Performative Validity of Recourse Explanations

Gunnar König[1], **Hidde Fokkema**[2], **Timo Freiesleben**[3], **Celestine Mendler-Dünner**[4], **Ulrike von Luxburg**[1]

[1]University of Tübingen and Tübingen AI Center, [2]University of Amsterdam, [3]LMU Munich, [4]MPI for Intelligent Systems and ELLIS Institute Tübingen

g.koenig.edu@pm.me

Summary

💬 **Recourse**: I was rejected, what can I do to get accepted?

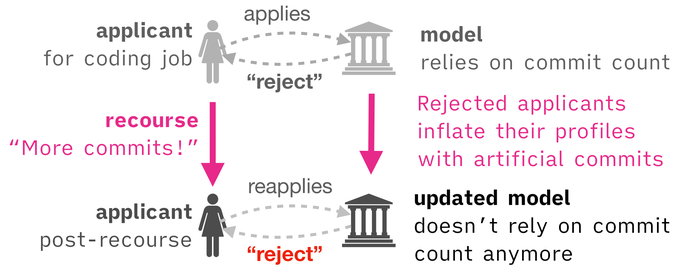
❗ Recourse explanations are *performative*: When people follow them they change the distribution and, once the model is updated, also the decision boundary.

❓ **Do recourse explanations induce a distribution shift that invalidate their own recommendations?**

💡 Recourse may invalidate itself if it is influenced by or if it intervenes on non-causal variables.

⇒ **We advise against using standard recourse methods** (counterfactual explanations, causal recourse) and suggest using improvement-focused recourse instead.

Example (intervention on effect)



💡 commit count is only a symptom/effect of qualification
→ intervening on it breaks the association with qualification

Notation

Basics: features X, target Y, binary label $L := \mathbf{1}[Y \geq t_L]$, decision model $\hat{L}(x) = \mathbf{1}[P(L = 1 \mid X = x)]$

Effect: Variable that is caused by Y.

Recourse: Causal intervention $a(x) := \text{do}(X_{I_a} := \theta_a)$.

Recourse methods: Counterfactual Explanations (**CE**), Causal Recourse (**CR**), Improvement-focused CR (**ICR**)

Performative validity

Definition: Recourse methods are **performatively valid** if recourse that got you accepted by the original model $\hat{L}$ also gets you accepted by the updated model $\hat{L}^m$.

Post-recourse (X^p, Y^p, L^p): State after implementing A, where

$$\text{Action variable } A \;:= \begin{cases} a(x) & \text{if } \hat{L}(x) = 0 \\ \text{do}(\emptyset) & \text{otherwise} \end{cases}.$$

Updated distribution: mix of pre- and post-recourse applicants

$$P(L^m, X^m) = \underbrace{\alpha P(L, X)}_{\text{original applicants}} + \underbrace{(1-\alpha) P(L^p, X^p \mid \hat{L} = 0)}_{\text{post-recourse reapplicants}}$$

Updated model: the optimal predictor in $P(L^m, X^m)$, that is

$$\hat{L}^m(x) = \mathbf{1}[P(L^m = 1 \mid X^m = x) \geq t_c].$$

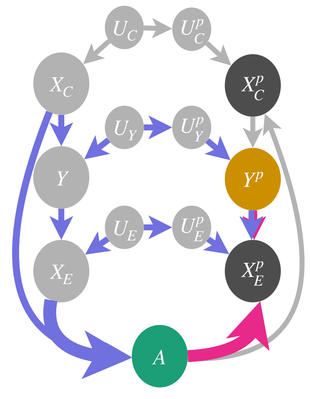
Two Sources of Performative Invalidity

Proposition: Recourse is performatively valid if observing the recommended action does not help in predicting the target in the post-recourse distribution, that is

$$A \perp\!\!\!\perp Y^p \mid X^p.$$

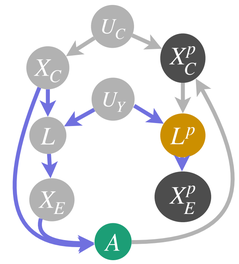
Theorem: There are at most two sources of performative invalidity.

1. Actions A that are influenced by effect variables, that is $X_E \to A$.
2. Actions A that intervene on effect variables, that is $A \to X_E^p$.



⚠️ **CAUTION:** Effect variables (symptoms) can be highly predictive.
→ Decision models and recourse are influenced by effects.
→ Standard recourse methods (CE, CR) may intervene on effects.

Example (influence by effect)



Degree	$X_C \sim \text{Bin}(0.5)$
Autonomy	$U_L \sim \text{Unif}(0,1)$
Qualification	$L := \begin{cases} \mathbf{1}[U_L > 0.55] & X_C = 0 \\ \mathbf{1}[U_L > 0.45] & X_C = 1 \end{cases}$
GitHub activity	$X_E := \begin{cases} L & X_C = 0 \\ 1 & X_C = 1 \end{cases}$

Rejected: applicants with $X = (0,0)$ and thus $u_L < 0.55$
Recourse: $\text{do}(X_C := 1)$
$P(L = 1 \mid X = (1,1)) = 0.55$ but $P(L^p = 1 \mid X^p = (1,1), \hat{L} = 0) \approx 0.22$

💡 Recourse leaks info about autonomy and invalidates recourse.

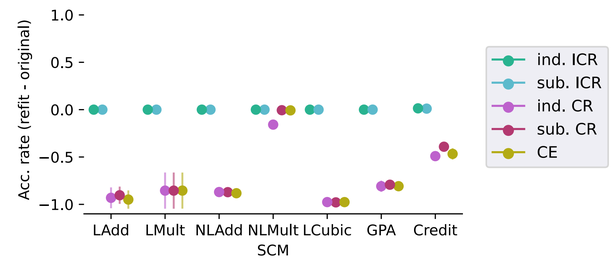
✅ We present assumptions given which the influence by effect variables is not problematic (for example, linear Gaussian data).

Experiments confirm: CE and CR fail

Data: Synthetic and real-world settings with predictive non-causal variables → decision model is influenced by effects

Recourse methods:

- CE and CR (intervene on effects)
- ICR (avoids interventions on effects)



✅ ICR maintains performative validity across all settings.

⚠️ CE and CR are commonly Invalidated.

→ Interventions on effects are particularly problematic.