Endogenous Dynamics in Algorithmic Recourse

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

Giovan Angela
EEMCS
Delft University of Technology
Delft, Netherlands
g.j.a.angela[at]student.tudelft.nl

Aleksander Buszydlik
EEMCS
Delft University of Technology
Delft, Netherlands
a.j.buszydlik[at]student.tudelft.nl

Karol Dobiczek
EEMCS
Delft University of Technology
Delft, Netherlands
k.t.dobiczek[at]student.tudelft.nl

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

***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. A meaningful recourse recommendation for a denied applicant could be: *"If your net savings rate had been 10% of your monthly income instead of the actual 8%, your application would have been successful. See if you can temporarily cut down on consumption."* In the remainder of this paper we will use both terminologies - recourse and counterfactual - interchangeably to refer to situations where counterfactuals are generated with the intent to provide individual recourse.

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 **??** 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.

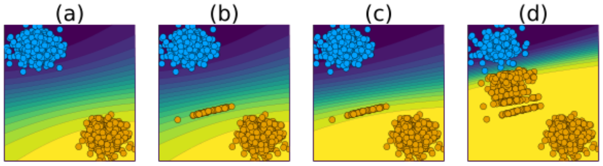

Figure 1: Dynamics in Algorithmic Recourse: we have a simple Bayesian model trained for binary classification (a); the implementation of AR for a random subset of individuals leads to a domain shift (b); as the classifier is retrained we observe a model shift (c); as this process is repeated, the decision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

We think that these types of endogenous dynamics may be problematic. Firstly, model shifts may inadvertently change classification outcomes for individuals who never received and implemented recourse. Secondly and relatedly, we observe in Figure **??** that as the decision boundary moves in the direction of the non-target class, counterfactual paths become shorter: in the consumer 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].

We find that ...

The paper is structured as follows: Section **??** places our work in the broader context of related literature. Section **??** presents our methodology and data. Section **??** presents our empirical findings which are then discussed in Section **??**. We also point to some of the limitations or our work as well as avenues for future research in Section **??**. Finally, Section **??** concludes.

II. Related Work

In this Section we provide a review of the relevant literature. First, Section **??** discusses the existing research within the domain of counterfactual explanations and algorithmic recourse. Then, Section **??** presents some of the previous work on the measurement of dataset and model shifts.

A. Algorithmic Recourse

A framework for Counterfactual Explanations was first proposed in 2017 by [5] and has served as the baseline for most methodologies that have been proposed since then. Let $M : \mathcal{X} \mapsto \mathcal{Y}$ denote some pre-trained model that maps from inputs $X \in \mathcal{X}$ to outputs $Y \in \mathcal{Y}$. Then we are interested in minimizing the complexity or effort $H = h(x')$ associated with moving an individual x to a counterfactual state x' such that the predicted outcome $M(x')$ corresponds to some target outcome t:

$$\min_{x' \in x} c(x') \quad \text{s. t.} \quad M(x') = t \tag{1}$$

For implementation purposes, Equation **??** is typically approximated through regularization:

$$x' = \arg\min_{x'} \ell(M(x'), t) + \lambda h(x') \tag{2}$$

In the baseline work and many subsequent approaches the complexity function $h : \mathcal{X} \mapsto \mathbb{R}$ is proxied by some distance metric based on the simple intuition that large perturbations of x are costly.

Many approaches for the generation of algorithmic recourse have been described in the literature. An October 2020 survey by Karimi et al. layed out 60 algorithms that have been proposed since 2014 [9]. Another survey published around the same time by Verma et al. described 29 algorithms [10]. Different approaches vary primarily in terms of the objective functions they impose, how they optimize said objective (from brute force through gradient-based approaches to graph traversal algorithms), and how the ensure that certain requirements for CE are met. Regarding the latter, the literature has produced an extensive list of desiderata each addressing different needs. To name but a few, we are interested in generating counterfactuals that close [5], actionable [11], realistic [12], sparse, diverse [7] and if possible causally founded [13].

Efforts so far have largely been directed at improving the quality of counterfactual explanations within a static context: given some pre-trained classifier $M : \mathcal{X} \mapsto \mathcal{Y}$ we are interested in generating one or multiple meaningful counterfactual explanations for some individual characterized by x_i. The ability of counterfactual explanations to handle dynamics like data and model shifts remains a largely unexplored research challenge at this point [10]. We have been able to identify only one recent work that considers the implications of **exogenous** domain and model shifts in the context of AR [14]. Exogenous shifts are strictly of external origin. For example, they might stem from data correction, temporal shifts or geospatial changes [14]. The authors of [14] propose framework for algorithmic recourse (ROAR) that evidently improves robustness to such exogenous shifts.

B. Domain and Model Shifts

Much attention has been paid to the detection of dataset shifts – situations where the distribution of data changes over time. Rabanser et al. suggest a framework to detect data drift from a minimal number of samples through the application of two-sample tests [15]. This task is a generalization of the anomaly detection problem for large datasets, which aims to answer the question if two sets of samples could have been generated from the same probability distribution. Numerous approaches to anomaly detection have been summarized [16]. Another well-established research topic is that of concept drift – situations where external variables influence the patterns between the input and the output of a model [17]. For instance, Gama et al. offer a review of the adaptive learning techniques which can handle concept drift [18]. Less previous work is available on the related topic of model drift - changes in model performance over time. Nelson et al. review how resistant different machine learning models are to the model drift [19]. Ackerman et al. offer a method to detect changes in model performance when ground truth is not available [20].

In the context of algorithmic recourse, domain and model shifts were first brought up by the authors behind ROAR [14]. In their work they refer to model shifts as simply any perturbation Δ to the parameters of the model in question: M. While this also sets the baseline for our analysis here, it is worth noting that in [14] these perturbations are mechanically introduced. In contrast we are interested in quantifying model shifts that arise endogenously as part of a dynamic recourse process. In addition to quantifying the magnitude of shifts Δ, we aim to also analyse the characteristics of changes to the model, such as the position of the decision boundary and the overall decisiveness of the model. We have not been able to identify previous work on this topic.

C. Benchmarking Counterfactual Generators

Despite the large and growing number of approaches to counterfactual search there have been surprisingly few benchmark studies that compare different methodologies. This may be partially due to limited software availability in this space. Recent work has started to address this gap: firstly, [21] run a large benchmarking study using different algorithmic aproaches and numerous tabular datasets; secondly, [4] introduce a Python framework that can be used to apply and benchmark different methodology; finally, [22] provides an extensible and language-agnostic implementation in Julia. For the experiments conducted in this work we relied on the latter.

III. Methodology

In the following we first set out a generalized framework for gradient-based counterfactual search in Section **??** to introduce the various counterfactual generators we have chosen to use in our experiments. We then describe the experimental setup in Section **??** and introduce several evaluation metrics used to benchmark the different generators.

A. A Generalized Framework for Gradient-Based Counterfactual Search

In this work we have chosen to focus on a number of gradient-based counterfactual generators to investigate the endogenous dynamics we introduced in Section **??**. Gradient-based counterfactual search is well-suited for differentiable black-box models like deep neural networks. We can restate Equation **??** in a more general form that encompasses most gradient-based approaches to counterfactual search:

$$\mathbf{s}^\prime = \arg \min_{\mathbf{s}^\prime \in \mathcal{S}} \left\{ \sum_{k=1}^{K} \ell(M(f(s^\prime_k)),t) + \lambda h(f(s^\prime_k)) \right\} \tag{3}$$

Here $\mathbf{s}^\prime = \{s^\prime_k\}_K$ is the stacked K-dimensional array of counterfactual states and $f : \mathcal{S} \mapsto \mathcal{X}$ maps from the counterfactual state space to the feature space. In the case of the baseline counterfactual generator [5] f is just the idendity function and the number of counterfactuals K is equal to one. This generator, which we shall refer to as **Wachter** in the following, shall serve as the baseline against which all other gradient-based methodologies will be compared. In particular, we include include the following generator in our benchmarking exercises: REVISE [6], CLUE [8], DICE [7] and a greedy approach that relies on probabilistic models [12].

Both **REVISE** and **CLUE** search counterfactuals in some latent embedding $S \subset \mathcal{S}$ instead of the feature space directly. The latent embedding is learned by a separate generative model that is tasked with learning the data generating process (DGP) of X. In this case f in Equation **??** corresponds to the decoder part of the generative model, in other words the deterministic function that maps back from the latent embedding to the feature space. Provided the generative model is well-specified, traversing the latent embedding typically results in realistic and plausible counterfactuals, because they are implicitly generated by the (learned) DGP [6]. CLUE distinguishes itself from REVISE and other counterfactual generators in that it aims to minimize the predictive uncertainty of the model in question M. To quantify predictive uncertainty the authors rely on entropy estimates for probabilistic models. The **Greedy** approach proposed by [12] also works with the subclass of models $\tilde{\mathcal{M}} \subset \mathcal{M}$ that can produce predictive uncertainty estimates. The authors