Automated black-box decision-making models are becoming increasingly pervasive in our society, but we cannot still understand or act on their recommendations. For example, if a machine learning model denies me a loan, it is impossible for me to challenge its decision. Counterfactual interventions are a powerful tool which can explain black-box model decisions and enable algorithmic recourse. However, current methods provide interventions without considering the user's preferences. We propose the first human-in-the-loop approach to perform algorithmic recourse by modelling and including users in the optimization process, following the preference elicitation theory. An experimental evaluation of synthetic and real-world datasets shows that a handful of queries allows for achieving a substantial reduction in the cost of interventions with respect to user-independent alternatives.

**Motivation**

**What is Algorithmic Recourse?**

Algorithmic Recourse is the ability to provide "explanations and recommendations to individuals who are unfavourably treated by automated decision-making systems" via counterfactual interventions. It implements the "right to an explanation" defined by Article 22 of the GDPR.

**How do we measure the recourse cost?**

The total cost of an intervention given the user preferences is:

\[ C(I|w) = \sum_{I \in O^{(1)}} P(I|w) C(I|w) \]

In the real world, features are causally related. We use a Structural Causal Model (SCM) to model the (linear) dependencies between features and the cost of an action given the user preferences.

**How do we ask the right questions?**

The Expected Utility of Selection (EUS) gives the maximally informative choice set that maximises the user's expected utility (minimizing the intervention costs). We model the user response model as noiseless or logistic (Bradley-Terry).

**Experiments**

Table 1. (Normalized) Average True Regret Improvement (1 – R/I(W)) when we ask q=1 and q=10 questions under all the response models and choice set sizes. With the minimal choice set (k = 2) and q = 1, we can provide interventions that are, on average, ~40% cheaper than the base model. In bold, we have the best result for each model.

(a) synthetic (b) GiveMeSomeCredit (c) Adult

**License**

This research was partially supported by TAILOR, a project funded by the EU Horizon 2020 research and innovation programme under GA No 952215 and partially supported by the project AI@Trento (FBK-Unitn).