Meaningful Causal Recourse (MCR): From Gaming to Improvement

ANONYMOUS AUTHOR(S)*

Algorithmic recourse recommendations, such as Karimi et al.'s (2020) causal recourse (CR), inform stakeholders of how to act to revert unfavorable predictions. We distinguish reverting the prediction of the model, referred to as acceptance, from reverting the underlying real-world state, which we call improvement. We observe that actions which lead to acceptance do not necessarily lead to improvement, but may rather game the predictor. Such gaming recourse recommendations are binding commitments to misclassification. With our method, Meaningful Causal Recourse (MCR), we propose a paradigm shift: Firstly, instead of directly targeting acceptance, MCR is focused on guiding individuals towards improvement. Secondly, instead of tailoring recourse to a specific predictor, we leverage the causal knowledge required for MCR to design decision systems that predict accurately pre- and post-recourse. As a result, the improvement guarantees for MCR can be translated into acceptance guarantees. We justify the detachment of recourse recommendations from the predictor by separating the recourse setting from situations where explanations shall enable individuals to contest incorrect or unfair decisions. On a simulated example we demonstrate that MCR recommendations are not only more meaningful than CR but also more robust to refits of post-recourse data.

CCS Concepts: • **Computing methodologies** → **Causal reasoning and diagnostics**; **Machine learning algorithms**; *Philosophical/theoretical foundations of artificial intelligence.*

Additional Key Words and Phrases: recourse, causality, robustness, explanation

ACM Reference Format:

Anonymous Author(s). 2022. Meaningful Causal Recourse (MCR): From Gaming to Improvement. In *FAccT '22: ACM FAccT Conference 2022, June 21–24, 2022, Seoul, South Korea.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Predictive systems are increasingly deployed in high-stakes environments such as hiring [39], recidivism prediction [53] or loan approval [48]. A range of work develops tools that offer individuals possibilities for so-called algorithmic recourse, i.e. actions that revert unfavorable decisions [5, 21, 22, 51]. In contrast to 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). We argue that recourse should be *meaningful*, i.e. lead to acceptance *and improvement*. Existing methods such as counterfactual explanations [51] or causal recourse [21] 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 a simple motivational example. The goal is to predict the Covid-19 risk of employees in order to restrict office access to low-risk individuals. In the example, the model's prediction $\hat{Y}$ represents whether someone is diagnosed with Covid $\hat{Y}$, whereas the prediction target represents whether someone is actually infected with Covid Y. Even if the model's prediction is accurate in a given test distribution, target and prediction differ in their causal roles and therefore in how they are affected by actions: Whether an individual is *vaccinated* causally influences the *Covid*

Manuscript submitted to ACM

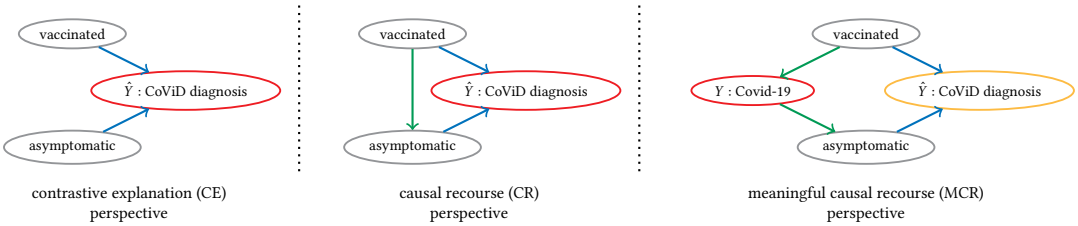

Fig. 1. Directed Acyclic Graph (DAG) illustrating the perspective on model and data taken by counterfactual explanations [51] (left) and causal recourse [22] (center) in contrast to meaningful recourse (right). Blue edges represent the causal relationships induced by the prediction model, green edges the causal relationships induced by the data generating process, gray nodes the input variables and the red (yellow) node the primary (secondary) recourse target. Causal recourse (CR) respects the causal relationships but only between input features. Meaningful causal recourse (MCR) is the only approach that also takes the target Y into account. While CE and CR aim to revert the prediction $\hat{Y}$, MCR aims to revert the underlying target Y. A more detailed description of the model and data generating process is given in Section 4.

risk Y. *Covid* Y causes typical *symptoms* such as dry cough. In contrast, both features are causal for the prediction $\hat{Y}$, since the ML model learns to exploit not only the direct cause but also the associated variable. Both counterfactual explanations (CE) [5, 51] and causal recourse (CR) [22] ignore the underlying target Y and only optimize for acceptance (Figure 1). Therefore, they may suggest to game the predictor by intervening on the non-causal variable *symptom* (e.g., by taking cough drops). This way they change the *Covid prediction* ($\hat{Y}$) without actually lowering the *Covid* risk (Y) and yield acceptance without improvement.

Such non-meaningful recourse does not only fail to guide towards improvement, but also lacks robustness. By robustness we refer to the problem that since implementing recourse actions takes time and decisions systems are regularly updated, the model which decides post-recourse[1] may be different from the pre-recourse model for which the recommendation was designed. As such, the recourse recommendation does not necessarily lead to acceptance. However, for recourse recommendations to be a *robust good* [49], it is vital that they are honored if acted upon [7, 34, 40, 51]. We argue that robustness and non-meaningfulness are related problems, since non-meaningful recourse itself is a cause of distribution shifts that deteriorate the model's performance. The drop in performance triggers refits, which then invalidate the non-meaningful recommendation.

In the motivational example, non-meaningful recourse actions fool the predictor to consider sick individuals who treated their symptoms as healthy. A refitted model would recognize that the association between symptom-state and sickness is reduced and adapt accordingly. As a consequence, individuals who where recommended to fool the predictor (by treating their symptoms) are not accepted by the refitted model. If recourse guarantees were made, the model's decision must be overruled and the symptom-treating individuals must be accepted against better knowledge.[2]

One may argue that those are not issues of the explanation technique, but of the prediction model that should not incentivize gaming (i.e., fooling the predictor to yield acceptance without the cost of improvement) in the first place. As a consequence, one may adapt the predictor strategically to make gaming less lucrative than improvement [32]. In our example, the model's reliance on the symptom-state would have to be reduced. However, gameable variables such as the symptom state are often highly predictive. Therefore, such so-called strategic prediction would come at the cost of predictive performance [42].

[1]Meaning when the individual applies again after implementing recourse, see Table 1 for an overview of important terms.
[2]A more detailed description of the example is given in Section 4.

Thus, we tackle the problem in the explanation domain. We propose Meaningful Causal Recourse (MCR) (Section 6), a method that guides individuals towards improvement of the underlying target Y. Following Karimi et al. [22] we distinguish two settings with different causal knowledge: One where the structural causal model (SCM) can be specified, and one where only the causal graph is known. For the first setting, we introduce an *individualized* variant that uses structural counterfactuals to make individualized predictions of causal effects. For the second setting, we suggest a *subpopulation-based* variant which is based on the computation of average causal effects for subgroups of individuals who are similar to the explainee. For each method we derive accurate post-recourse predictors with acceptance guarantees. The theoretical results are supported with a simulation on the motivational example (Section 7).

In order to justify the suggested detachment of recourse explanations from the model, we disentangle two goals of explanations that should be regarded separately (Section 5). On the one hand, we need explanations that enable individuals to contest algorithmic decisions (contestability). They must be maximally faithful to the model. For instance, in the motivational example, they should expose the model's reliance on the symptom-state. On the other hand, we need explanations that guide individuals to revert unfavorable algorithmic decisions, assuming that the model is trustworthy (recourse), which is the focus of this work. Since recourse recommendations must not explain how the model works they do not have to explain how the model could be gamed, but should rather offer a maximum variety of options for improvement.

Table 1. Overview of important terms and their meanings.

term	meaning
explainee	the individual for which the explanation is generated, e.g. loan applicant
model authority	the decision-making entity, e.g. credit institute
recourse	action of the explainee that reverts unfavorable decision
acceptance	desirable model prediction ($\hat{Y} = 1$), e.g. model predicts *healthy*
improvement	(yield) desirable state of the underlying target ($Y = 1$), e.g. person is *healthy*
gaming	actions that yield acceptance but no improvement, e.g. treating the symptoms
pre-/post-recourse	before/after implementing recourse recommendation
contestability	the explainee's ability to contest an algorithmic decision
robustness of recourse	probability that a recourse action leads to acceptance despite shifts in model and data

2 RELATED WORK AND CONTRIBUTIONS

2.1 Constrastive Explanations

So-called contrastive explantions explain the decision of a predictor $\hat{f}(x)$ by assessing the minimal actions that would revert the model's decision [5, 21, 47, 51]. However, counterfactual explanations are ignorant of causal dependencies in the data and therefore in general fail to guide action [21]. In contrast, the causal recourse (CR) framework by Karimi et al. [21, 22] takes the causal dependencies between covariates into account: Karimi et al. [21] use structural causal models to guide individuals towards acceptance.

Our main contributions are of conceptual nature: In contrast to previous work on recourse, we separate reverting the prediction $\hat{Y}$ (acceptance) from reverting the underlying real-world state Y (improvement). We demonstrate that even though causal recourse (CR) recommendations lead to acceptance by the predictor ($\hat{Y} = 1$), they may fail to improve the underlying target Y (Section 4). We argue that in recourse settings, improvement is beneficial for both model authorities

and explainees and therefore should be the central goal of recourse. Furthermore, improvement options should not be limited by a model's inability to predict accurately, such that we do not constrain on acceptance by an observational predictor (Section 5).

Thus, we introduce *Meaningful Causal Recourse* (MCR) that recommends actions that exclusively focus on the improvement of the prediction target Y (Sections 6). In contrast to CR, MCR does not directly target acceptance. Instead, we demonstrate how to design decision-making systems such that acceptance naturally ensues improvement. Following Karimi et al. [22], we propose both an individualized version (that is more precise but requires knowledge of the structural causal model), and a subpopulation-based variant (that only requires access to the causal graph).

Our technical contributions include theoretical results on the estimation of the structural counterfactual given unobserved variable Y, the estimation of an individualized-post recourse predictor, the derivation of conditions under which optimal observational models predict accurately in post-recourse environments as well as conditions for CR and MCR to coincide. While the acceptance probability for CR is encoded in a non-interpretable hyperparameter, we derive interpretable post-recourse acceptance bounds, thereby enabling the explainee to make an informed choice.

2.2 Robust algorithmic recourse

According to Barocas et al. [2] and Venkatasubramanian and Alfano [49] counterfactual explanations (CE) only provide reliable information about relevant alternative predictions if the model is stable over time. Recourse, on the other side, should be a *robust good*: For individuals who implement recourse recommendations acceptance should be guaranteed even if model and data have shifted. In a similar vein, Wachter et al. [51] suggest guaranteeing counterfactual-based recourse within a pre-specified period of time.

The robustness of counterfactuals [51] and causal recourse [21] has been investigated before [34, 40, 46], yet only with respect to generic shifts of model and data. Rawal et al. [40] and Upadhyay et al. [46] assess the impact of data correction, temporal and geospatial shifts. Upadhyay et al. [46] introduce ROAR, an algorithm inspired by adversarial training that optimizes a novel objective that incentivizes robustness of recourse. Pawelczyk et al. [34] demonstrate that counterfactuals that lie within the support of the observational distribution are more robust to model multiplicity [31]. Dominguez-Olmedo et al. [7], Pawelczyk et al. [35] introduce causal recourse that is robust to uncertainty in the input features.

Inspired by Goodhard's Law [14] and work in strategic prediction (Section 2.3), we are the first to assess the robustness of recourse w.r.t. distribution shifts that were induced by the recourse actions themselves. We argue that MCR recommendations are more robust regarding refits of the model on post-recourse data than CR (Section 6.4) and support the claim on a simulated example (Section 7).

2.3 Strategic Classification

The related field of strategic modeling investigates how the prediction mechanism incentivizes rational agents [16]. Miller et al. [32] thereby distinguish 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). Chen et al. [4] suggest adapting predictors such that incentivizing improvement and predictive accuracy are traded-off. They forgo a causal formalism and assume linear relationships and causally independent covariates. Bechavod et al. [3] demonstrate that strategic agents behaviour in combination with model refits can help to distinguish causal from non-causal features and to learn a causal model that does not incentivize gaming. Tsirtsis and Gomez Rodriguez [45] link the field with contrastive explanations and investigate how to jointly design decision policies and counterfactual

explanations to maximize utility. Strategic classification is not only studied in the context of ML but also in ecnomomics. For instance, Haghtalab et al. [15], Kleinberg and Raghavan [24] discuss the design of mechanisms such that improvement is supported but gaming is not. All the aforementioned methods are concerned with adapting the model strategically, where except for special cases the following three goals are in conflict: incentivizing improvement, predictive accuracy and retrieving the true underlying mechanism [42].

We build on the distinction between gaming and improvement [32] and adopt the perspective that explanations inform the actions of strategic agents [45]. However, in contrast to work in strategic modeling, we do not adapt the predictive model strategically, but maximize predictive accuracy. To ensure that individuals are nevertheless guided towards improvement, we suggest to strategically adapt the explanations instead.

3 BACKGROUND AND NOTATION

3.1 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 sometimes abbreviate as $p(y|x)$. We denote the estimated score function as $\hat{h}(x)$, which can be transformed into the binary decision function $\hat{f}(x)$. For the decision threshold t the decision is positive $\hat{f}(x) = 1$ if $\hat{h}(x) \geq t$ and vice versa $\hat{f}(x) = 0$ if $\hat{h}(x) < t$. Given $t = 0.5$ the most probable class is chosen.

3.2 Causal data model

We model the data generating process using a structural causal model (SCM) $\mathcal{M} \in \Pi$ [36, 37]. 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 $\mathbb{F} : \mathcal{U} \rightarrow \mathcal{X}$. 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 $asc(S)$, all non-ascendant indexes as $nasc(S)$, all descendant indexes as $d(S)$ and all non-descendant indexes as $nd(S)$. The set of observed variables is denoted as O where $O \subseteq D$. For the most part, we either assume $O = D$ or equivalently causal sufficiency of O, meaning that there is no variable $j \notin O$ that is a common cause of two variables $k, l \in O$ (see e.g., [37]). The Markov blanket $MB_O(Y)$ is the minimal subset of O that allows for an optimal prediction of Y, i.e. for which $X_O \perp\!\!\!\perp Y|X_{MB_O(Y)}$.[3]

SCMs allow to answer causal questions. I.e. they cannot only be used to describe (conditional) distributions (observation, rung 1 on Pearl's ladder of causation [36]), but can also be used to predict the (average) effect of actions $do(x)$ (intervention, rung 2) and image the results of alternative actions in light of factual observation $(x, y)^F$ (counterfactuals, rung 3).

As such, we model actions as structural interventions $a : \Pi \rightarrow \Pi$, 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. The interventional distribution for the action can be modeled by replacing the respective structural equations $f_I := \theta_I$. Counterfactuals are computed in three steps [36]: First, the factual distribution of exogenous variables U given the factual observation of the endogenous variables x^F is

[3] Sometimes the Markov blanket is defined as the minimal d-separating set. If faithfulness and the Markov property are fulfilled, both definitions coincide.

inferred (abduction), i.e., $P(U_j|X^F)$. Second, the structural interventions corresponding to $do(a)$ are performed (action). Finally, the counterfactual distribution $P(X^{SCF}|X = x^F, do(a))$ is computed from the abducted noise distribution and the intervened-upon structural equations (prediction).

4 NEGATIVE RESULT: ALGORITHMIC RECOURSE IS NEITHER MEANINGFUL NOR ROBUST

In the introduction we claimed that CR recommendations [21, 22] may not be meaningful and may not be robust with respect to refits of the model. Now, we formally demonstrate the case on the Covid office admission example (Figure 1) which we extend with the full structural causal model (Example 1). All code is publicly available.

Example 1. *Let V indicate whether someone is fully vaccinated, Y indicate whether someone is free of Covid-19 and S whether someone is asymptomatic. The data is generated by the following structural causal model (SCM) entailing the causal graph depicted in Figure 1:*

$$V := U_V, \qquad U_V \sim Bern(0.5) \tag{1}$$

$$Y := V + U_Y \mod 2, \qquad U_Y \sim Bern(0.09) \tag{2}$$

$$S := Y + U_S \mod 2, \qquad U_S \sim Bern(0.05) \tag{3}$$

For prediction, a `sklearn` *logistic regression model is fit on* 2000 *samples, yielding $\hat{h}$ with $\beta_v \approx 3.7$, $\beta_s \approx 5.1$, $\beta_0 \approx -4.3$. Employees are allowed to enter the office if $\hat{h} < 0.5$. Intervening on (flipping) V and S costs* 0.5 *and* 0.1 *respectively.*

Lack of meaningfulness: Given a decision threshold of 0.5, the model admits everyone without symptoms ($S = 1$), irrespective of their vaccination status V. Therefore, in order to revert rejections ($S = 0$), both individualized and subpopulation-based CR suggest removing the symptoms S ($do(S = 1)$, for instance by taking cough drops). However, since they only treat the symptoms S, the actual Covid risk Y is unaffected: none of the recourse-implementing individuals actually improve. We say the predictor is *gamed*.

Lack of robustness: For individuals who implement recourse the association between symptom-state S and Covid risk Y is broken. Thus, the predictive power of the model for recourse-seeking individual drops from $\approx$ 95 percent pre-recourse to $\approx$ 5 percent post-recourse.[4] A refit of the model on a mix pre- and post-recourse data (2000 samples each) yields $\hat{h}$ with $\beta_V \approx 4.1, \beta_S \approx 3.3, \beta_0 \approx -4.8$. Since the association between symptom state and disease status is broken post-recourse, the new model rejects individuals if they are not vaccinated, irrespective of their symptom-state. For that reason, recourse recommendations that were designed for the original model only lead to acceptance by the refitted model for those individuals who happened to be vaccinated anyway.

In conclusion, CR recommendations are prone to gaming the predictor and therefore may neither lead to improvement nor be robust to model refits. In this paper we suggest an alternative. Instead of guiding towards acceptance, our method guides towards improvement. The philosophical foundation of our method is developed in Section 5. The technical details of our method, meaningful causal recourse (MCR), are introduced in Section 6. We revisit Example 1 in Section 7, where we show that MCR is more meaningful and more robust to refits.

[4] The previously wrongly-rejected individuals are correctly classified after implementing recourse.

5 MEANINGFULNESS AND THE TWO TALES OF CONTRASTIVE EXPLANATIONS

In this Section, we lay the philosophical foundations for our method. More specifically, we argue that for recourse the acceptance constraint of CR should be replaced with an improvement constraint. Therefore, we first separate two purposes of contrastive explanations — *contestability of algorithmic decisions* and *meaningful, actionable recourse*. Then we argue that improvement is an essential requirement for recourse (Section 5.2). Moreover, recourse recommendations should not be constrained to lead to acceptance by a potentially instable predictor (Section 5.3). Clearly, in order to contest algorithmic decisions, more diverse explanations should be offered as well (Section 5.4).

5.1 Contestability and recourse as distinct goals

We first recall that a multitude of goals may be pursued with contrastive explanations [51]. As follows we argue that those goals may be in conflict. More specifically, we identify contestability and recourse as distinct, incompatible goals 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 give explainees grounds to contest unfavorable decisions [9, 51]. What is contested is the decision itself or the decision-making process. For contestability it is therefore crucial that explanations reflect the reasons for an algorithmic decision, i.e., explanations must be maximally faithful to the prediction model. Since not all violations of standards can be characterized exhaustively, it is important to offer diverse explanations [41] such that the explainee can contest decisions based on their personal beliefs and opinions.

Recourse recommendations on the other hand need to satisfy various constraints that are not related to the model. For instance, causal recourse takes causal dependencies between variables into account, which are not reflected in the prediction model [21]. Moreover, changes in features like age, ethnicity or height are commonly prohibited in recourse since they are not actionable for the explainee [21, 47]. Also, recourse recommendations must be plausible, i.e., make realistic suggestions that are jointly satisfiable and prefer sparse over widespread action recommendations [5, 20].

In conclusion, explanations geared to contest are more complete and faithful to the model while recourse recommendations are more selective and faithful to the underlying process and account for the limitations of the explainee.[5] We believe that the selectivity and reliance of recourse recommendations on factors besides the model itself is not a limitation but is rather indespensable for making explanations more relevant to the explainee.

5.2 In the context of recourse, improvement is desirable for model authority and explainee

In the same vein, we consider meaningfulness to be a further important requirement for recourse. Non-meaningfulness is disadvantageous for both explainee and model authority: For model authorities, to recommend (and guarantee) non-meaningful recourse is a (binding[6]) commitment to misclassification. For explainees, following non-meaningful recommendations means putting effort into actions that are rewarded in the short term but are not robust in the long term: Gaming actions are only honored by models that rely on the same associations. For instance, as demonstrated in Example 1, gaming actions may not be honored by model refits on a post-recourse population. In contrast, reversal of the underlying target Y is honored by any accurate predictor. We conclude that given the advantages for both model authority and explainee, recourse recommendations should aim to improve the underlying target Y.

[5] Yet, it could be argued that contesting decisions additionally offers an alternative route towards recourse in cases where model authorities admit mistakes and adapt their algorithmic decision-making system accordingly.
[6] It has been suggested that recourse must be granted even if the model changes [2, 49].

5.3 Improvement first, acceptance second

Taken that we constrain the optimization on improvement ($Y = 1$), it remains an open question how to guarantee acceptance ($\hat{Y} = 1$). One approach would be to constrain the optimization on both improvement and acceptance. However, additionally restricting on acceptance is either redundant or questionable: if improvement already implies acceptance, constraining on acceptance is redundant. In the remaining cases, improvement options would be withheld from the explainee because of the model's inability to classify the cases correctly.[7] Therefore, we suggest to optimize for improvement only and will demonstrate how to design prediction systems such that acceptance guarantees naturally ensue. In layman's terms, given that recourse leads to improvement, to ensure that recourse leads to acceptance one must simply make sure that the predictor is accurate. More specifically, since recourse leads to distribution shifts, the model must predict accurately in interventional environments. In order to design such intervention-stable predictors, the causal knowledge that is anyway required to guide towards improvement is sufficient. The approach is explained in more detail in Section 6. For the explainee to maintain agency it is important that they can make an informed choice. As such, the acceptance probability for an action must be communicated to the individual.

5.4 Separate explanations to contest algorithmic decisions

Meaningful recourse guides individuals towards actions that help them to improve, e.g., it recommends a vaccination to lower the risk to get infected with Covid. If, however, a explainee is more interested in contesting the algorithmic decision, (meaningful) recourse recommendations are not suitable. 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 place of birth[8]. In such situations, we suggest to additionally show explainees diverse explanations, which reflect the model's prediction mechanism and therefore enable to contest the decision. For example, such an explanation could be: if your place of birth would be different, your predicted Covid risk would have been lower.

6 MEANINGFUL CAUSAL RECOURSE (MCR)

We continue with the technical introduction of an explanation technique that targets improvement ($Y = 1$) instead of acceptance ($\hat{Y} = 1$), which we call Meaningful Causal Recourse (MCR). Like previous work in the field [22] we distinguish two settings: one where knowledge of the SCM can be assumed, for which we propose the counterfactual-based individualized MCR (inMCR, Section 6.1), and a setting where only the causal graph is known, for which we propose the interventional, subpopulation-based MCR (sMCR, Section 6.2).

For each setting we introduce a notion of meaningfulness. Further, we demonstrate how to design accurate decision systems such that in expectation the post-recourse model predicts the respective improvement probability. As a consequence, we are able to show that implementing MCR recommendations (which target improvement) also lead to acceptance (Section 6.3). Furthermore, we argue that meaningful recourse is more robust than non-meaningful recourse (Section 6.4).

6.1 Individualized Meaningful Causal Recourse (inMCR)

For individualized MCR we exploit knowledge of an SCMs for recourse generation since they cannot only be used to answer interventional (rung 2 on Pearl's ladder of causation), but also counterfactual questions (rung 3). In contrast to rung 2 predictions, counterfactuals are tailored to the individual and their situation [36]: they ask what would have been

[7] Respectively, we would argue that if improvement does not lead to acceptance the explainee is given grounds contest the decision.
[8] E.g., due to a spurious association with the causal variable *recent traveling*.

if one had acted differently and therefore 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 [21] we include both the prediction $\hat{Y}$ and the target variable Y as separate variables in the SCM. As a result, the SCM can also be used to model the individualized probability of improvement, and therefore to introduce an individualized notion of meaningfulness (Section 6.1.1). Moreover we suggest to exploit the SCM not only to generate explanations, but also for accurate post-recourse decision making (Section 6.1.2), such that acceptance guarantees can be derived (Section 6.3).

6.1.1 Individualized meaningfulness. We regard an action a as meaningful for an individual with observation x^{pre} if the counterfactual outcome $y^{post,a}$ is favorable. This counterfactual may be probabilistic.[9] Therefore we require it to be reverted with user-specified probability γ.[10]

Definition 6.1 (Individualized Meaningfulness Objective). For a pre-recourse observation x^{pre} a recourse recommendation a is γ-individually-meaningful with confidence γ if

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

inMCR differs from inCR (Equation 6) in its target: we optimize the counterfactual outcome of Y instead of $\hat{Y}$. Since the pre-recourse (factual) target Y cannot be observed standard counterfactual prediction cannot be directly applied. Details on how individualized meaningfulness can nevertheless be estimated are provided in Appendix B.2.

6.1.2 Accurate individualized post-recourse prediction. Recourse recommendations should not only guarantee improvement of Y but also revert the decision $\hat{Y}$. Whether acceptance guarantees naturally ensue from γ-meaningfulness depends on the ability of the predictor to recognize the made improvements.

If we were to use a standards predictive model post-recourse, there would be an imbalance in predictive capability between ML model and individualized MCR: MCR individualized its predictions using x^{pre} and the SCM. This knowledge is not accessible by observational predictors $\hat{h}^*$, such that improvement that was accurately predicted by MCR is not necessarily recognized by $\hat{h}^*$.[11] As a result, the γ-meaningfulness property could not be directly translated into an acceptance bound. We demonstrate the issue in more detail in Appendix B.3.

In order to settle the imbalance between MCR and the predictor, we suggest to leverage the SCM not only when generating individualized MCR recommendations but also for individualized post-recourse prediction, such that the predictor is at least as accurate as the pre-recourse MCR estimate. More formally, we suggest to estimate the post-recourse conditional distribution of Y given x^{pre}, $do(a)$ and the post-recourse observation $x^{post,a}$ (Definition 6.2). It resembles the counterfactual distribution, only that we additionally take the factual post-recourse observation of the covariates into account.

Definition 6.2 (Individualized post-recourse predictor). We define the individualized post-recourse predictor as

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

[9] Especially since Y cannot be observed when the explanation is requested, as elaborated in Appendix B.2.

[10] Building on the deterministic criterion in [21], [22] present probabilistic causal recourse optimization problems (Equations 6 and 7). They require the expectation of the counterfactual prediction to exceed some context-dependent threshold. The formulation is similar to our formulation since the threshold can be translated into a confidence γ. For more details refer to Appendix B.1.

[11] One may also argue that standards 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 demonstrates. We could refute this criticism using results that we present in Section 6.2.2, where we show that $\hat{h}^*$ is stable with respect to MCR actions.

Details on the estimation of the individualized post-recourse predictor are given in Appendix B.4. For the individualized post-recourse predictor, improvement probability and prediction are tightly linked (Proposition 6.3). More specifically, the expected post-recourse prediction $h^{*,ind}$ is equal to the individualized improvement probability $\gamma(x^{pre}, a)$. This property will be exploited in Section 6.3, where we derive acceptance guarantees for MCR.

Proposition 6.3. *The expected individualized post-recourse score is equal to the individualized pre-recourse improvement probability* $\gamma(x^{pre}, a)$.

$$E[\hat{h}^{*,ind}(x^{post})|x^{pre}, do(a)] = \gamma(x^{pre}, a) := P(Y^{post} = 1|x^{pre}, do(a)).$$

6.2 Subpopulation-based Meaningful Causal Recourse (sMCR)

With individualized MCR we can guide individuals towards improvement. However, individualized MCR requires a fully specified SCM, which is hard to infer. If the SCM is not specified, but the causal graph is known instead, we can still estimate the effect of interventions (rung 2) from observational data. Interventional distributions describe the whole population and therefore provide limited insight into the effect for a specific individual. Following Karimi et al. [22] we narrow the distribution down to a subpopulation of similar individuals for which we then estimate the causal effect.[12]

In contrast to existing work we include the underlying target Y and prediction $\hat{Y}$ in the causal graph as two separate variables and target improvement instead of acceptance. Curiously, we are able to show that optimal observational predictors maintain their predictive power in environments where MCR actions have changed the environment (Section 6.2.2), which is not the case for CR. As a consequence, we can show that in settings where only causes of Y are observed, subpopulation-based CR and MCR coincide (Section 6.2.3) and can derive acceptance guarantees (Section 6.3).

6.2.1 Subpopulation meaningfulness. Subpopulation-based γ-meaningfulness of an action a is achieved if the action leads to the desired improvement of Y within a subgroup of similar individuals with at least probability γ. Like Karimi et al. [22], we consider individuals to belong to the same subgroup if the variables that are not affected by the intervention take the same values. More formally, these are the non-descendants of the intervened-upon variables in the causal graph. For action a we denote the set of subgroup characteristics as $G_a := nd(I_a)$.

The set G_a is chosen for practical reasons. In order to make the estimation more accurate, we would like to make the subgroup as small as possible, and therefore condition on as many characteristics of the individual as we can. However, without access to a SCM, one can only identify interventional distributions for subgroups of the population by conditioning on their post-intervention characteristics [13, 36]. Their post-intervention state is not observed and differs from the observed pre-intervention state if the respective variables are affected by the intervention (G_a). Therefore, the post-intervention values are only known for variables that are not affected by the intervention. In order to make the estimation as accurate as possible without requiring access to the SCM (and while making sure that the individual of interest is actually part of the subgroup), we therefore condition on G_a.

In our illustrative example, for an intervention on the vaccination status, all remaining variables are affected by the action and therefore the set of group characteristics would be empty. If we were to observe a cause of the vaccination status, like the individual's trust in science T, other causes of Y, like the number of people in their office N, or unaffected

[12] This causal effect resembles the conditional average treatment effect[1, 22].

causes of the symptom-state S like whether a person has a chronic infection C, those would be subgroup characteristics. So instead of evaluating the effect of a vaccination for the overall population, we would evaluate the effect for a subgroup of people who share the trust in science and the number of people in the office and whether they have chronic infections.

For the evaluation of meaningfulness, we are only interested in the effect on Y. As such, interventions on variables that do not cause Y are irrelevant for the estimation. Furthermore, including them in the subpopulation-based estimation may reduce the set of variables that are not affected by the action and thus the set of subpopulation characteristics G. Therefore, we suggest to reduce the action to interventions on causes of Y and denote the resulting action as a'.

Definition 6.4 (Subpopulation-based meaningfulness). A recourse recommendation a is γ-subpopulation-meaningful if

$$P(Y^{post} = 1|do(a'), x^{pre}_{G_{a'}}) \geq \gamma. \tag{4}$$

Given that there are no unobserved confounders (causal sufficiency), subpopulation-based meaningfulness can be estimated with the causal graph $\mathcal{G}$ and the observational distribution (Appendix E.5).

6.2.2 Accurate post-recourse prediction. In order to enable recourse guarantees, the decision model must honor the made improvements and therefore predict accurately despite the distribution shift induced by recourse. We recall that in general predictive models are not stable with respect to interventions, meaning that CR actions may not affect prediction and the underlying target coherently (as demonstrated on Example 1).

In contrast to the negative results for CR, we are able to prove that MCR recommendations revert both the target as well as the prediction. The reason is that optimal observational predictors are stable w.r.t MCR interventions, which is not the case for CR. The key difference to CR is that MCR actions exclusively intervene on causes of Y: Interventions on effects 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 and therefore $MB(Y) \subseteq S$). In contrast, given causal sufficiency, the conditional $P(Y|X_S)$ is invariant to interventions on causes of Y.

Proposition 6.5. *Given nonzero cost for all interventions, MCR exclusively suggests actions on causes of Y.*

Proposition 6.6. *Given causal sufficiency, any set S that allows for an optimal prediction (i.e., $MB(Y) \subseteq S$) is stable with respect to MCR actions. As a consequence, the expected subgroup-wide score h^* is equal to the subgroup-wide improvement probability $\gamma(x^{pre}_{G_{a'}}, a)$.*

$$E[\hat{h}^*(x^{post})|x^{pre}_{G_{a'}}, do(a)] = \gamma(x^{pre}_{G_{a'}}, a) := P(Y^{post} = 1|do(a), x^{pre}_{G_{a'}}).$$

As Proposition 6.6 demonstrates, the expected prediction and the subgroup-based improvement probability coincide. We will use the property to link sMCR to sCR (Section 6.2.3) and to derive acceptance guarantees (Section 6.3).

6.2.3 Link to causal recourse. Since under causal sufficiency optimal predictors reliably estimate the improvement under interventions on causes we can link subpopulation-based CR and MCR.

Proposition 6.7. *For actions a that only intervene on causes of Y, $p^{pre}(x^{post}) > 0$, a cross-entropy optimal binary predictor and causal sufficiency, we can equivalently define γ-subpopulation meaningfulness as*

$$E[h^*(\theta_{I_a}, x^{pre}_{G_{a'}}, x^{post}_{-(G_{a'} \cup I_a)})|do(a), x^{pre}_{G_{a'}}] \geq \gamma.$$

Proposition 6.7 can be interpreted in two directions: Firstly, we see that in scenarios where all observed variables are causes of Y, CR recommendations are $\texttt{thresh}(a)$-subpopulation meaningful.[13] Secondly, we can transform subpopulation-based CR into a $\texttt{thresh}(a)$ subpopulation-based meaningfulness technique by adding a constraint that only allows interventions on causes of Y.

6.3 Acceptance guarantees

For the presented accurate post-recourse predictors, acceptance naturally ensues from γ-meaningfulness. The reason is that the post-recourse prediction is linked with the improvement confidence γ (Propositions 6.3 and 6.6). More specifically, we are able to derive a lower bound on the acceptance probability 6.8.

Proposition 6.8. *Let x_S^{pre} be the subset of the pre-recourse observed variables that were taken into account to compute MCR, i.e. $S = D$ for individualized recourse and $S = G_{a'}$ for subpopulation-based recourse. For subpopulation-based MCR, we furthermore assume that $p^{pre}(x^{post}) > 0$. For pre-recourse state x^{pre}, a γ-confident action a, the (individualized) optimal predictor h^* and a global decision threshold t the post-recourse acceptance probability η is*

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

where $\gamma(x_S^{pre}, a) = p(Y^{post,a} = 1 | x_S^{pre}, a)$.

We can now tune the model's decision threshold and confidence γ accordingly to yield the desired acceptance rate guarantees. For instance, given $\gamma = 0.95$ and a global decision threshold $t = 0.5$ we can guarantee acceptance with probability $\eta \geq 0.9$. If the model authority would like to guarantee acceptance with certainty, the decision threshold can be set to $t = 0$.[14]

This acceptance guarantee is vital for the agency of the explainee. Only if the acceptance probability is made explicit, they can make an informed choice.

6.4 Robustness

As demonstrated in Example 1, CR may not alter prediction $\hat{Y}$ and target Y coherently. As a consequence, the predictor's performance deteriorates. In order to counteract the drop, the model authority must refit the model on post-recourse data. This refit may not honor recourse recommendations that correspond to gaming anymore.

In contrast, inMCR recommendations do not target one specific predictor, but focus on improving the underlying target. Furthermore, instead of using an (observational) predictor in interventional environments, we introduce a SCM-based decision system that is robust to the shifts induced by MCR.

Furthermore, we demonstrate that given causal sufficiency optimal predictors are robust with respect to sMCR actions. As a consequence, optimal pre-recourse predictors remain optimal in the respective interventional distributions. Therefore, we conjecture that sMCR recommendations are more robust to model refits than sCR recommendations.

6.5 Optimization

In order to generate individually meaningful recourse recommendations, we can optimize

$$\text{argmin}_{a=do(X_I=\theta)} \quad \text{cost}(a, x^{pre}) \quad \text{subject to} \quad a \text{ is } \gamma\text{-meaningful.} \tag{5}$$

[13] $\texttt{thresh}(a)$ is the CR equivalent to γ.

[14] We conjecture that the model's decision threshold can be adapted to the individual or subgroup to yield more exact acceptance, false negative or false positive rate guarantees. We sketch an approach in Appendix B.6.

Table 2. Overview of the results for the Covid admission example. ζ denotes the percentage of recourse-seeking individuals for which a recommendation for the respective confidence level could be made. $\gamma_{\text{obs.}}$ is the percentage of recourse-implementing individuals for which improvement was achieved. $\eta_{\text{obs.}}$ denotes the observed acceptance rate for recourse-implementing individual for the original model and the post-recourse refit øcost denotes the average recourse cost among recourse-seeking individuals and øintv. the average number of interventions on causal/non-causal variables.

Office adm.	ζ		$\gamma_{\text{obs.}}$		$\eta_{\text{obs.}}$		$\eta^{\text{refit}}_{\text{obs.}}$		øcost		øintv.	
γ/thresh :	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	causes	other
ind. CR	1.00	1.00	0.05	0.05	1.00	1.00	0.13	0.13	0.10	0.10	0.00	1.00
sub. CR	1.00	1.00	0.05	0.05	1.00	1.00	0.13	0.13	0.10	0.10	0.00	1.00
ind. MCR	0.87	0.87	1.00	1.00	1.00	1.00	1.00	1.00	0.50	0.50	1.00	0.00
sub. MCR	0.87	0.00	1.00	-	1.00	-	1.00	-	0.50	-	1.00	0.00

Like the optimization problems for counterfactuals explanations [45, 51] or CR [22], the optimization problem is computationally challenging. It can be seen as a two-stage problem, where in the first stage the intervention targets I_a, and in the second stage the corresponding intervention values θ_a are optimized [22]. For the selection of intervention targets I_a, $2^{d'}$ combinations exist, with $d' < d$ being the number of causes of Y. We use the *Nondominated Sorting Genetic Algorithm II* (NSGA-II) [6]. The computational complexity of optimizing the intervention value θ depends on the type of data. For mixed categorical and continuous data, previous work in the field [5] suggests to use NSGA-II [6] in combination with *mixed integer evoluation strategies* [27].

7 SIMULATED EXAMPLE

In this Section, we revisit the Covid office admission example (Example 1) to illustrate the differences between CR and MCR. More specifically, we asses whether the approaches lead to acceptance and improvement, as captured by the observed improvement rate ($\gamma_{\text{obs.}}$) and the observed acceptance rates for the pre-recourse predictor ($\eta_{\text{obs.}}$) and a refit ($\eta^{\text{refit}}_{\text{obs.}}$). We compare the average recourse cost as well as the average number of interventions on causal and non-causal variables. All of the aforementioned statistics only take those individuals into account, for which a recourse recommendation could be made. The percentage of recourse-seeking individuals for which a recourse recommendation could be suggested is denoted as ζ. We compare individualized causal recourse (ind. CR), subpopulation-based causal recourse (sub. CR), individualized MCR (ind. MCR) and subpopulation-based MCR (sub. MCR). Implementation details are provided in Appendix D. All code is publicly available. The results are summarized in Table 2.

While CR recommends to treat the symptoms, MCR suggests to get vaccinated. Both strategies lead to acceptance, but only MCR actually leads to a decreased risk of catching Covid. Although the CR recommendations are cheaper, the suggested recourse recommendations are not robust to refits on mixed pre- and post-recourse data: only around 13 percent were honored. In contrast, all MCR recommendations were honored by a refit on pre- and post-recourse data or by the individualized post-recourse predictor.

Since for the sMCR actions ($do(V = 1)$) the set of subgroup-characteristics is empty, the information about the pre-recourse symptom-state cannot be used to predict individualized treatment effects. As a result, no 0.95-confident sMCR recommendation can be made. In contrast, individualized MCR takes the pre-recourse symptom-state into account and is thereby able to accurately detect cases for which recommendations can be made with confidence $\gamma = 0.95$.

8 LIMITATIONS AND DISCUSSION

8.1 Causal knowledge and assumptions

inMCR requires a fully specified SCM. sMCR relaxes the assumptions of inMCR to requiring knowledge of the causal graph and causal sufficiency. However, both SCM and causal graph are hard to identify [37] and causal sufficiency is difficult to test [19].

However, guiding action without causal assumptions is impossible. Assuming that one can estimate causal effects without taking causal relationships into account is an even more unrealistic causal assumption, namely that all covariates are causal for the target and that they do not influence each other (causal independence). Thereby, we join a broad range of work that demonstrates the necessity of causal knowledge in explainability [10, 18, 52, 55] and fairness [23, 25, 29, 54]. Due to its importance in many applications, causal discovery is an active area of research [12, 17, 30, 37, 44]. Therefore, we hope that future research on causal inference will make MCR feasible in practice.

8.2 Intervention stability and extrapolation

The acceptance guarantees for sMCR are based on the insight that the conditional $P(Y|X)$ is stable with respect to sMCR actions. However, even if all assumptions are met and the conditional distribution is invariant, the joint distribution of the variables is affected if individuals act on recourse recommendations. If the interventional joint distribution extends the support of the observational distribution, the model would be forced to predict outside of the training distribution, where we cannot expect it to accurately model the conditional expectation. Therefore Propositions 6.8 and 6.7 are restricted to areas with $p(x) > 0$. However, we conjecture that the problem can be mitigated: One could leverage access to the causal graph to simulate post-recourse data, thereby extending the support of the model without affecting its performance in the observational support.[15]

8.3 Invariance assumptions

Like CR [22], MCR makes invariance assumptions. For inMCR the unobserved influences (as captured by the exogeneous variables U) are assumed to be invariant within individuals over time. Only given this invariance does the factual observation x^{pre} contain invariant information about the explainee and only then structural counterfactuals cannot only explain *what would have been in the past*, but also predict *what would be in the future*. Furthermore, sMCR implicitly assumes that the conditioned-upon variables, namely the variables that are not affected by the intervention, are invariant within individuals. If we think of the variables as characteristics of the individual or an individual's static context, the assumption can be regarded as reasonable. If the variables represent a volatile situational influence, the information is not predictive. In principle, it is possible to base the predictions on the subset of variables that is invariant.

8.4 Ability to provide recourse recommendations

As the experimental results show, it is not always possible to find γ-meaningful recourse recommendations. One may argue that it is unrealistic to expect explanations that guarantee improvement with near certainty. However, in order to provide a variety of options, model authorities should aim to observe actionable causes that explain the variation of Y, allowing a greater variety of improvement options.

[15] As the refit scenario in Example 1 demonstrates, such an extension of the support would not be possible for CR.

9 CONCLUSION

In this work, 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 CR may suggest to game the predictor. Such recourse recommendations are binding commitments to misclassification. Furthermore, we showed that such recourse is not robust to refits of the model on post-recourse data. We tackled the problem in the explanation domain and introduced Meaningful Causal Recourse (MCR), an explanation technique that aims to improve the underlying prediction target. Although MCR does not constrain the recommendations to lead to acceptance by the predictor, we were able to derive acceptance guarantees: We demonstrated how the causal knowledge that is required for MCR can also be exploited to design decision systems that honor improvements. On a simulated example we demonstrated that MCR is more meaningful and more robust to refits than the acceptance-focused recourse method CR.

As of today, the applicability of the method is limited by the lack of causal knowledge in a range of real-world applications. However, if one aims to guide individuals towards improvement, causal assumptions must be made. By introducing MCR, we make these assumptions explicit.

REFERENCES

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. 2015. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* 33, 4 (2015), 485–505.

[2] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 80–89.

[3] Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. 2020. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024* (2020).

[4] Yatong Chen, Jialu Wang, and Yang Liu. 2020. Linear Classifiers that Encourage Constructive Adaptation. *arXiv preprint arXiv:2011.00355* (2020).

[5] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI*, Thomas Bäck, Mike Preuss, André Deutz, Hao Wang, Carola Doerr, Michael Emmerich, and Heike Trautmann (Eds.). Springer International Publishing, Cham, 448–469.

[6] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.

[7] Ricardo Dominguez-Olmedo, Amir-Hossein Karimi, and Bernhard Schölkopf. 2021. On the Adversarial Robustness of Causal Algorithmic Recourse. *arXiv preprint arXiv:2112.11313* (2021).

[8] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13 (jul 2012), 2171–2175.

[9] Timo Freiesleben. 2021. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines* (Oct 2021).

[10] Christopher Frye, Colin Rowat, and Ilya Feige. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems* 33 (2020), 1229–1239.

[11] Dan Geiger, Thomas Verma, and Judea Pearl. 1990. Identifying independence in Bayesian networks. *Networks* 20, 5 (1990), 507–534.

[12] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.

[13] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer.* John Wiley & Sons.

[14] Charles AE Goodhart. 1984. Problems of monetary management: the UK experience. In *Monetary theory and practice.* Springer, 91–121.

[15] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z Wang. 2020. Maximizing welfare with incentive-aware evaluation mechanisms. *arXiv preprint arXiv:2011.01956* (2020).

[16] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science.* 111–122.

[17] Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. 2018. Causal structure learning. *Annual Review of Statistics and Its Application* 5 (2018), 371–391.

[18] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems* 33 (2020), 4778–4789.

[19] Dominik Janzing, Eleni Sgouritsa, Oliver Stegle, Jonas Peters, and Bernhard Schölkopf. 2012. Detecting low-complexity unobserved causes. *CoRR* abs/1202.3737 (2012). arXiv:1202.3737

[20] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-Agnostic Counterfactual Explanations for Consequential Decisions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, Online, 895–905.

[21] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 353–362.

[22] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., virtual, 265–277.

[23] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems* 30 (2017).

[24] Jon Kleinberg and Manish Raghavan. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.

[25] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[26] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The Dangers of Post-Hoc Interpretability: Unjustified Counterfactual Explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. AAAI Press, Macao, China, 2801–2807.

[27] Rui Li, Michael TM Emmerich, Jeroen Eggermont, Thomas Bäck, Martin Schütz, Jouke Dijkstra, and Johan HC Reiber. 2013. Mixed integer evolution strategies for parameter optimization. *Evolutionary computation* 21, 1 (2013), 29–64.

[28] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2020. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. arXiv:1912.03277 [cs.LG]

[29] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553* (2020).

[30] Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass* 13, 1 (2018), e12470.

[31] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.

[32] John Miller, Smitha Milli, and Moritz Hardt. 2020. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Online, 6917–6926.

[33] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* 22, 57 (2021), 1–64.

[34] Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) (Proceedings of Machine Learning Research, Vol. 124)*, Jonas Peters and David Sontag (Eds.). PMLR, Online, 809–818.

[35] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. 2022. Algorithmic Recourse in the Face of Noisy Human Responses. *arXiv preprint arXiv:2203.06768* (2022).

[36] Judea Pearl. 2009. *Causality* (2 ed.). Cambridge University Press, Cambridge, UK.

[37] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[38] Niklas Pfister, Evan G. Williams, Jonas Peters, Ruedi Aebersold, and Peter Bühlmann. 2021. Stabilizing variable selection and regression. *The Annals of Applied Statistics* 15, 3 (2021), 1220 – 1246.

[39] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481.

[40] Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. 2021. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. arXiv:2012.11788 [cs.LG]

[41] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 20–28.

[42] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. 2020. Causal Strategic Linear Regression. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, virtual, 8676–8686.

[43] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[44] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, Vol. 3. SpringerOpen, 1–28.

[45] Stratis Tsirtsis and Manuel Gomez Rodriguez. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems* 33 (2020), 16749–16760.

[46] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 2021. Towards Robust and Reliable Algorithmic Recourse. arXiv:2102.13620 [cs.LG]

[47] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 10–19.

[48] Bart Van Liebergen et al. 2017. Machine learning: a revolution in risk management and compliance? *Journal of Financial Transformation* 45 (2017), 60–67.

[49] Suresh Venkatasubramanian and Mark Alfano. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 284–293.

[50] Julius von Kügelgen, Nikita Agarwal, Jakob Zeitler, Afsaneh Mastouri, and Bernhard Schölkopf. 2021. Algorithmic Recourse in Partially and Fully Confounded Settings Through Bounding Counterfactual Effects. *arXiv preprint arXiv:2106.11849* (2021).

[51] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[52] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. 2021. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 721–729.

[53] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689–722.

[54] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. Issue: 1.

[55] Qingyuan Zhao and Trevor Hastie. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics* 39, 1 (2021), 272–281.

A EXTENDED BACKGROUND

A.1 d-separation

Two variable sets X, Y are called d-separated [11, 43] by the variable set Z in a graph $\mathcal{G}$ ($X \perp\!\!\!\perp_{\mathcal{G}} Y|Z$), if, and only if, for every path p holds either (i) p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ where $m \in Z$ or (ii) p contains a collider $i \to m \leftarrow j$ such that m and all of its descendants n it holds that $m, n \notin Z$.

A.2 Generalizability and intervention stability

We leverage necessary conditions for invariant conditional distributions as derived in [38]. The authors introduce a d-separation based intervention stability criterion that is applied to a modified version of $\mathcal{G}$. For every intervened upon variable X_l an auxiliary intervention variable, denoted as I_l, is added as direct cause of X_l, yielding $\mathcal{G}^*$. The intervention variable can be seen as a switch between different mechanisms. A set $S \subseteq \{1, \dots, d\}$ is called *intervention stable* regarding a set of actions if for all intervened upon variables X_l (where $l \in I^{\text{total}}$) the d-separation[16] $I^l \perp\!\!\!\perp_{\mathcal{G}^*} Y|X_S$ holds in $\mathcal{G}^*$. The authors show that intervention stability implies an invariant conditional distribution, i.e., for all actions $a, b \in \mathbb{A}$ with $I^a, I^b \subseteq I^{\text{total}}$ it holds that $p(y^a|x_S) = p(y^b|x_S)$ (Pfister et al. [38], Appendix A).

A.3 Causal recourse

MCR is closely related to the CR framework [21, 22], but differs substantially in its motivation and target. In order to allow for a direct comparison we briefly sketch the main ideas and the central CR definitions in our notation. Like MCR, CR aims to guide individuals to revert unfavorable algorithmic decisions (recourse). Therefore, they suggest to search for cost-efficient actions that lead to acceptance by the prediction model. Actions are modeled as structural interventions $a : \Pi \to \Pi$, which can be constructed as $a = do(\{X_i := \theta_i\}_{i \in I})$, where I is the index set of features to be intervened upon [21]. The conservativeness of the suggested actions can be adjusted using the hyperparameter γ_{LCB}, that determines the adaptive threshold $\texttt{thresh}(a)$ and thereby how many standard deviations the expected prediction shall be away from the model's decision threshold t. In order to accommodate different levels of causal knowledge, two probabilistic versions of CR were introduced [22]: While individualized recourse assumes knowledge of the SCM, subpopulation-based CR only assumes knowledge of the causal graph.

A.3.1 *Individualized recourse.*

Individualized recourse predicts the effect of actions using structural counterfactuals [21], which require a full specification of the SCM, but allow for individualized prediction. [17]

Given a function that evaluates the cost of actions, the optimization goal for individualized causal recourse is given in Equation 6. We abbreviate the counterfactual post-recourse distribution of the covariates given pre-recourse observation x^{pre} and action $a \in \mathbb{A}$, commonly $p(x|x^{pre}, do(a))$, as $p(x^{post,a})$.[18]

$$a^* \in \underset{a \in \mathbb{A}}{\text{argmin}}\ \text{cost}(a, x^{pre}) \text{ subject to } \mathbb{E}[\hat{h}(x^{post,a})] \geq \texttt{thresh}(a) \quad \text{with} \quad \texttt{thresh}(a) = 0.5 + \gamma_{LCB}\sqrt{\text{Var}[\hat{h}(x^{post,a})]} \tag{6}$$

[16] Background on d-separation in Appendix A.1.
[17] For individualized CR, the exogenous variables U are implicitly assumed to capture static unobserved characteristics rather than varying situational influences. In such a setting, counterfactuals do not only entail a retrospective but also predictive meaning.
[18] More specifically, the authors denote the counterfactual as $p(x^{SCF})$Further constraints have been suggested, e.g., $x^{post,a} \in \mathcal{P}$lausible or $a \in \mathcal{F}$easible [5, 21, 26, 28, 47].

A.3.2 Subpopulation-based recourse: If no knowledge of the SCM is given, counterfactual distributions cannot be estimated and consequently individualized recourse recommendations cannot be computed. Given knowledge of the causal graph $\mathcal{G}$, subpopulation-based recourse recommendations can be made. Subpopulation-based recourse is based on the average treatment effect within a subgroup of similar individuals [22]. More specifically individuals belong to the same group if the non-descendants $nd(I)$ of intervention variables (which ceteris paribus remain constant despite the intervention) take the same value. This grouping is designed to make use of as many characteristics as possible while still allowing observational identifiability of the effect. The subpopulation-based objective is given in Equation 7.

$$a^* \in \underset{a \in \mathbb{A}}{\operatorname{argmin}} \operatorname{cost}(a, x^{pre}) \text{ subject to } \mathbb{E}_{X_{d(I)}|do(X_I=\theta), x^{pre}_{nd(I)}}[\hat{h}(x^{pre}_{nd(I)}, \theta, X_{d(I)})] \geq \texttt{thresh}(a). \tag{7}$$

B DETAILS

B.1 Why we chose a non-adaptive improvement confidence γ

Karimi et al. [22] phrase the optimization objective for CR in terms of an expectation over the counterfactual prediction distribution and an action-adaptive threshold. The authors introduce an adaptive threshold that also takes the standard deviation of the counterfactual prediction distribution into account. The cost and acceptance probability can then be traded off with the hyperparameter of the adaptive threshold.

We decided not to use fixed user-defined threshold for MCR, since in the context of MCR improvement is prioritized over acceptance. Furthermore, γ can be intuitively interpreted as improvement probability (whereas the expected prediction cannot be interpreted as acceptance probability). In order to adaptively achieve the desired acceptance rate for every individual, in addition to altering γ (Sections **??** and 6.2.2) we sketch how to fine-tune the model's decision threshold (instead of γ) to the individual (Appendix B.6).

B.2 Estimation of individualized meaningfulness

In order to optimize individualized meaningfulness, we need to be able to estimate the counterfactual $P(Y^{post,a} = 1)$. In principle, given the SCM, the computation of counterfactual distributions with SCMs is a common task [36]. They are computed in three steps: abduction, intervention and prediction (Section A.3). In the context of MCR, only intervention and prediction can be performed as usual.

For abduction, the exogenous influences U are usually reconstructed from the all endogenous variables by modeling the conditional $P(U|X = x, Y = y)$. In the context of MCR, we do not observe Y.[19] Therefore, we suggest to resort to abduction using the observed endogenous variables, i.e. $P(U|X = x)$.

Although a purely observational task, estimating the conditionals is challenging [33]. For situations where all endogenous variables are observed analytical solutions are available. For every node X_j abduction is a function of its direct parents and its observed value:[20]

$$p(u_j|x^{pre}) = p(u_j|x_j^{pre}, x_{pa(j)}^{pre}) = \frac{p(x_j^{pre}|u_j, x_{pa(j)}^{pre})p(u_j)}{\int p(x_j^{pre}|u_j, x_{pa(j)}^{pre})p(u_j)} = \frac{\delta_{f_j(x_{pa(j)}^{pre}, u_j), x_j^{pre}} p(u_j)}{\int \delta_{f_j(x_{pa(j)}^{pre}, u_j), x_j^{pre}} p(u_j) du_j}. \tag{8}$$

Leveraging knowledge about the SCM the estimation can be simplified further. For instance, for invertible structural equations, the (deterministic) abduction function is given as

$$u_j = f^{-1}(x_j^{pre}; x_{pa(j)}^{pre})$$

.

These analytical formulas can be applied for the abduction of all nodes except Y and X_j with $j \in ch(Y)$ (for which we would need to observe Y). Since we are only interested in the counterfactual distribution of Y, only the abduction of U_Y and its ancestors is necessary (and the abduction of U_j with $j \in ch(Y)$ is not). It remains to show how to (analytically) abduct U_Y from X.[21]

[19] Otherwise we would have needed a prediction model in the first place.
[20] Given independent noise terms, u_j is independent from all other variables given x_j and $x_{pa(j)}$, which can easily be proven using d-separation.
[21] The reconstructed noise terms $U_Y, U_{asc(Y)}$ are pairwise independent given X. Therefore, they can be reconstructed one by one.

Therefore, we reformulate $P(U_Y|X)$ in terms of tractable expressions[22] that are given by the SCM. For instance, for binary targets and invertible structural equations, we can rewrite the abducted probability as product of marginal noise terms for the unobserved variable and its children, as Proposition B.1 demonstrates. A proof for Proposition B.1 is given in Appendix E.2.

Proposition B.1. *In general, abduction of u_Y can be performed using*

$$p(u_Y|x^{pre}) = \frac{p(u_Y)\prod_{i\in ch(Y)} p(x_i^{pre}|x_{pa(i)}^{pre}, f_Y(x_{pa(Y)}^{pre}, u_Y))}{\int_{\mathcal{U}_Y} p(u'_Y)\prod_{i\in ch(Y)} p(x_i^{pre}|x_{pa(i)}^{pre}, f_Y(x_{pa(Y)}^{pre}, u'_Y))du'_Y}. \tag{9}$$

Given invertible structural equations, a binary target variable Y and given that all endogenous variables except for Y can be observed, u_y can be abducted using

$$p(u_y|x) = \frac{p(u_Y)\prod_{i\in ch(Y)} p(u_i(u_Y))}{\sum_{y'\in\{0,1\}} p(u_Y(y'))\prod_{i\in ch(Y)} p(u_i(u_Y(y')))} \tag{10}$$

where $u_Y(y) = f^{-1}(y; x_{pa(Y)})$ and $u_i(u_y) = f_i^{-1}(x_i; f_Y(x_{pa(Y)}, u_Y), x_{pa(i)})$.

Since we can sample from $P(U)$, the integral in Equation 9 can be approximated using Monte Carlo integration.

B.3 Imbalance between standard predictors and individualized MCR recommendations

Example 2. *Let there be a three variable chain $X_1 \to Y \to X_2$ where at every step the value is incremented by one with 50% chance and the maximum value is set to 2 ($X_1 := U_1, Y := X_1 + U_Y, X_2 := min(2, Y + U_2)$ where $U_1, U_2, U_Y \sim Bern(0.5)$). Let us assume a factual observation $x^{pre} = (0, 2)$ and action $a = do(X_1 = 1)$ yielding $x^{post,a} = (1, 2)$. For the observation $x^{pre} = (0, 2)$ we can infer that U_Y must have been 1, since two increments are needed to get from 0 to 2. However, from the post-intervention observation $x^{post} = (1, 2)$ we cannot infer where the increment happened (U_Y or U_2). As a consequence, an optimal predictive model that only has access to x^{post} would predict that $y^{post,a}$ for $x^{post} = (1, 2)$ could be 1 or 2 with equal likelihood. In contrast, with access to x^{pre} and the SCM we can infer that $y^{post,a} = 2$ since $U_Y = 1$.*

In the above example, given knowledge of the SCM, the pre-intervention observation x^{pre} and the performed action a we can already abduct U_Y perfectly and therefore correctly determine the post-intervention state of Y (even without access to the post-intervention observation $x^{post,a}$). In contrast, with the post-recourse observation alone it is impossible to reconstruct U_Y and therefore impossible to determine the post-intervention state of Y.[23] In the context of MCR this means that the observational predictor's post-recourse predictions are not directly linked with γ: they may not honor the implementation of a γ-individually-meaningful action, even though we can guarantee improvement with certainty ($\gamma = 1$).

B.4 Estimation of the individualized post-recourse predictor

In order to allow the estimation of the post-recourse prediction, we decompose Definition 6.2 into tractable components. More specifically, the post-recourse prediction can be composed of the conditional distributions of the endogenous variables given a state of the exogenous variables, and the abducted probabilities of the exogenous variables given the

[22] Given the SCM, we have access to the marginal distribution of the noise terms $P(U_j)$ as well as the structural equations f_j. The structural equations determine the value of a variable x_j given the state of the parents $(x, y)_{pa(j)}$ and the noise value u_j. Therefore, we can also sample from the conditional distribution of variables given their parents. Furthermore, given invertability, the conditional distribution of the noise term given the corresponding variable and its parents $P(U_j|X_j, (X, Y)_{pa(j)})$ reduces to evaluating the marginal noise distribution.

[23] The optimal pre-recourse predictor $\hat{h}^*(x^{post})$ predicts 0.5 for both $y = 1$ and $y = 2$.

pre-recourse observation. A general formula as well as a simplified version for binary decision problems with invertible structural equations is given in Proposition B.2. A proof can be found in Appendix E.3.

Proposition B.2. *In general, the individualized post-recourse predictor can be estimated as*

$$p(y^{post,a}|x^{pre}, x^{post,a}) = \frac{\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})}{\int_{\mathcal{Y}} \left(\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})du\right) dy} \tag{11}$$

Given binary decision problems with invertible structural equations, the individualized post-recourse prediction function reduces to

$$p(y^{post,a}|x^{post,a}, x^{pre}) = \frac{p(U_{-I} = f^{-1}(y^{post,a}, x^{post,a})|x^{pre})}{\sum_{y' \in \{0,1\}} p(U_{-I} = f^{-1}(y', x^{post,a})|x^{pre})}. \tag{12}$$

The integrals in Equation 11 can be approximated using Monte Carlo integration. For binary Y, the integral over Y reduces to a sum, as in Equation 12.

B.5 Observational identifiability of sMCR

Given causal sufficiency, subpopulation-meaningfulness is observationally identifiable (Proposition B.3, Proof in Appendix E.5).

Proposition B.3. *Given causal sufficiency of $\mathcal{G}$, the conditional interventional distribution is observationally identifiable.*

$$p(y|do(a), x^{pre}_{G_{a'}}) = \int_{\mathcal{X}_\Gamma} p(y|x_{pa(Y)}) \prod_{r \in \Gamma} p(x_r|x_{pa(r)}) dx_\Gamma \Big|_{do(a), x^{pre}_{nd(I^a)}}$$

Here $\Gamma := \{r : r \in asc(Y) \wedge r \in d(I)\}$ is the set of ancestors of Y that are descendants of X_I.
For actions a on non-causes of Y ($I^a_{asc} = \emptyset$) it holds that $p(Y = 1|do(X_{I^a} = \theta), x^{pre}_{G_{a'}}) = p(Y = 1|x^{pre})$.

B.6 Fine tuning decision thresholds to yield exact acceptance guarantees

We conjecture that in order to yield exact acceptance rates, to tune false positive, false negative rates or more generally to calibrate the decision function, the decision threshold can be adapted to the individual or subgroup.
Therefore, the estimation of the prediction distribution is required. As follows we sketch the estimation procedure, but leave a more detailed analysis for future work.

B.6.1 Acceptance guarantees for individualized decision thresholds: In order to estimate $p(h^{*,ind}|x^{pre}, do(a))$, we need to estimate the counterfactual distribution of all covariates (and not only of ancestors of y).
Abduction for children of Y is a further challenge. Since Y is unknown, the noise terms for the children of Y become dependent with u_Y. Therefore the abducted distribution of U_j for $j \in ch(Y)$ must be constructed conditional on u_y. We leave a detailed analysis for future work.

B.6.2 Acceptance guarantees for subgroup-based decision thresholds: If the assumptions for subpopulation-based CR are fulfilled, meaning that causal sufficiency holds (despite not observing Y), then we can estimate $p(x|do(a), x^{pre}_{G_a})$ as explained in [22]. Otherwise, approaches that bound the intervention effect despite unobserved confounding must be developed, see e.g. [50] for first steps.

C MODEL MULTIPLICITY EXPERIMENT

In this Section, we analyze the robustness of CR and MCR with respect to model multiplicity. Model multiplicity refers to the problem that many (mechanistically) different models may all constitute optimal predictors for a given problem. Although all bayes optimal predictors behave similarly within the observational domain, they may behave differently outside the training distribution. As a consequence, recourse recommendations that extrapolate from the data support may not be recognized by other equivalently performant models [34].

We conjecture that in contrast to non-meaningful contrastive explanations, meaningful causal recourse is indeed more robust to model multiplicity, since predictors are at least stable with respect to MCR recommendations (see Proposition 6.6). As a consequence, the acceptance guarantees hold for any optimal predictor. Extrapolation may nevertheless be problematic, as discussed in Section 8.2.

We leave a detailed investigation of the robustness of MCR with respect to model multiplicity for future work. We support the point with a small motivational example.

Therefore, we adapted Example 1 such that refits of unregularized model on the same distribution yield strongly varying results. More specifically, we added two highly correlated variables, that are independent of the remaining variables (Example 3).

Example 3. *Let the SCM contain the three variables V, Y, S as well as the respective structural equations as defined in Example 1. In addition, we introduce two variables X_1 and X_2 that are unrelated to V, Y, S, but highly correlated with each other.*

$$X_1 = U_1, \qquad U_1 \sim Bern(0.86) \tag{13}$$

$$X_2 = X_1 + U_2 \qquad U_2 \sim Bern(0.001) \tag{14}$$

The fitted linear model includes the two variables such that they cancel each other out, i.e. $\hat{f}(x) = \beta_0 + \beta_S x_S + \beta_V x_V + \beta_1 x_1 + beta_2 x_2$, where $\beta_1 \approx -\beta_2$. On different bootstrap samples of the training dataset, we yield very different parameters. E.g., for the 5-th iteration of the experiment with $\gamma = 0.9$, for the five model multiplicity refits β_2 ranged from 0.1 to -14.9. As such an intervention on X_2 may have a very small or a rather strong effect on the prediction.

As the results presented in Table 3 show, meaningful causal recourse is more robust to model multiplicity than causal recourse.

Table 3. Overview of the results for the Covid admission example. ζ denotes the percentage of recourse-seeking individuals for which a recommendation for the respective confidence level could be made. $\gamma_{\text{obs.}}$ is the percentage of recourse-implementing individuals for which improvement was achieved. $\eta_{\text{obs.}}$ denotes the observed acceptance rate for recourse-implementing individual for the original model and the refit. øcost denotes the average recourse cost among recourse-seeking individuals and øintv. the average number of interventions on causal/non-causal variables.

Multiplicity	ζ		$\gamma_{\text{obs.}}$		$\eta_{\text{obs.}}$		$\eta_{\text{obs.}}^{\text{refit}}$		$\eta_{obs.}^{\text{multipl.}}$		øintv.	
γ/thresh :	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	causes	other
ind. CR	1.00	1.00	0.05	0.05	1.00	1.00	0.09	0.12	0.89	0.98	0.00	1.00
sub. CR	1.00	1.00	0.05	0.05	1.00	1.00	0.09	0.12	0.90	0.98	0.00	1.00
ind. MCR	0.87	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
sub. MCR	0.87	0.00	1.00	-	1.00	-	1.00	-	1.00	-	1.00	0.00

D IMPLEMENTATION

In order to enable a more direct comparison of the CR and MCR targets, we equalize the optimization thresholds for MCR and CR. Given the intuitive interpretation of γ, we fixed the threshold for CR and MCR to γ.[24] In our simulations, we use binary variables with binomial noise. The full specifications of the SCMs are given in Example 1.

We use the evolutionary optimization library deap [8] and NSGA-II [6] to solve the combinatorial optimization problem. All code is publicly available.

The causal recourse [22] framework assumes causal sufficiency, meaning that there are no two endogeneous variables that share an unobserved cause. If the target variable Y is exogeneous then any causal model with more than one endogeneous direct effect of Y violates the assumptions. As such, causal recourse cannot be applied in scenarios where more than one direct effect is observed/modeled. Therefore, a comparison on examples with more than one direct effect is not possible.

[24] A short comment on the choice of a non-adaptive threshold can be found in Appendix B.1.

E PROOFS

E.1 Abduction, proof of Equation 8

We recall the Equation below:

$$p(u_j|x^{pre}) = p(u_j|x_j^{pre}, x_{pa(j)}^{pre}) = \frac{p(x_j^{pre}|u_j, x_{pa(j)}^{pre})p(u_j)}{\int p(x_j^{pre}|u_j, x_{pa(j)}^{pre})p(u_j)} = \frac{\delta_{f_j(x_{pa(j)}^{pre}, u_j), x_j^{pre}} p(u_j)}{\int \delta_{f_j(x_{pa(j)}^{pre}, u_j), x_j^{pre}} p(u_j) du_j}. \tag{8}$$

Proof: According to d-separation, U_j is independent of all other nodes given X_j and $X_{pa(j)}$. Consequently, it holds that

$$p(u_j|x) \overset{d\text{-sep.}}{=} p(u_j|x_j, x_{pa(j)}) \tag{15}$$

$$\overset{\text{Bayes}}{=} \frac{p(u_j, x_j|x_{pa(j)})}{p(x_j|x_{pa(j)})} \tag{16}$$

$$\overset{\text{chain rule}}{=} \frac{p(x_j|u_j, x_{pa(j)})p(u_j|x_{pa(j)})}{\int p(x_j|u_j, x_{pa(j)})p(u_j|x_{pa(j)})du_j} \tag{17}$$

$$\overset{d\text{-sep.}}{=} \frac{p(x_j|u_j, x_{pa(j)})p(u_j)}{\int p(x_j|u_j, x_{pa(j)})p(u_j)du_j} \tag{18}$$

$$\overset{\text{SCM}}{=} \frac{\delta_{f_j(x_{pa(j)}, u_j), x_j} p(u_j)}{\int \delta_{f_j(x_{pa(j)}, u_j), x_j} p(u_j) du_j} \tag{19}$$

where $\delta_{k,l}$ is the Kronecker delta ($\delta_{k,l} = 1$ if $k = l$ an d $\delta_{k,l} = 0$ otherwise).

E.2 Abduction of noise for prediction target, Proof of Proposition B.1

Proposition B.1. *In general, abduction of u_Y can be performed using*

$$p(u_Y|x^{pre}) = \frac{p(u_Y)\prod_{i\in ch(Y)} p(x_i^{pre}|x_{pa(i)}^{pre}, f_Y(x_{pa(Y)}^{pre}, u_Y))}{\int_{\mathcal{U}_Y} p(u'_Y)\prod_{i\in ch(Y)} p(x_i^{pre}|x_{pa(i)}^{pre}, f_Y(x_{pa(Y)}^{pre}, u'_Y))du'_Y}. \tag{9}$$

Given invertible structural equations, a binary target variable Y and given that all endogenous variables except for Y can be observed, u_y can be abducted using

$$p(u_y|x) = \frac{p(u_Y)\prod_{i\in ch(Y)} p(u_i(u_Y))}{\sum_{y'\in\{0,1\}} p(u_Y(y'))\prod_{i\in ch(Y)} p(u_i(u_Y(y')))} \tag{10}$$

where $u_Y(y) = f^{-1}(y; x_{pa(Y)})$ and $u_i(u_y) = f_i^{-1}(x_i; f_Y(x_{pa(Y)}, u_Y), x_{pa(i)})$.

Proof: We prove the more general formula, i.e. Equation 9. Equation 10 directly follows from Equation 9.

$$p(u_Y|x) \overset{\text{Bayes}}{=} \frac{p(x|u_Y)p(u_Y)}{p(x)} \tag{20}$$

$$\overset{\text{Markov fact.}}{=} \frac{p(u_Y)\prod_{i\in ch(Y)} p(x_i|x_{pa(i)}, u_y, x_{pa(Y)})\prod_{i\notin ch(Y)} p(x_i|x_{pa(i)})}{\int_{\mathcal{U}_Y}\prod_{i\in ch(Y)} p(x_i|x_{pa(i)}, u'_Y, x_{pa(Y)})du'_Y\prod_{i\notin ch(Y)} p(x_i|x_{pa(i)})} \tag{21}$$

$$= \frac{p(u_Y)\prod_{i\in ch(Y)} p(x_i|x_{pa(i)}, f_Y(x_{pa(Y)}, u_Y))}{\int_{\mathcal{U}_Y} p(u'_Y)\prod_{i\in ch(Y)} p(x_i|x_{pa(i)}, f_Y(x_{pa(Y)}, u'_Y))du'_Y} \tag{22}$$

The proof uses Bayes theorem and the markov factorization for the joint distribution of bayesian networks/structural causal models.

E.3 Individualized post-recourse prediction, proof of Proposition B.2

Proposition B.2. *In general, the individualized post-recourse predictor can be estimated as*

$$p(y^{post,a}|x^{pre}, x^{post,a}) = \frac{\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})}{\int_{\mathcal{Y}} \left(\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})du\right) dy} \tag{11}$$

Given binary decision problems with invertible structural equations, the individualized post-recourse prediction function reduces to

$$p(y^{post,a}|x^{post,a}, x^{pre}) = \frac{p(U_{-I} = f^{-1}(y^{post,a}, x^{post,a})|x^{pre})}{\sum_{y' \in \{0,1\}} p(U_{-I} = f^{-1}(y', x^{post,a})|x^{pre})}. \tag{12}$$

Proof: It holds that

$$p(y^{post,a}|x^{pre}, x^{post,a}) \overset{\text{def. cond.}}{=} \frac{p(y^{post,a}, x^{post,a}|x^{pre})}{p(x^{post,a}|x^{pre})} \tag{23}$$

$$\tag{24}$$

We can reformulate the conditional distribution $p(y^{post,a}, x^{post,a}|x^{pre})$ as two parts, one that describes the probability of a state of the context given x^{pre}, and one that describes the probability of a post-recourse state $x^{post,a}, y^{post,a}$ given a certain noise state u.

$$p(y^{post,a}, x^{post,a}|x^{pre}) \overset{\text{marginal.}}{=} \int_{\mathcal{U}} p(y^{post,a}, x^{post,a}, u|x^{pre}) \tag{25}$$

$$\overset{\text{chain rule}}{=} \int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u, x^{pre})p(u|x^{pre}) \tag{26}$$

$$\overset{(y,x)^{post} \perp x^{pre}|u}{=} \int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre}). \tag{27}$$

In combination we yield

$$p(y^{post,a}|x^{pre}, x^{post,a}) = \frac{\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})}{\int_{\mathcal{Y}} \left(\int_{\mathcal{U}} p(y^{post,a}, x^{post,a}|u)p(u|x^{pre})du\right) dy} \tag{28}$$

For a setting with invertible structural equations and binary prediction, this reduces to

$$p(y^{post,a}|x^{post,a}, x^{pre}) = \frac{p(y^{post,a}, x^{post,a}|x^{pre})}{p(x^{post,a}|x^{pre})} \tag{29}$$

$$= \frac{p(U_{-I} = f^{-1}(y^{post,a}, x^{post,a})|x^{pre})}{\sum_{y' \in \{0,1\}} p(U_{-I} = f^{-1}(y^{post,a}, x^{post,a})|x^{pre})}. \tag{30}$$

where $-I$ is the index set for variables that have not been intervened on (since the noise terms for the intervened upon variables are isolated variables in the interventional graph).

E.4 Proof of Proposition 6.3

Proposition 6.3. *The expected individualized post-recourse score is equal to the individual pre-recourse improvement probability.* $\gamma(x^{pre}, a)$

$$E[\hat{h}^{*,ind}(x^{post})|x^{pre}, a] = \gamma(x^{pre}, a) = P(Y^{post,a} = 1).$$

Proof: It holds that

$$E[\hat{h}(x^{post})^{*,ind}|x^{pre}, a] = E[E[Y|x^{pre}, x^{post}]|x^{pre}, a] \overset{\text{total exp.}}{=} E[Y|x^{pre}, a] = \gamma(x^{pre}, a).$$

E.5 Observational identifiability of subpopulation meaningfulness, Proposition B.3

Proposition B.3. *Given causal sufficiency of* $\mathcal{G}$*, the conditional interventional distribution is observationally identifiable.*

$$p(y|do(a), x^{pre}_{G_{a'}}) = \int_{X_\Gamma} p(y|x_{pa(Y)}) \prod_{r\in\Gamma} p(x_r|x_{pa(r)}) dx_\Gamma \Big|_{do(a), x^{pre}_{nd(I^a)}}$$

Here $\Gamma := \{r : r \in asc(Y) \wedge r \in d(I)\}$ *is the set of ancestors of* Y *that are descendants of* X_I*.*
For actions a *on non-causes of* Y *(*$I^a_{asc} = \emptyset$*) it holds that* $p(Y = 1|do(X_{I^a} = \theta), x^{pre}_{G_{a'}}) = p(Y = 1|x^{pre})$*.*

Proof: Let I_{asc} be the set of interventions on causes of Y and I_{nasc} be interventions that do not affect Y. Clearly, interventions on non-ancestors of Y do not affect Y. Consequently

$$p(y|do(a), x_{nd(I_{asc})}) = p(y|do(I^a_{asc} = \theta_{asc}), do(I^a_{nasc} = \theta_{nasc}), x_{nd(I_{asc})}) \tag{31}$$
$$= p(y|do(I^a_{asc} = \theta_{asc}), x_{nd(I_{asc})}) \tag{32}$$

If $Y \in nd(I^a)$, then $I^a_{asc} = \emptyset$. Therefore, in such cases, we get

$$p(y|do(I^a_{asc} = \theta), x_{nd(I_{asc})}) = p(y|x) \tag{33}$$

This conditional is observationally identifiable because of the factorization properties of causally sufficient causal models. The following proof is inspired by the proof of Proposition 7 in [22].

We reformulate the joint distribution as

$$p(x, y) = p(y|x_{pa(y)}) \prod_{r\in D} p(x_r|(x, y)_{pa(r)}) \tag{34}$$
$$= p(y|x_{pa(y)}) \prod_{r\in I^a_{asc}} p(x_r|(x, y)_{pa(r)}) \prod_{r\in D\setminus I^a_{asc}} p(x_r|(x, y)_{pa(r)}) \tag{35}$$
$$= p(y|x_{pa(y)}) \prod_{r\in I^a_{asc}} p(x_r|(x, y)_{pa(r)}) \prod_{r\in d(I^a_{asc})} p(x_r|(x, y)_{pa(r)}) \prod_{r\in nd(I^a_{asc})} p(x_r|(x, y)_{pa(r)}) \tag{36}$$
$$\tag{37}$$

Conditioning on the intervention we yield

$$p(y, x_{d(I^a_{asc})}, x_{nd(I^a_{asc})} | do(x_{I^a_{asc}} = \theta_{asc})) \tag{38}$$

$$= p(y|x_{pa(y)}) \prod_{r \in d(I^a_{asc})} p(x_r|(x,y)_{pa(r)}) \prod_{r \in nd(I^a_{asc})} p(x_r|(x,y)_{pa(r)}) \Big|_{do(x_{I^a_{asc}} = \theta_{asc})} \tag{39}$$

Conditioning on $x_{nd(I^a_{asc})}$ we get

$$p(y, x_{d(I^a_{asc})} | do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}) = p(y|x_{pa(y)}) \prod_{r \in d(I^a_{asc})} p(x_r|(x,y)_{pa(r)}) \Big|_{do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}} \tag{40}$$

Since we are only interested in y we can drop all non-ascendants of Y from the factorization[25] to yield

$$p(y, x_{d(I^a_{asc}) \cap anc(Y)} | do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}) \tag{41}$$

$$= p(y|x_{pa(y)}) \prod_{r \in d(I^a_{asc}) \cap anc(Y)} p(x_r|(x,y)_{pa(r)}) \Big|_{do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}} \tag{42}$$

By marginalizing out $\Gamma := d(I^a_{asc}) \cap anc(Y)$ we get

$$p(y|do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}) = \int_{X_\Gamma} p(y|x_{pa(y)}) \prod_{r \in \Gamma} p(x_r|(x,y)_{pa(r)}) dx_\Gamma \Big|_{do(x_{I^a_{asc}} = \theta_{asc}), x_{nd(I^a_{asc})}} \tag{43}$$

E.6 MCR only suggests interventions on causes, Proposition 6.5

Proposition 6.5. *Given nonzero cost for all interventions, then MCR exclusively suggests actions on causes of Y.*

Proof: The goal of meaningful recourse is to improve Y with minimal cost. Only interventions on causes alter Y. Consequently, actions on non-causes of Y would not be suggested by meaningful recourse.

E.7 Intervention stability, proof of Proposition 6.6

Proposition 6.6. *Given causal sufficiency, any set S that allows for an optimal prediction (i.e., $MB(Y) \subseteq S$) is stable with respect to actions on causes of Y.*

Proof: We prove the statement in five steps.

Given causal sufficiency, a graph $\mathcal{G}$ and an endogenous Y, the set of endogeneous direct parents, direct effects and direct parents of effects are the minimal d-separating set $S_{\mathcal{G}}$: Standard result, see e.g. Peters et al. [37], Proposition 6.27.

The set $S_{\mathcal{G}^*}$ in the augmented graph $\mathcal{G}^*$ coincides with $S_{\mathcal{G}}$: The minimal d-separating set contains direct causes, direct effects and direct parents of direct effects. I_l is never a direct cause of X_l. Also, since I_l has no endogenous causes, it cannot be a direct effect. Furthermore, since we restrict interventions to be performed on causes, I_l cannot be a direct parent of a direct effect.

[25] We can think of it as the factorization of a (causally consistent) subgraph of Y and its ancestors.

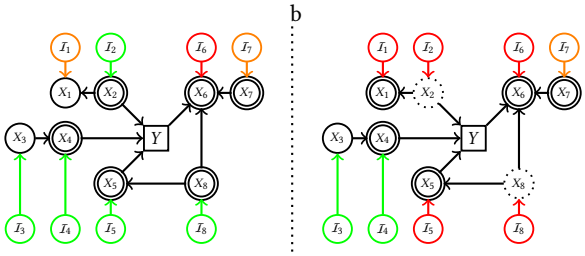

Fig. 2. A schematic drawing illustrating under which interventions $I_1, \ldots, I_8$ the Markov blanket (double circle) is intervention stable. In this setting, we consider the intervention variables to be independent treatment variables: We would like to know how the different actions influence the conditional distribution, irrespective of how likely they are to be applied. Therefore, they are modeled as parent-less variables. Green indicates intervention stability, red indicates no intervention stability. Orange indicates intervention stability of non-causal variables. Dotted variables are not observed. *Left:* Since all endogenous variables are observed, $MB_O(Y)$ is stable w.r.t. interventions on every endogenous cause of Y (Proposition 6.6). *Right:* Unobserved variables (X_2, X_8) open paths between interventions on causes and Y.

$S_{\mathcal{G}}$ *is intervention stable:* As follows, all intervention variables are d-separated from Y in $\mathcal{G}^*$ by $S_{\mathcal{G}}$. Therefore $S_{\mathcal{G}}$ is intervention stable. An example is given in Figure 2.

Then also the markov blanket is intervention stable: Since d-separation implies independence $MB(Y) \subseteq S_{\mathcal{G}}$. Therefore, if $X_T \perp\!\!\!\perp Y | X_{MB(Y)}$ then also $X_T \perp\!\!\!\perp Y | S_{\mathcal{G}}$. If any element $s \in S_{\mathcal{G}}$ it holds that $s \notin MB(Y)$, then it must hold that $X_s \perp\!\!\!\perp Y | X_{MB(Y)}$. Therefore, if $X_T \perp\!\!\!\perp Y | X_{MB(Y)}, X_s$ then also $X_T \perp\!\!\!\perp Y | X_{MB(Y)}$ and therefore any independence entailed by $S_{\mathcal{G}}$ also holds for $MB(Y)$. Since Pfister et al. [38] only require the independence that is implied by d-separation in their invariant conditional proof, the same implication holds for the $MB(Y)$. As follows, $P(Y|X_{MB(Y)})$ is invariant with respect to interventions on any set of endogenous causes.

Then any superset of the markov blanket is intervention stable: We prove the statement by contradiction. The markov blanket d-separates the target variable Y from any other set of variables. If adding a set of variables S_1 to the markov blanket would open a path to any other set of variables S_2, then it would hold that $S := S_1 \cup S_2$ is not d-separated from Y ($P(Y|MB(Y)) = P(Y|MB(Y), S_1, S_2) \neq P(Y|MB(Y), S_1) = P(Y|MB(Y))$)

And the expected prediction equals the improvement probability:

$$E[\hat{h}(x^{post,a})^* | x^{pre}_{nd(I)}, a] = E[E[Y|x^{post,a}]|x^{pre}_{nd(I)}, a] \overset{\text{total exp.}}{=} E[Y|x^{pre}_{nd(I)}, a] \overset{\gamma\text{-meaningful}}{=} \gamma(x^{pre}_{nd(I)}, a).$$

E.8 Proof of proposition 6.7

Proposition 6.7. *Given a binary predictor that is optimal w.r.t. cross-entropy in the pre-recourse observational distribution and stable with respect to interventions on causes, we can equivalently define γ-subpopulation meaningfulness as*

$$E[h^*(\theta, x^{pre}_{G_{a'}}, x^{post}_{d(I^a_{asc})})|do(a), x^{pre}_{G_{a'}}] \geq \gamma.$$

Proof: We can reformulate the expected prediction as

$$E[h^*(\theta, x^{pre}_{G_{a'}}, x^{post}_{d(I^a_{asc})})|do(X_{I^a_{asc}} = \theta), x^{pre}_{G_{a'}}] \tag{44}$$

$$= E[E[Y|\theta, x^{pre}_{G_{a'}}, x^{post}_{d(I^a_{asc})}]|do(X_{I^a_{asc}} = \theta), x^{pre}_{G_{a'}}] \tag{45}$$

$$= E[Y|do(X_{I^a_{asc}} = \theta), x^{pre}_{G_{a'}}] \tag{46}$$

$$= p(Y = 1|do(X_{I^a_{asc}} = \theta), x^{pre}_{G_{a'}}) \tag{47}$$

and yield the interventional subpopulation probability of $Y = 1$. The claim is a direct consequence.

E.9 Proof of Proposition 6.8

Proposition 6.8. *Let x^{pre}_S be the subset of the pre-recourse observed variables that were taken into account to compute MCR, i.e. $S = D$ for individualized recourse and $S = G_{a'}$ for subpopulation-based recourse. For subpopulation-based MCR, we furthermore assume that $p^{pre}(x^{post}) > 0$. For pre-recourse state x^{pre}, a γ-confident action a, the (individualized) optimal predictor h^* and a global decision threshold t the post-recourse acceptance probability η is*

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

where $\gamma(x^{pre}_S, a) = p(Y^{post,a} = 1|x^{pre}_S, a)$.

Proof: The assumption that $p^{pre}(x^{post}) > 0$ is necessary for subpopulation-based MCR since only then we can assume that the model is actually optimal for any input that it receives. The problem is discussed in more detail in Section 8.2.

As follows we denote $\hat{h}^*$ as the random variable indicating the predictions of the post-recourse predictors described in Section 6.

From Propositions 6.3 and 6.6, for both individualized and subpopulation-based post-recourse predictors we know that

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

We decompose the expected prediction

$$\gamma(x^{pre}_S, a) = E[\hat{h}^*|x^{pre}_S, a] \tag{48}$$

$$= E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) + E[\hat{h}^*|\hat{h}^* \leq t]P(\hat{h}^* \leq t) \Big|_{x^{pre}_S, a} \tag{49}$$

$$= E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) + E[\hat{h}^*|\hat{h}^* \leq t](1 - P(\hat{h}^* > t)) \Big|_{x^{pre}_S, a} \tag{50}$$

$$= E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) + E[\hat{h}^*|\hat{h}^* \leq t] - P(\hat{h}^* > t)E[\hat{h}^*|\hat{h}^* \leq t] \Big|_{x^{pre}_S, a} \tag{51}$$

$$= P(\hat{h}^* > t)\Big(E[\hat{h}^*|\hat{h}^* > t] - E[\hat{h}^*|\hat{h}^* \leq t]\Big) + E[\hat{h}^*|\hat{h}^* \leq t] \Big|_{x^{pre}_S, a} \tag{52}$$

$$\tag{53}$$

which can be reformulated to yield the acceptance rate η:

$$\frac{\gamma - E[\hat{h}^*|\hat{h}^* \leq t]}{E[\hat{h}^*|\hat{h}^* > t] - E[\hat{h}^*|\hat{h}^* \leq t]}\Bigg|_{x^{pre}_S, a} = P(\hat{h}^* > t|x^{pre}_S, a) = \eta(x^{pre}_S, a). \tag{54}$$

It holds that $E[\hat{h}^{*,ind}|\hat{h}^* \leq t] = FNR(t)$ and $E[\hat{h}^*|\hat{h}^* > t] = TPR(t)$.

We can show that $E[\hat{h}^*|\hat{h}^* \leq t] \leq t$

$$0 \leq FNR(t|x_S^{pre}, a) = P(Y^{a,post} = 1|h^* \leq t, x_S^{pre}, a) \tag{55}$$
$$= E[Y^{a,post}|h^* \leq t, x_S^{pre}, a] \tag{56}$$
$$= E[E[Y^{a,post}|x^{post,a}]|h^* \leq t, x_S^{pre}, a] \tag{57}$$
$$= E[h^*|h^* \leq t, x_S^{pre}, a] \tag{58}$$
$$\leq t \tag{59}$$

and analog that $1 \geq TPR(t) \geq t$. Therefore

$$\eta(t, x_S^{pre}, a) = \left.\frac{\gamma - FNR(t)}{TPR(t) - FNR(t)}\right|_{x_S^{pre},a} \geq \frac{\gamma(x_S^{pre}, a) - FNR(t)}{1 - FNR(t)} \geq \frac{\gamma(x_S^{pre}, a) - t}{1 - t}. \tag{60}$$