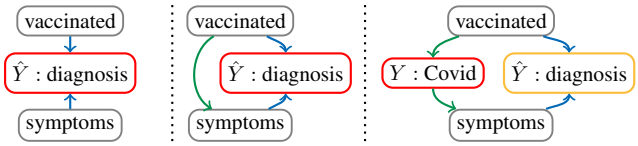
Improvement-Focused Causal Recourse (ICR)

Gunnar König[1,2], **Timo Freiesleben**[4,5,6], **Moritz Grosse-Wentrup**[2,3]
[1] Munich Center for Machine Learning (MCML), LMU Munich
[2] Research Group Neuroinformatics, University of Vienna
[3] Data Science @ Uni Vienna, Vienna CogSciHub
[4] Munich Center for Mathematical Philosophy (MCMP), LMU Munich
[5] Cluster of Excellence Machine Learning, University of Tübingen
[6] Graduate School of Systemic Neurosciences, LMU Munich
g.koenig.edu@pm.me

Abstract

Algorithmic recourse recommendations inform stakeholders of how to act to revert unfavorable decisions. However, existing methods may recommend actions that lead to acceptance (i.e., revert the model's decision) but do not lead to improvement (i.e., may not revert the underlying real-world state). To recommend such actions is to recommend fooling the predictor. We introduce a novel method, Improvement-Focused Causal Recourse (ICR), which involves a conceptual shift: Firstly, we require ICR recommendations to guide toward improvement. Secondly, we do not tailor the recommendations to be accepted by a specific predictor. Instead, we leverage causal knowledge to design decision systems that predict accurately pre- and post-recourse, such that improvement guarantees translate into acceptance guarantees. Curiously, optimal pre-recourse classifiers are robust to ICR actions and thus suitable post-recourse. In semi-synthetic experiments, we demonstrate that given correct causal knowledge ICR, in contrast to existing approaches, guides toward both acceptance and improvement.



counterfactual expl. (CE) | causal recourse (CR) | improvement-focused CR (ICR)

Figure 1: Causal graph illustrating the perspectives of counterfactual expl. (CE, left) and causal recourse (CR, center) in contrast to improvement-focused CR (ICR, right). Green edges represent real-world causal links, and blue edges the prediction model. Gray nodes represent covariates, and the red (yellow) node is the primary (secondary) recourse target. CR respects causal relationships but solely between features; only ICR takes Y into account. While CE and CR aim to revert the prediction $\hat{Y}$, ICR aims to revert the target Y.

1 Introduction

Predictive systems are increasingly deployed for high-stakes decisions, for instance in hiring (Raghavan et al. 2020), judicial systems (Zeng, Ustun, and Rudin 2017), or when distributing medical resources (Obermeyer and Mullainathan 2019). A range of work (Wachter, Mittelstadt, and Russell 2017; Ustun, Spangher, and Liu 2019; Karimi, Schölkopf, and Valera 2021) develops tools that offer individuals possibilities for so-called algorithmic recourse (i.e., actions that revert unfavorable decisions). Joining previous work in the field, we distinguish between reverting the model's prediction $\hat{Y}$ (acceptance) and reverting the underlying real-world state Y (improvement) and argue that recourse should lead to acceptance *and improvement* (Ustun, Spangher, and Liu 2019; Barocas, Selbst, and Raghavan 2020). Existing methods, such as counterfactual explanations (CE; Wachter, Mittelstadt, and Russell (2017)) or causal recourse (CR; Karimi, Schölkopf, and Valera (2021)), ignore the underlying real-world state and only optimize for acceptance. Since ML models are not designed to predict accurately in interventional environments (i.e., environments where actions have changed the data distribution), acceptance does not necessarily imply improvement.

Let us consider an example. We aim to predict whether hospital visitors without test certificate are infected with Covid to restrict access to tested and low-risk individuals. Here, the model's *prediction* $\hat{Y}$ represents whether someone is classified to be infected, whereas the *target* Y represents whether someone is actually infected. Target and prediction differ in how they are affected by actions: Intervening on the *symptoms* may change the model's diagnosis $\hat{Y}$, but will not affect whether someone is infected (Y).

Both counterfactual explanations (CE) and causal recourse (CR) only target $\hat{Y}$ (Figure 1). Therefore, CE and CR may suggest altering the *symptoms* (e.g., by taking cough drops) and thereby may recommend to *game* the predictor: Although the intervention leads to acceptance, the actual Covid risk Y is not improved.[1]

One may argue that this is an issue of the prediction model and may adapt the predictor to make gaming less lucrative than improvement (Miller, Milli, and Hardt 2020). However, such adaptions would come at the cost of predictive performance – even in light of causal knowledge. The reason is that gameable variables can be highly predictive (Shavit, Edelman, and Axelrod 2020); In our example, the model's reliance on the symptom state would need to be reduced. Thus,

[1] In E.1, the case is formally demonstrated.

we tackle the problem by adjusting the explanation instead.

Contributions We present improvement-focused causal recourse (ICR), the first recourse method that targets improvement instead of acceptance. Since estimating the effects of actions is a causal problem, causal knowledge is required. More specifically, we show how to exploit either knowledge of the structural causal model (SCMs) or the causal graph to guide toward improvement (Section 5). On a conceptual level, we argue that the individual's improvement options should not be limited by an acceptance constraint (Section 4). To nevertheless yield acceptance, we show how to exploit said causal knowledge to design post-recourse decision systems that recognize improvement (Section 6), such that improvement guarantees translate into acceptance guarantees (Section 7). On synthetic and semi-synthetic data, we demonstrate that ICR, in contrast to existing approaches, leads to improvement and acceptance (Section 8).

2 Related Work

Constrastive Explanations Contrastive explanations explain decisions by contrasting them with alternative decision scenarios (Karimi et al. 2020a; Stepin et al. 2021); a well-known example are counterfactual explanations (CE) that highlight the minimal feature changes required to revert the decision of a predictor $\hat{f}(x)$ (Wachter, Mittelstadt, and Russell 2017; Dandl et al. 2020). However, CEs are ignorant of causal dependencies in the data and thus, in general, fail to guide action (Karimi, Schölkopf, and Valera 2021). In contrast, the causal recourse (CR) framework by Karimi et al. (2022) takes the causal dependencies between covariates into account: More specifically, Karimi et al. (2022) use structural causal models or causal graphs to guide individuals towards acceptance.[2] The importance of improvement was discussed before (Ustun, Spangher, and Liu 2019; Barocas, Selbst, and Raghavan 2020), but as of now, no improvement-focused recourse method has been proposed.

Strategic Classification The related field of strategic modeling investigates how the prediction mechanism incentivizes rational agents (Hardt et al. 2016; Tsirtsis and Gomez Rodriguez 2020). A range of work (Bechavod et al. 2020; Chen, Wang, and Liu 2020; Miller, Milli, and Hardt 2020) thereby distinguishes models that incentivize *gaming* (i.e., interventions that affect the prediction $\hat{Y}$ but not the underlying target Y in the desired way) and *improvement* (i.e., actions that also yield the desired change in Y). Strategic modeling is concerned with adapting the model, where except for special cases, the following three goals are in conflict: incentivizing improvement, predictive accuracy, and retrieving the true underlying mechanism (Shavit, Edelman, and Axelrod 2020).

3 Background and Notation

Prediction model We assume binary probabilistic predictors and cross-entropy loss, such that the optimal score function $h^*(x)$ models the conditional probability $P(Y = 1|X = x)$, which we abbreviate as $p(y|x)$. We denote the estimated score function as $\hat{h}(x)$, which can be transformed into the binary decision function $\hat{f}(x) := [\hat{h}(x) \geq t]$ via the decision threshold t.

Causal data model We model the data generating process using a structural causal model (SCM) $\mathcal{M} \in \Pi$ (Pearl 2009; Peters, Janzing, and Schölkopf 2017). The model $\mathcal{M} = \langle X, U, \mathbb{F} \rangle$ consists of the endogenous variables $X \in \mathcal{X}$, the mutually independent exogenous variables $U \in \mathcal{U}$, and structural equations $\mathbb{F} : \mathcal{U} \to \mathcal{X}$. Each structural equation f_j specifies how X_j is determined by its endogenous causes and the corresponding exogenous variable U_j. The SCM entails a directed graph $\mathcal{G}$, where variables are connected to their direct effects via a directed edge.

The index set of endogenous variables is denoted as D. The parent indexes of node j are referred to as $pa(j)$, and the children indexes as $ch(j)$. We refer to the respective variables as $X_{pa(j)}$. We write $X_{pa(j)}$ to denote all parents excluding Y and $(X, Y)_{pa(j)}$ to denote all parents including Y. All ascendant indexes of a set S are denoted as $asc(S)$, its complement as $nasc(S)$, all descendant indexes as $d(S)$, and its complement as $nd(S)$.

SCMs allow answering causal questions. This means that they cannot only be used to describe (conditional) distributions (observation, rung 1 on Pearl's ladder of causation (Pearl 2009)) but can also be used to predict the (average) effect of actions $do(x)$ (intervention, rung 2) and imagine the results of alternative actions in light of factual observation $(x, y)^F$ (counterfactuals, rung 3).

As such, we model actions as structural interventions $a : \Pi \to \Pi$, which can be constructed as $do(a) = do(\{X_i := \theta_i\}_{i \in I})$, where I is the index set of features to be intervened upon. A model of the interventional distribution can be obtained by fixing the intervened upon values to θ_I (e.g., by replacing the structural equation $f_I := \theta_I$). Counterfactuals can be computed in three steps (Pearl 2009): First, the factual distribution of exogenous variables U given the factual observation of the endogenous variables x^F is inferred (*abduction*) (i.e., $P(U_j|X^F)$). Second, the structural interventions corresponding to $do(a)$ are performed (*action*). Finally, we can sample from the counterfactual distribution $P(X^{SCF}|X = x^F, do(a))$ using the abducted noise and the intervened-upon structural equations (*prediction*).

4 The Two Tales of Contrastive Explanations

In the introduction, we demonstrated that CE and CR might suggest gaming the predictor (i.e., guide towards acceptance without improvement). To tackle the issue, we will introduce a new explanation technique called improvement-focused causal recourse (ICR) in Section 5.

In this section, we lay the conceptual justification for our method. More specifically, we argue that for recourse, the acceptance constraint of CR should be *replaced* by an improvement constraint. Therefore, we first recall that a multitude of goals may be pursued with contrastive explanations (Wachter, Mittelstadt, and Russell 2017) and separate two purposes of contrastive explanations: *contestability of algorithmic decisions* and *actionable recourse*. We then argue

[2] For the interested reader, we formally introduce CR in our notation in A.4.

that improvement is an essential requirement for recourse and that the individual's options for improvement should not be limited by acceptance constraints.

Contestability and recourse are distinct goals. *Contestability* is concerned with the question of whether the algorithmic decision is correct according to common sense, moral or legal standards. Explanations may help model authorities to detect violations of such standards or enable explainees to contest unfavorable decisions (Wachter, Mittelstadt, and Russell 2017; Freiesleben 2021). Explanations that aim to enable contestability must reflect the model's rationale for an algorithmic decision. *Recourse recommendations*, on the other hand, need to satisfy various constraints unrelated to the model, such as causal links between variables (Karimi, Schölkopf, and Valera 2021) or their actionability (Ustun, Spangher, and Liu 2019). Consequently, explanations geared to contest are more complete and true to the model, while recourse recommendations are more selective and true to the underlying process.[3] We believe that the selectivity and reliance of recourse recommendations on factors besides the model itself is not a limitation but an indispensable condition for making explanations more relevant to the explainee.

In the context of recourse, improvement is desirable for model authority and explainee. We consider improvement an important normative requirement for recourse, both with respect to explainee and model authority. Valuable recourse recommendations enable explainees to plan and act; thus, such recommendations must either provide indefinite validity or a clear expiration date (Wachter, Mittelstadt, and Russell 2017; Barocas, Selbst, and Raghavan 2020; Venkatasubramanian and Alfano 2020). Problematically, when model authorities give guarantees for non-improving recourse, this constitutes a binding commitment to misclassification. However, if model authorities do not provide recourse guarantees over time, this diminishes the value of recourse recommendations to explainees. They might invest effort into non-improving actions that ultimately do not even lead to acceptance because the classifier changed.[4] In contrast, improvement-focused recourse is honored by any accurate classifier. We conclude that, given these advantages for both model authority and explainee, recourse recommendations should help to improve the underlying target Y.[5]

Improvement should come first, acceptance second. Taken that we constrain the optimization on improvement, how to guarantee acceptance remains an open question.
One approach would be to constrain the optimization on both improvement and acceptance. However, a restriction on acceptance is either redundant or, from our moral standpoint, questionable: If improvement implies acceptance, the constraint is redundant; In the remaining cases, we can predict improvement with the available causal knowledge but would withhold these (potentially less costly) improvement options because of the limitations of the observational predictor.
To guarantee acceptance without restricting improvement options, we do not restrict the optimization on acceptance but ensure that the post-recourse predictor can recognize improvements (rendering the acceptance constraint redundant). More specifically, we exploit the assumed causal knowledge to design accurate post-recourse predictors (Section 6) for which acceptance guarantees follow from improvement guarantees (Section 7).

[3]We do not claim that recourse and contestability always diverge; we only describe a difference in focus. If contesting is successful, it may even provide an alternative route toward recourse.

[4]For instance, in the introductory example, an intervention on the symptom state would only be honored by a refit of the model on pre- and post-recourse data for the small percentage of individuals who were already vaccinated, as documented in more detail in E.1. Also, gaming actions may not be robust concerning model multiplicity, as seen in the experiments (Section 8).

[5]We do not claim that gaming is necessarily bad; it may be justified when predictors perform morally questionable tasks.

5 Improvement-Focused Causal Recourse (ICR)

We continue with the formal introduction of ICR, an explanation technique that targets improvement ($Y = 1$) instead of acceptance ($\hat{Y} = 1$). Therefore we first define the improvement confidence γ, which can be optimized to yield ICR. Like previous work in the field (Karimi et al. 2020b), we distinguish two settings: In the first setting, knowledge of the SCM can be assumed, such that we can leverage structural counterfactuals (rung 3 on Pearl's ladder of causation) to introduce the individualized improvement confidence γ^{ind}. In the second setting only the causal graph is known, which we exploit to propose the subpopulation-based improvement confidence γ^{sub} (rung 2).

Individualized improvement confidence For the individualized improvement confidence γ^{ind} we exploit knowledge of a SCM. SCMs can be used to answer counterfactual questions (rung 3). In contrast to rung-2-predictions, counterfactuals are tailored to the individual and their situation (Pearl 2009): They ask what would have been if one had acted differently and thereby exploit the individual's factual observation. Given unchanged circumstances, counterfactuals can be seen as individualized causal effect predictions.
In contrast to existing SCM-based recourse techniques (Karimi et al. 2022) we include both the prediction $\hat{Y}$ and the target variable Y as separate variables in the SCM. As a result, the SCM can be used not only to model the individualized probability of acceptance but also the individualized probability of improvement.

Definition 1 (Individualized improvement confidence)**.** *For pre-recourse observation x^{pre} and action a we define the individualized improvement confidence as*

$$\gamma^{ind}(a) = \gamma(a, x^{pre}) := P(Y^{post} = 1 | do(a), x^{pre}).$$

Since the pre-recourse (factual) target Y cannot be observed, standard counterfactual prediction cannot be applied directly. However, we can regard the distribution as a mixture with two components, one for each possible state of Y. We can estimate the mixing weights using h^* and each component using standard counterfactual prediction. Details, including pseudocode, are provided in B.1.

Subpopulation-based improvement confidence For the estimation of the individualized improvement confidence γ^{ind}, knowledge of the SCM is required. If the SCM is not specified, but the causal graph is known instead, and there are no unobserved confounders (causal sufficiency), we can still estimate the effect of interventions (rung 2).
In contrast to counterfactual distributions (rung 3), interventional distributions describe the whole population and therefore provide limited insight into the effects of actions on specific individuals. Building on Karimi et al. (2020b), we thus narrow the population down to a subpopulation of similar individuals, for which we then estimate the subpopulation-based causal effect. More specifically, we consider individuals to belong to the same subgroup if the variables that are not affected by the intervention take the same values. For action a, we define the subgroup characteristics as $G_a := nd(I_a)$ (i.e., the non-descendants of the intervened-upon variables in the causal graph).[6] More formally, we define the subpopulation-based improvement confidence γ^{sub} as the probability of Y taking the favorable outcome in the subgroup of similar individuals (Definition 2).

Definition 2 (Subpopulation-based improvement confidence)**.** *Let a be an action that potentially affects Y, i.e. $I_a \cap asc(Y) \neq \emptyset$.[7] Then we define the subpopulation-based improvement confidence as*

$$\gamma^{sub}(a) = \gamma(a, x^{pre}_{G_a}) := P(Y^{post} = 1 | do(a), x^{pre}_{G_a}).$$

The set G_a is chosen for practical reasons. To make the estimation more accurate, we would like to condition on as many characteristics as possible. However, without access to the SCM, one can only identify interventional distributions for subgroups of the population by conditioning on their (unobserved) post-intervention characteristics (but not by conditioning on their pre-intervention characteristics) (Pearl 2009; Glymour, Pearl, and Jewell 2016). If we were to select a subgroup from a post-recourse distribution by conditioning on pre-recourse characteristics that are affected by a (e.g., strong pre-recourse symptoms), we yield a group that the individual may not be part of (e.g., people with strong post-recourse symptoms). In contrast, for X_{G_a} pre- and post-intervention values coincide, such that we can estimate γ^{sub}: Assuming causal sufficiency, the standard procedure to sample interventional distributions can be applied, only that additionally $X^{post}_{G_a} := x^{pre}_{G_a}$. Based on the sample, γ^{sub} can be estimated (as detailed in B.3).
The estimation of γ^{sub} does not require knowledge of the SCM but is less accurate than γ^{ind}. In the introductory example, for the action *get vaccinated*, the set of subgroup characteristics G_a is empty. As such, γ^{sub} is concerned with the effect of a vaccination on the whole population. If we were to observe *zip code*, a variable that is not affected by *vaccination*, γ^{sub} would indicate the effect of vaccination for subjects that share the explainee's *zip code*. In contrast, γ^{ind} also takes the explainee's *symptom state* into account.

Optimization problem To generate ICR recommendations, we can optimize Equation 1. We aim to find actions that meet a user-specified improvement target confidence $\overline{\gamma}$ with minimal cost for the recourse seeking individual. The cost function $\text{cost}(a, x^{pre})$ captures the effort the individual requires to perform action a (Karimi et al. 2020b).
As for CE or CR, the optimization problem for ICR is computationally challenging (B.4). It can be seen as a two-level problem, where on the first level the intervention targets I_a, and on the second level the corresponding intervention values θ_a are optimized (Karimi et al. 2020b). Since we target improvement, we can restrict I_a to causes of Y. Following Dandl et al. (2020), we use the genetic algorithm NSGA-II (Deb et al. 2002) for optimization.

$$\text{argmin}_{a=do(X_I=\theta)} \quad \text{cost}(a, x^{pre}) \quad \text{s.t.} \quad \gamma(a) \geq \overline{\gamma}. \quad (1)$$

6 Accurate Post-Recourse Prediction

Recourse recommendations should not only lead to improvement Y but also revert the decision $\hat{Y}$. Whether acceptance guarantees naturally ensue from γ depends on the ability of the predictor to recognize improvements. As follows, we demonstrate how the assumed causal knowledge can be exploited to design accurate post-recourse predictors. We find that an individualized post-recourse predictor is required to translate γ^{ind} into an individualized acceptance guarantee, but curiously that the observational predictor is sufficient in supopulation-based settings.

Individualized post-recourse prediction If we were to use the optimal pre-recourse observational predictor h^* for post-recourse prediction, there would be an imbalance in predictive capability between ML model and individualized ICR: ICR individualizes its predictions using x^{pre} and the SCM. This knowledge is not accessible by the predictor h^*, which only makes use of x^{post}. As such, improvement that was accurately predicted by ICR is not necessarily recognized by h^*, and γ^{ind} cannot be directly translated into an acceptance bound. We demonstrate the issue at an Example in E.3.[8]
To settle the imbalance between ICR and the predictor, we suggest leveraging the SCM not only when generating individualized ICR recommendations but also when predicting post-recourse, such that the predictor is at least as accurate as γ^{ind}. More formally, we suggest estimating the post-recourse distribution of Y conditional on x^{pre}, $do(a)$, and the post-recourse observation $x^{post,a}$ (Definition 3). This post-recourse prediction resembles the counterfactual distribution, except that we additionally take the factual post-recourse observation of the covariates into account.

[6] The estimand resembles the conditional treatment effect with G_a being effect modifiers (Hernán MA 2020).

[7] If a cannot affect Y, we can predict $P(Y|x^{pre}, do(a)) = P(Y|x^{pre})$ using the optimal observational predictor h^*.

[8] One may also argue that standard predictive models are not suitable since optimality of the predictor in the pre-recourse distribution does not necessarily imply optimality in interventional environments (as Example 1, E.1 demonstrates). We can refute this criticism using Proposition 3, where we learn that $\hat{h}^*$ is stable with respect to ICR actions.

Definition 3 (Individualized post-recourse predictor)**.** *We define the individualized post-recourse predictor as*

$$h^{*,ind}(x^{post}) = P(Y^{post} = 1|x^{post}, x^{pre}, do(a))$$

For SCMs with invertible equations, $h^{*,ind}$ can be estimated using a closed form solution. Otherwise, we can sample from the counterfactual post-recourse distribution $p(y^{post}, x^{post}|x^{pre}, do(a))$ (as we did for the estimation of γ^{ind}), select the samples that conform with x^{post} and compute the proportion of favorable outcomes (details in B.2).
For the individualized post-recourse predictor, improvement probability and prediction are closely linked (Proposition 1). More specifically, the expected post-recourse prediction $h^{*,ind}$ is equal to the individualized improvement probability $\gamma(x^{pre}, a)$. We will exploit Proposition 1 in Section 7, where we derive acceptance guarantees for ICR.

Proposition 1. *The expected individualized post-recourse score is equal to the individualized improvement probability $\gamma^{ind}(x^{pre}, a) := P(Y^{post} = 1|x^{pre}, do(a))$, i.e.*

$$E[\hat{h}^{*,ind}(x^{post})|x^{pre}, do(a)] = \gamma^{ind}(a).$$

Subpopulation-based post-recourse prediction Curiously we find that for ICR actions a the optimal observational pre-recourse predictor h^* remains accurate: in the subpopulation of similar individuals, the expected post-recourse prediction corresponds to the improvement probability $\gamma^{sub}(a)$ (Proposition 3). This allows us to derive acceptance guarantees for h^* in Section 7.
This result is in contrast to the negative results for CR, where actions may not affect prediction and the underlying target coherently, such that the predictive performance deteriorates (as demonstrated in the introduction, and more formally in E.1). The key difference to CR is that ICR actions exclusively intervene on causes of Y: Interventions on non-causal variables may lead to a shift in the conditional distribution $P(Y|X_S)$ (where $S \subseteq D$ is any set of variables that allows for optimal prediction). In contrast, given causal sufficiency, the conditional $P(Y|X_S)$ is stable to interventions on causes of Y.

Proposition 2. *Given nonzero cost for all interventions, ICR exclusively suggests actions on causes of Y. Assuming causal sufficiency, for optimal models, the conditional distribution of Y given the variables X_S that the model uses (i.e., $P(Y|X_S)$) is stable w.r.t interventions on causes. Therefore, optimal predictors are intervention stable w.r.t. ICR actions.*

Proposition 3. *Given causal sufficiency and positivity[9], for interventions on causes the expected subgroup-wide optimal score h^* is equal to the subgroup-wide improvement probability $\gamma^{sub}(a) := P(Y^{post} = 1|do(a), x^{pre}_{G_a})$, i.e.*

$$E[\hat{h}^*(x^{post})|x^{pre}_{G_a}, do(a)] = \gamma^{sub}(a).$$

[9]Positivity ensures that the post-recourse observation lies within the observational support (Neal 2020), where the model was trained (i.e., $p^{pre}(x^{post}) > 0$)).

Link between CR and ICR: Proposition 2 has further interesting consequences. For CR actions a that only intervene on causes of Y and that are guaranteed to yield a predicted score ζ in the subpopulation, we can infer that $\gamma^{sub}(a) \geq \zeta$. For instance, if acceptance with respect to a 0.5 decision threshold can be guaranteed, that implies improvement with at least 50% probability. As such, in subpopulation-based settings (1) improvement guarantees can be made for CR if only interventions on causes are lucrative, and (2) CR can be adapted to also guide towards improvement by restricting actions to intervene on causes.

7 Acceptance Guarantees

For the presented accurate post-recourse predictors, improvement guarantees translate into acceptance guarantees (Proposition 4). The reason is that the post-recourse prediction is linked to γ (Propositions 1 and 3).

Proposition 4. *Let g be a predictor with $E[g(x^{post})|x^{pre}_S, do(a)] = \gamma(x^{pre}_S, a)$. Then for a decision threshold t the post-recourse acceptance probability $\eta(t; x^{pre}_S, a) := P(g(x^{post}) > t|x^{pre}_S, do(a))$ is lower bounded by the respective improvement probability:*

$$\eta(t; x^{pre}_S, a, g) \geq \frac{\gamma(x^{pre}_S, a) - t}{1 - t}.$$

Proof (sketch): We decompose the expected prediction (γ) into true positive rate (TPR), false negative rate (FNR) and acceptance rate. By bounding TPR and FNR we yield the presented acceptance bound. The proof is provided in D.4.

Using Proposition 4, we can tune confidence γ and the model's decision threshold to yield a desired acceptance rate. For instance, we can guarantee acceptance with (subgroup-wide) probability $\eta \geq 0.9$ given $\gamma = 0.95$ and a global decision threshold $t = 0.5$.
Furthermore, we can leverage the sampling procedures that we use to compute γ to estimate the individualized or subpopulation-based acceptance rate $\eta(t; x^{pre}_S, a, g)$ (as detailed in B.1 and B.3). To guarantee acceptance with certainty, the decision threshold can be set to $t = 0$.
For the explainee, it is vital that the acceptance guarantee is presented in a human-intelligible fashion. In contrast to previous work in the field, we suggest communicating the acceptance guarantee in terms of a probability.[10] Furthermore, for subpopulation-based recourse, the set of subgroup characteristics should be transparent. In the hospital admission example, the subpopulation-based acceptance guarantee could be communicated as follows: *Within a group of individuals that share your zip code, a vaccination leads to acceptance with at least probability η.*

8 Experiments

In the experiments we evaluate the following questions, assuming correct causal knowledge and accurate models of the conditional distributions in the data:

[10]For CR, the acceptance confidence is encoded in a hyperparameter, as explained in E.2.

Q1: Do CE, CR, and ICR lead to improvement?
Q2: Do CE, CR, and ICR lead to acceptance (by pre- and post-recourse predictor)?
Q3: Do CE, CR, and ICR lead to acceptance by other predictors with comparable test error?[11]
Q4: How costly are CE, CR and ICR recommendations?

Setup We evaluate CE, individualized and subpopulation-based CR, and ICR with various confidence levels, over multiple runs, and on multiple synthetic and semi-synthetic datasets with known ground truth (listed below).[12] Random forests were used for prediction, except in the *3var* settings where logistic regression models were used. Following Dandl et al. (2020), we use NSGA-II (Deb et al. 2002) for optimization. For a full specification of the SCMs including the linear cost functions, we refer to C.2. Details on the implementation and access to the code are provided in C.1.

3var-causal: A linear gaussian SCM with binary target Y, where all features are causes of Y.
3var-noncausal: The same setup as *3var-causal*, except that one of the features is an effect of Y.
5var-skill: A categorical semi-synthetic SCM where programming skill level is predicted from causes (e.g. *university degree*) and non-causal indicators extracted from GitHub (e.g. *commit count*).
7var-covid: A semi-synthetic dataset inspired by a real-world covid screening model (Jehi et al. 2020; Wynants et al. 2020).[13] The model includes typical causes like *covid vaccination* or *population density* and symptoms like *fever* and *fatigue*. The variables are mixed categorical and continuous with various noise distributions. Their relationships include nonlinear structural equations.

Results The results are visualized in Figures 3-5 and provided in tabular form in C.3. For each setting CE, CR, and ICR explanations were computed over 10 runs on 200 individuals each. For CR and ICR the confidences $0.75, 0.85, 0.9, 0.95$ were targeted (for CR: $\overline{\eta}$, for ICR: $\overline{\gamma}$). For CE no slack is allowed, such that the results correspond to a confidence level of 1.0. Values are plotted on quadratic scales.

Q1 (Figure 3): In scenarios where gaming is possible and lucrative (*3var-noncausal*, *5var-skill* and *7var-covid*) ICR reliably guides towards improvement, but CE and CR game the predictor and yield improvement rates close to zero. For instance, on *5var-skill* CE and CR exclusively suggest tuning the GitHub profile (e.g. by adding more commits). Since the employer offered recourse it should be honored although the applicants remain unqualified. In contrast, ICR suggests getting a degree or gaining experience, such that recourse

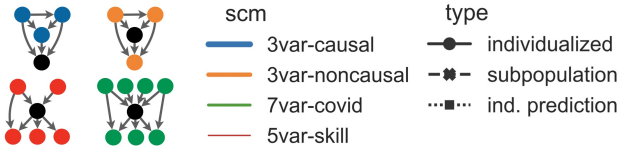


Figure 2: Left: Causal graphs. Right: Legend for color (SCM) and linestyle (recourse type) in Figures 3, 4 and 5.

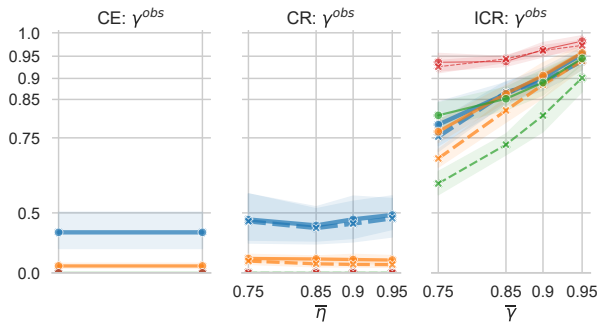


Figure 3: Observed improvement rates γ^{obs} (Q1).

implementing individuals are suited for the job.
On *3var-causal*, where gaming is not possible, CR also achieves improvement. However, since acceptance w.r.t to a decision threshold $t = 0.5$ is targeted, only improvement rates close to 50% are achieved (the expected predicted score translates into γ^{sub} (Proposition 3)).
For subp. ICR, γ^{obs} is below $\overline{\gamma}$, because the subpopulation may include individuals that were already accepted pre-recourse, such that γ^{sub} and γ^{obs} may not coincide.

Q2 (Figure 4): All methods yield the desired acceptance rates w.r.t. to the pre-recourse predictor.[14] For CE and CR η^{obs} is higher than for ICR, and for ind. recourse higher than for subp. recourse. Curiously, although no acceptance guarantees could be derived for the pre-recourse predictor and ind. ICR, we find that both pre- and ind. post-recourse predictor reliably lead to acceptance.[15]

Q3 (Figure 5): We observe that CE and CR actions are unlikely to be honored by other model fits with similar performance on the same data. This result is highly relevant to practitioners since models deployed in real-world scenarios are regularly refitted. As such, individuals that implemented acceptance-focused recourse may not be accepted after all, since the decision model was refitted in the meantime. In contrast, ICR acceptance rates are nearly unaffected by refits. The result confirms our argument that improvement-focused recourse may be more desirable for explainees (Section 4).

Q4 (Table 1): CR actions are cheaper than ICR actions, since improvement may require more effort than gaming. As

[11]The problem that refits on the same data with similar performance have different mechanism is known as the Rashomon problem or model multiplicity (Breiman 2001; Pawelczyk, Broelemann, and Kasneci 2020; Marx, Calmon, and Ustun 2020).

[12]For ground-truth counterfactuals, simulations are necessary (Holland 1986).

[13]The real-world screening model is used to decide whether individuals need a test certificate to enter a hospital. It can be accessed via https://riskcalc.org/COVID19/.

[14]ICR holds the acceptance rates from Proposition 4, as analyzed in more detail in C.3.

[15]Given that the ind. post-recourse predictor is much more difficult to estimate, the pre-recourse predictor in combination with individualized acceptance guarantees (B.1) may cautiously be used as fallback.

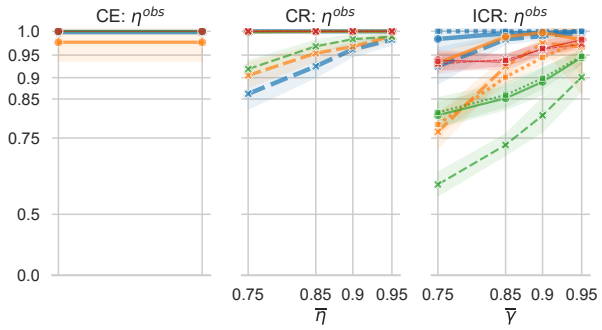


Figure 4: Observed acceptance rates η^{obs} w.r.t. h^*; for ind. ICR additionally w.r.t. $h^{*,ind}$ (Q2).

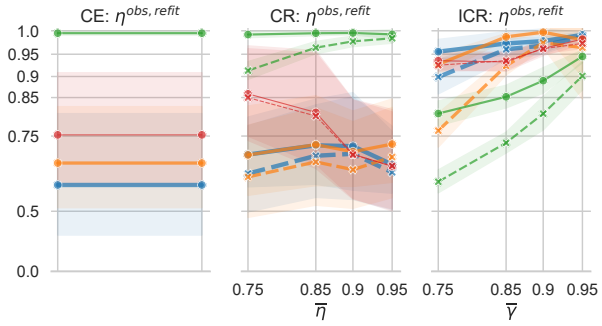


Figure 5: Observed acceptance rates for other fits with comparable test set performance $\eta^{obs,\text{refit}}$ (Q3).

such, CR has benefits for the explainee: For instance, on *5var-skill*, CR suggests tuning the GitHub profile (e.g. by adding more commits), which requires less effort than earning a degree or gaining job experience. Detailed results on cost are reported in C.3.

In conclusion, ICR actions require more effort than CR, but lead to improvement and acceptance while being more robust to refits of the model.

9 Limitations and Discussion

Causal knowledge and assumptions Individualized ICR requires a fully specified SCM; Subpopulation-based ICR is less demanding but still requires the causal graph and causal sufficiency. SCMs and causal graphs are rarely readily available in practice (Peters, Janzing, and Schölkopf 2017) and causal sufficiency is difficult to test (Janzing et al. 2012). Research on causal inference gives reason for cautious optimism that the difficulties in constructing SCMs and causal graphs can eventually be overcome (Spirtes and Zhang 2016; Peters, Janzing, and Schölkopf 2017; Heinze-Deml, Maathuis, and Meinshausen 2018; Malinsky and Danks 2018; Glymour, Zhang, and Spirtes 2019).

There are further foundational problems linked to causality that affect our approach: causal cycles, an ontologically vague target Y (e.g. in hiring), disparities in our data, or causal model misspecification (Barocas and Selbst 2016; Barocas, Hardt, and Narayanan 2017; Bongers et al. 2021). All of these factors are considered difficult open problems and may have detrimental impact on our, as well as on any other, recourse framework.

Guiding action without causal knowledge is impossible; when causal knowledge is available, our work provides a normative framework for improvement-focused recourse recommendations. Thus, we join a range of work in explainability (Frye, Rowat, and Feige 2020; Heskes et al. 2020; Wang, Wiens, and Lundberg 2021; Zhao and Hastie 2021) and fairness (Kilbertus et al. 2017; Kusner et al. 2017; Zhang and Bareinboim 2018; Makhlouf, Zhioua, and Palamidessi 2020) that highlights the importance of causal knowledge.

CE	ind. CR	sub. CR	ind. ICR	sub. ICR
1.8 ± 1.1	1.3 ± 1.1	1.7 ± 1.0	4.3 ± 3.3	4.2 ± 3.3

Table 1: Recourse cost (Q4).

Contestability Improvement-focused recourse guides individuals towards actions that help them to improve, e.g., it recommends a vaccination to lower the risk of getting infected with Covid. If, however, an explainee is more interested in contesting the algorithmic decision, (improvement-focused) recourse recommendations are not sufficient. Think of an individual who is denied entrance to an event because of their high Covid risk prediction, which is based on a non-causal, spurious association with their country of origin[16]. In such situations, we suggest to additionally show explainees diverse explanations, which enable to contest the decision. For example, such an explanation could be: if your country of origin was different, your predicted Covid risk would have been lower.

10 Conclusion

In the present paper, we took a causal perspective and investigated the effect of recourse recommendations on the underlying target variable. We demonstrated that acceptance-focused recourse recommendations like CE or CR might not improve the underlying target but game the predictor instead. The problem stems from predictive but non-causal relationships, which are abundant in ML applications.[17]

We introduced Improvement-Focused Causal Recourse (ICR), an explanation technique that exploits causal knowledge to guide toward improvement. To guarantee acceptance, we ensured that improvements are recognized by the post-recourse predictor: For cases where we individualize the recommendation using knowledge of the SCM, we proposed an individualized post-recourse predictor; In the remaining cases, post-recourse acceptance guarantees hold for any predictor that is accurate pre-recourse. In experiments we support the theoretical advantages of ICR.

With our proposal, we hope to inspire a shift from acceptance- to improvement-focused recourse.

[16] E.g., due to a spurious association with the *type of vaccine*.

[17] E.g. in hiring, some keywords in the CV are predictive, but adding them to the CV does not improve aptitude (Strong 2022).

Acknowledgements

This project is supported by the German Federal Ministry of Education and Research (BMBF), the Carl Zeiss Foundation (project on "Certification and Foundations of Safe Machine Learning Systems in Healthcare") and by the Graduate School of Systemic Neurosciences (GSN) Munich. The authors of this work take full responsibility for its content. We thank the anonymous reviewers for their feedback, which guided us toward improvement (and acceptance).

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Barocas, S.; Selbst, A. D.; and Raghavan, M. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 80–89. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Bechavod, Y.; Ligett, K.; Wu, Z. S.; and Ziani, J. 2020. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024*.

Bongers, S.; Forré, P.; Peters, J.; and Mooij, J. M. 2021. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5): 2885–2915.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–231.

Chen, Y.; Wang, J.; and Liu, Y. 2020. Linear Classifiers that Encourage Constructive Adaptation. *arXiv preprint arXiv:2011.00355*.

Dandl, S.; Molnar, C.; Binder, M.; and Bischl, B. 2020. Multi-Objective Counterfactual Explanations. In Bäck, T.; Preuss, M.; Deutz, A.; Wang, H.; Doerr, C.; Emmerich, M.; and Trautmann, H., eds., *Parallel Problem Solving from Nature – PPSN XVI*, 448–469. Cham: Springer International Publishing. ISBN 978-3-030-58112-1.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2): 182–197.

Freiesleben, T. 2021. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33: 1229–1239.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Heinze-Deml, C.; Maathuis, M. H.; and Meinshausen, N. 2018. Causal structure learning. *Annual Review of Statistics and Its Application*, 5: 371–391.

Hernán MA, R. J. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33: 4778–4789.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Janzing, D.; Sgouritsa, E.; Stegle, O.; Peters, J.; and Schölkopf, B. 2012. Detecting low-complexity unobserved causes. *CoRR*, abs/1202.3737.

Jehi, L.; Ji, X.; Milinovich, A.; Erzurum, S.; Rubin, B. P.; Gordon, S.; Young, J. B.; and Kattan, M. W. 2020. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4): 1364–1375.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020a. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 353–362. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020b. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 265–277. virtual: Curran Associates, Inc.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2022. Towards Causal Algorithmic Recourse. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 139–166. Springer.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.

Malinsky, D.; and Danks, D. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1): e12470.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6917–6926. Online: PMLR.

Neal, B. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*.

Obermeyer, Z.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency*, 89–89.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On Counterfactual Explanations under Predictive Multiplicity. In Peters, J.; and Sontag, D., eds., *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 809–818. Online: PMLR.

Pearl, J. 2009. *Causality*. Cambridge, UK: Cambridge University Press, 2 edition. ISBN 978-0-521-89560-6.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 469–481. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Causal Strategic Linear Regression. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8676–8686. virtual: PMLR.

Spirtes, P.; and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, 1–28. SpringerOpen.

Stepin, I.; Alonso, J. M.; Catala, A.; and Pereira-Fariña, M. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9: 11974–12001.

Strong, J. 2022. MIT Technology Review: Beating the AI hiring machines. https://www.technologyreview.com/2021/08/04/1030513/podcast-beating-the-ai-hiring-machines/. Accessed 2022-07-15.

Tsirtsis, S.; and Gomez Rodriguez, M. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33: 16749–16760.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 10–19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Venkatasubramanian, S.; and Alfano, M. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 284–293. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

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

Wang, J.; Wiens, J.; and Lundberg, S. 2021. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, 721–729. PMLR.

Wynants, L.; Van Calster, B.; Collins, G. S.; Riley, R. D.; Heinze, G.; Schuit, E.; Bonten, M. M.; Dahly, D. L.; Damen, J. A.; Debray, T. P.; et al. 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.

Zeng, J.; Ustun, B.; and Rudin, C. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3): 689–722.

Zhang, J.; and Bareinboim, E. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Issue: 1.

Zhao, Q.; and Hastie, T. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1): 272–281.