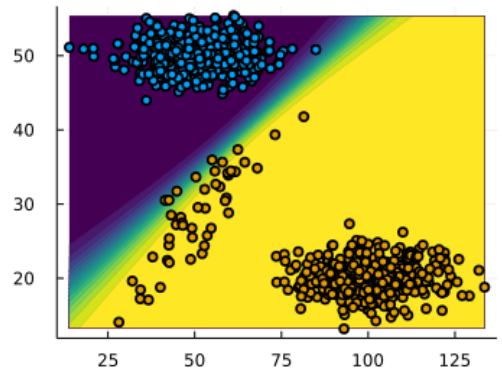

Figure 1: PLACEHOLDER: The dynamics of algorithmic recourse.

2 RELATED PROJECTS

Alongside my research I have developed open-source implementations related to explainable AI. CounterfactualExplanations.jl is 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 like Python and R. I have recently submitted the package along with a companion paper as a proposal for a main talk at JuliaCon. BayesLaplace.jl is a small Julia package that can be used to recover Bayesian representations of deep neural networks through Laplace approximation in a post-hoc manner. It is inspired by a recent paper (Daxberger et al. 2021) and has also been submitted to JuliaCon. Finally, deepvars is an R package that implements an approach towards vector autoregression that leverages deep learning. This was originally my master's thesis and later presented at the NeurIPS 2021 MLECON workshop. I have also published several blog posts on explainable AI and probabilisitic ML in an effort to make my research accessible to a broad audience.

3 FUTURE PROJECTS

Data sets in finance and economics typically involve time series data. Therefore, I am naturally interested in the application of explainable AI to sequential data, an area which has so far not been explored extensively. In the future, I want to work on counterfactual explanations for time series models. I am also interested in seeing if and how Laplace approximation can be used for Bayesian deep learning with time series data. I hope that the findings from both of these projects can ultimately be used to build complex but interpretable time series models for classification and forecasting in finance and economics. For example, I would like to leverage effortless Bayesian deep learning to make our proposed Deep Vector Autoregression model explainable.

REFERENCES

Antorán, Javier, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. 2020. "Getting a Clue: A Method for Explaining Uncertainty Estimates." *arXiv Preprint arXiv:2006.06848*.

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.

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.