User-Aware Algorithmic Recourse with Preference Elicitation

Giovanni De Toni[1,2], Paolo Viappiani[3], Bruno Lepri[1], Andrea Passerini[2]

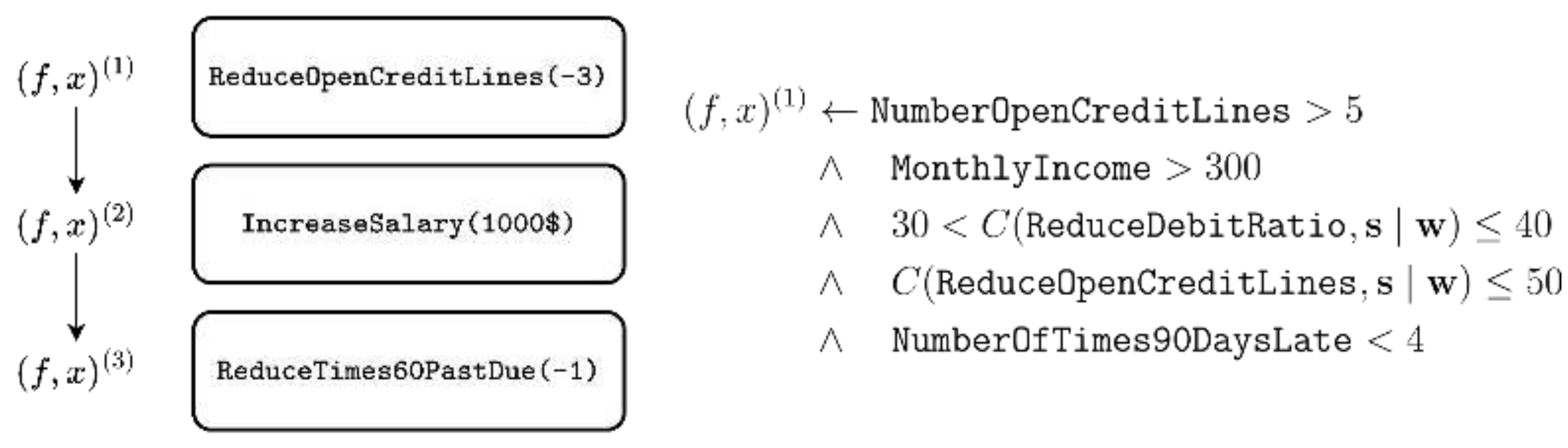
[1] Fondazione Bruno Kessler (FBK), Italy [2] DISI, University of Trento, Italy [3] CNRS and University Paris Dauphine, France

Motivation

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.

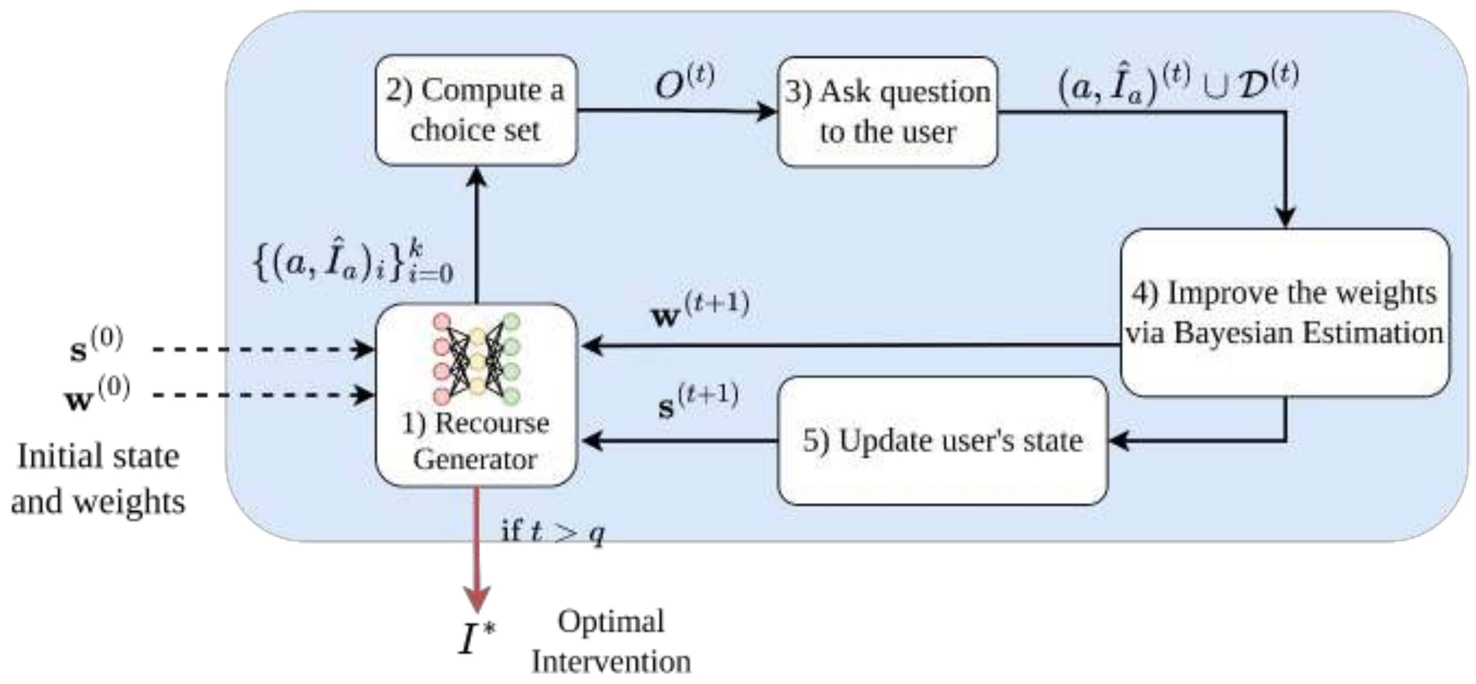
What is Algorithmic Recourse?

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.

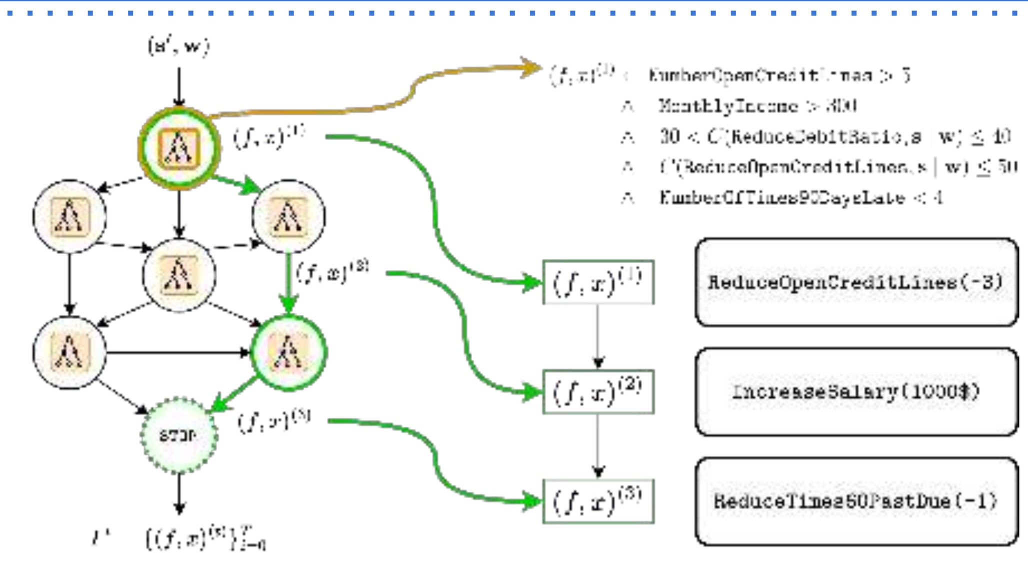


a) Counterfactual intervention (left) and associated explanations (right)

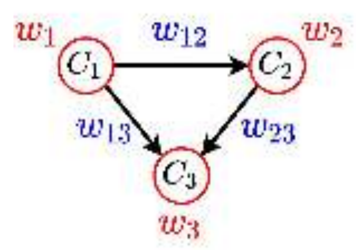
Human-in-the-loop Algorithmic Recourse (WFARE)



User-Aware Explainable Interventions (W-EFARE)



How do we measure the recourse cost?



$$C(a_1, \mathbf{s}|\mathbf{w}) = w_1(s_1' - s_1)$$
$$C(a_2, \mathbf{s}|\mathbf{w}) = w_{12}s_1 + w_2(s_2' - s_2)$$
$$C(a_3, \mathbf{s}|\mathbf{w}) = w_{13}s_1 + w_{23}s_2 + w_3(s_3' - s_3)$$

$$C(I|\mathbf{w}) = \sum_{t=0}^{T} C(a^{(t)}, \mathbf{s}^{(t)}|\mathbf{w})$$

Total cost of an intervention given the user preferences.

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?

$$\mathrm{EUS}_L(O^{(t)}|\mathcal{D}^{(t)}) = -\int_{\mathbf{w}} \left[\sum_{a, \hat{I}_a \in O^{(t)}} P_L(O^{(t)} \rightsquigarrow a|\mathbf{w}) C(\hat{I}_a|\mathbf{w}) \right] P(\mathbf{w}|\mathcal{D}^{(t)})\, d\mathbf{w}$$

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 $P_L(O^{(t)} \rightsquigarrow a|\mathbf{w})$ as **noiseless** or **logistic (Bradley-Terry)**.

How do we discover a successful intervention?

$$\underset{I=(a_1,\ldots,a_k),\ k \leq k_{\max}}{\operatorname{argmin}} \quad C(I, \mathbf{s}|\mathbf{w}) \quad \text{s.t.} \quad p(h(\mathbf{s}') = 1 \mid \mathbf{s}' = I(\mathbf{s})) \geq \tau$$

Given a binary black-box classifier h, we want to find the intervention with the **minimum cost**, which maximizes the probability of achieving a positive classification (e.g., you will get the loan).

Experiments

Table 1. (Normalized) Average True Regret Improvement ($1 - R(\mathbf{I}|\mathbf{W})$) 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 baseline. In **bold**, we have the best result for each model.

(a) synthetic

Model	Noise	$\lvert O^{(t)}\rvert = 2$, $q=1$	$\lvert O^{(t)}\rvert = 2$, $q=10$	$\lvert O^{(t)}\rvert = 4$, $q=1$	$\lvert O^{(t)}\rvert = 4$, $q=10$
W-FARE	R_L	0.33	0.75	0.36	**0.83**
	R_{NL}	0.38	0.79	0.51	**0.83**
W-EFARE	R_L	0.22	0.59	0.34	**0.72**
	R_{NL}	0.37	0.70	0.55	**0.73**

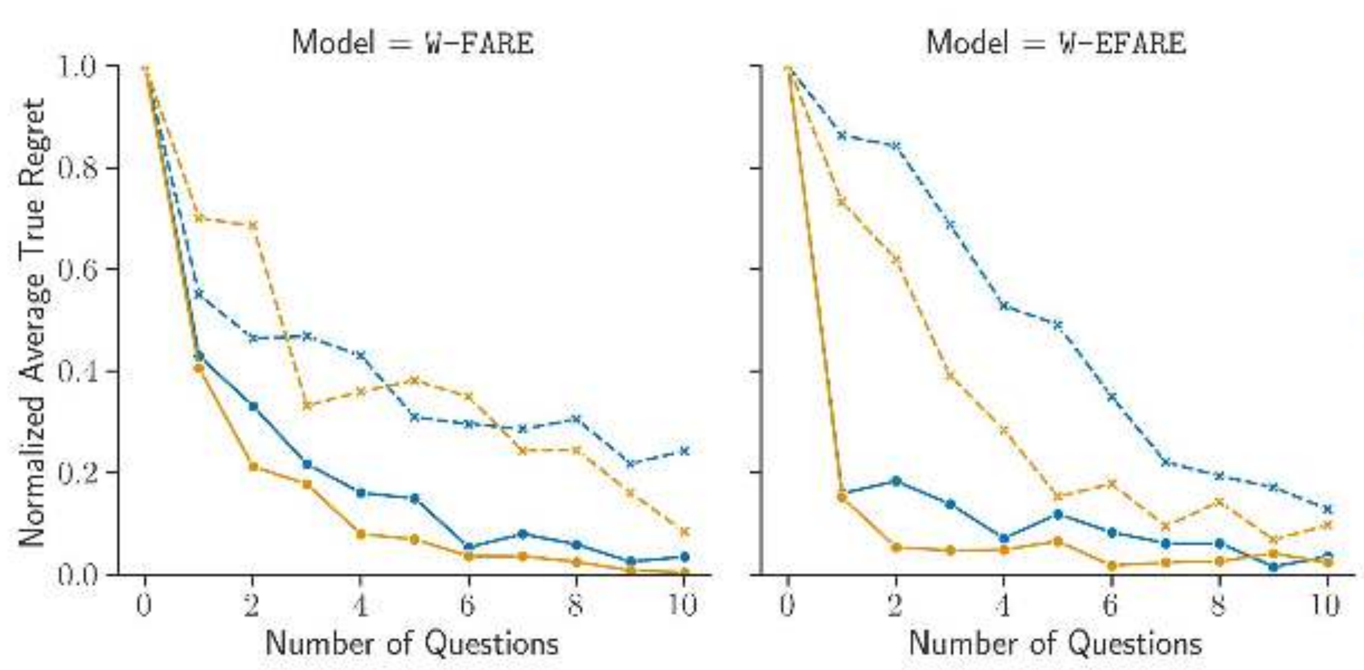
(b) GiveMeSomeCredit

Model	Noise	$\lvert O^{(t)}\rvert = 2$, $q=1$	$\lvert O^{(t)}\rvert = 2$, $q=10$	$\lvert O^{(t)}\rvert = 4$, $q=1$	$\lvert O^{(t)}\rvert = 4$, $q=10$
W-FARE	R_L	0.30	0.80	0.30	**0.86**
	R_{NL}	0.58	0.84	0.71	**0.88**
W-EFARE	R_L	0.37	**0.86**	0.34	0.80
	R_{NL}	0.40	**0.93**	0.66	0.91

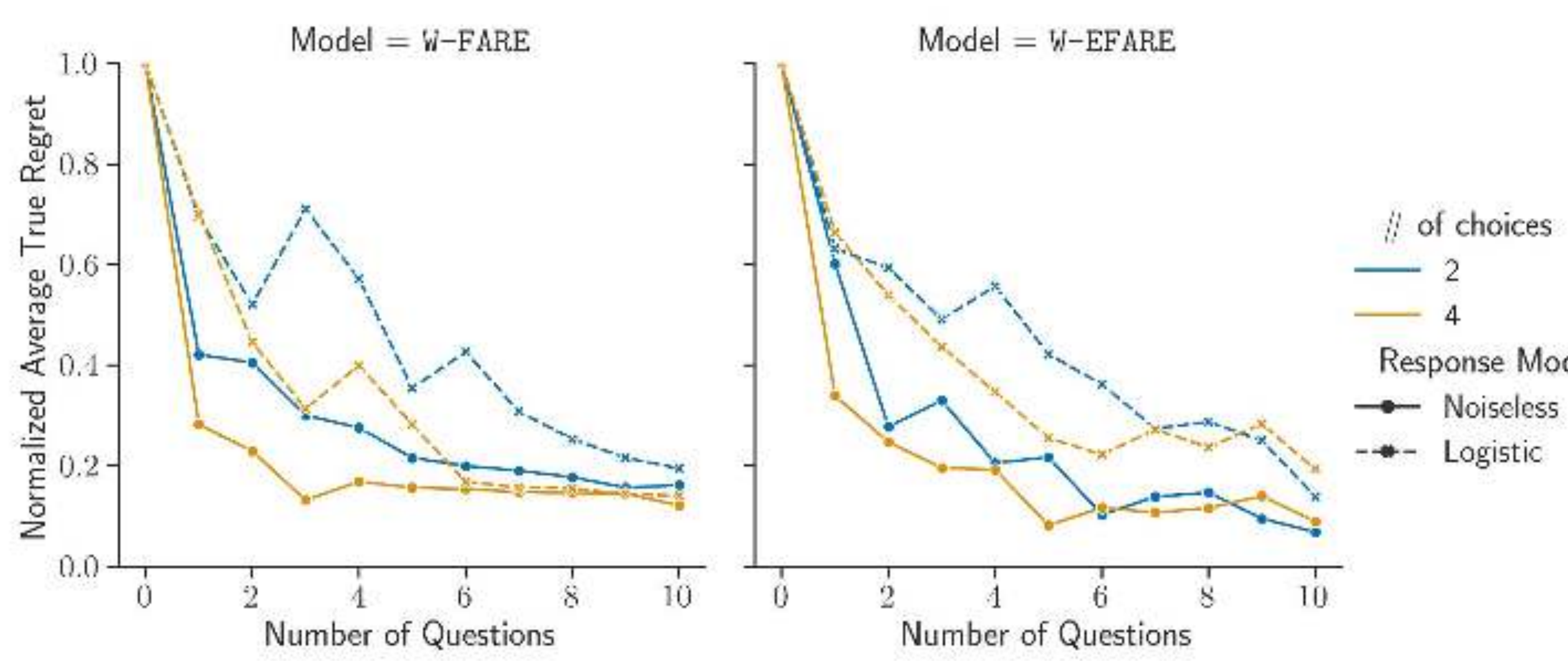
(c) Adult

Model	Noise	$\lvert O^{(t)}\rvert = 2$, $q=1$	$\lvert O^{(t)}\rvert = 2$, $q=10$	$\lvert O^{(t)}\rvert = 4$, $q=1$	$\lvert O^{(t)}\rvert = 4$, $q=10$
W-FARE	R_L	0.45	0.75	0.30	**0.92**
	R_{NL}	0.57	0.96	0.59	**1.00**
W-EFARE	R_L	0.14	0.87	0.26	**0.90**
	R_{NL}	0.84	0.96	0.85	**0.98**

Adult Income



GiveMeSomeCredit



License

Acknowledgements

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).