Endogenous Dynamics in Algorithmic Recourse

Patrick Altmeyer
Delft University of Technology
EEMCS
Delft, Netherlands
p.altmeyer[at]tudelft.nl

Cynthia C. S. Liem
Delft University of Technology
EEMCS
Delft, Netherlands
c.c.s.liem[at]tudelft.nl

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Existing work on Counterfactual Explanations (CE) and Algorithmic Recourse (AR) has largely been limited to the static setting: given some classifier we are interested in finding close, actionable, realistic, sparse, diverse and ideally causally founded counterfactuals. The ability of CE to handle dynamics like data and model drift remains a largely unexplored research challenge at this point. Only one recent work considers the implications of exogenous domain and model shifts. This project instead focuses on endogenous dynamics, that is shifts that occur when AR is actually implemented by a proportion of individuals. Early findings suggest that the involved shifts may be large with important implications on the validity of AR and the overall characteristics of the sample population.**

I. Introduction

Recent advances in Artificial Intelligence (AI) have propelled its adoption in scientific domains outside of Computer Science including Healthcare, Bioinformatics, Genetics and the Social Sciences. While this has in many cases brought benefits in terms of efficiency, state-of-the-art models like Deep Neural Networks (DNN) have also given rise a new type of principal-agent problem in the context of data-driven decision-making. It involves a group of **principals** - i.e. human stakeholders - that fail to understand the behaviour of their **agent** - i.e. the model used for automated decision-making [1].

Models or algorithms that fall into this category are typically referred to **black-box** models. Despite their shortcomings, black-box models have grown in popularity in recent years and have at times created undesirable societal outcomes [2]. The scientific community has tackled this issue from two different angles: while some have appealed for a strict focus on inherently iterpretable models [3], others have investigated different ways to explain the behaviour of black-box models. These two sub-domains can be broadly referred to as **interpretable AI** and **explainable AI** (XAI), respectively.

Among the approaches to XAI that have recently grown in popularity are **Counterfactual Explanations** (CE). They explain how inputs into a model need to change for it to produce different outputs. 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 Social Sciences 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 in this field has largely worked in a static setting: various approaches have been proposed to generate counterfactuals for a given individual that is subject to some pre-trained model. More recent work has compared different approaches within this static setting [4]. In this work we go one step further and ask ourselves: what happens if recourse is provided and implemented repeatedly? What types of dynamics are introduced and how do different counterfactual generators compare in this context?

Figure 1 illustrates this idea for a binary problem involving a probabilistic classifier and the counterfactual generator proposed by [5]: 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). We refer to these types of dynamics as **endogenous** because they are induced by the implementation of recourse itself.

We find reasons to believe that these types of endogenous dynamics may be problematic. In Figure 1, as the decision boundary moves in the direction of the non-target class, counterfactual paths become shorter: in the consumer

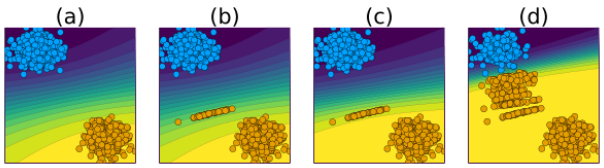

Figure 1: Dynamics in Algorithmic Recourse: we have a simple Bayesian model trained for binary classification (a); the implementation of AR for a random subset of individuals leads to a domain shift (b); as the classifier is retrained we observe a model shift (c); as this process is repeated, the decision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

credit 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 our knowledge this is the first work investigating endogenous dynamics in AR. Our contributions to the state of knowledge are two-fold. Firstly, we introduce an experimental framework extending previous work by [4], which allows for benchmarking different counterfactual generators in terms of the endogenous domain and model shifts they introduce. To this end we propose a number of novel evaluation metrics. Secondly, we use this framework to provide the first in-depth analysis of endogenous recourse dynamics induced by various popular counterfactual generators including [5], [6], [7] and [8].

The paper is structured as follows: Section II … .

II. Related Work

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 [5], actionable [9], realistic [10], sparse, diverse [7] and ideally causally founded counterfactual explanations [11] 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 [12]. Only one recent work considers the implications of **exogenous** domain and model shifts [13]. 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.

III. Methodology

IV. Experiments

V. Discussion

Acknoledgment

P. A. thanks …

References

[1] C. Borch, "Machine learning, knowledge risk, and principal-agent problems in automated trading," *Technology in Society*, p. 101852, 2022.

[2] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[4] M. Pawelczyk, S. Bielawski, J. van den Heuvel, T. Richter, and G. Kasneci, "Carla: A python library to benchmark algorithmic recourse and counterfactual explanation algorithms," *arXiv preprint arXiv:2108.00783*, 2021.

[5] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[6] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[7] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 607–617.

[8] J. Antorán, U. Bhatt, T. Adel, A. Weller, and J. M. Hernández-Lobato, "Getting a clue: A method for explaining uncertainty estimates," *arXiv preprint arXiv:2006.06848*, 2020.

[9] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 10–19.

[10] L. Schut *et al.*, "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties," in *International conference on artificial intelligence and statistics*, 2021, pp. 1756–1764.

[11] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 353–362.

[12] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[13] S. Upadhyay, S. Joshi, and H. Lakkaraju, "Towards robust and reliable algorithmic recourse," *arXiv preprint arXiv:2102.13620*, 2021.

VI. Tables

...

...

...