PREDICTION WITHOUT PRECLUSION: RECOURSE VERIFICATION WITH REACHABLE SETS

Avni Kothari[†] UCSF
Bogdan Kulynych[†] EPFL
Tsui-Wei Weng UCSD
Berk Ustun UCSD

ABSTRACT

Machine learning models are often used to decide who receives a loan, a job interview, or a public benefit. Models in such settings use features without considering their *actionability*. As a result, they can assign predictions that are *fixed* – meaning that individuals who are denied loans and interviews are, in fact, *precluded from access* to credit and employment. In this work, we introduce a procedure called *recourse verification* to test if a model assigns fixed predictions to its decision subjects. We propose a model-agnostic approach for recourse verification with *reachable sets* – i.e., the set of all points that a person can reach through their actions in feature space. We develop methods to construct reachable sets for discrete feature spaces, which can certify the responsiveness of *any model* by simply querying its predictions. We conduct a comprehensive empirical study on the infeasibility of recourse on datasets from consumer finance. Our results highlight how models can inadvertently preclude access by assigning fixed predictions and underscore the need to account for actionability in model development.

1 INTRODUCTION

Machine learning models routinely assign predictions to *people* – be it to approve an applicant for a loan [24], a job interview [5, 51], or a public benefit [66, 13, 16]. Models in such applications use features about individuals without accounting for how individuals can change them. In turn, they may assign predictions that are not *responsive* to the *actions* of their decision subjects. In effect, even the most accurate model can assign a prediction that is *fixed* (see Fig. 1).

The responsiveness of machine learning models to our actions is vital to their safety in consumer-facing applications. In applications like content moderation, models *should* assign fixed predictions to prevent malicious actors from circumventing detection by manipulating their features [25, 42, 31]. In lending and hiring, however, predictions should exhibit *some* sensitivity to our actions. Otherwise, models that deny loans and interviews may *preclude access* to credit and employment, thus violating basic rights such as equal opportunity [3] and universal access [8].

In this work, we introduce a formal verification procedure to test the responsiveness of a model's predictions with respect to the actions of its decision subjects. Our procedure – *recourse verification* – is grounded in a stream of work on algorithmic recourse [57, 59, 28]. While much of the work in this area focuses on *recourse provision* – i.e., providing a person with actions to obtain a desired prediction from a model – we focus on recourse verification – i.e., certifying that a model assigns predictions that each person can change. Unlike provision, verification is a model auditing procedure that practitioners can use to flag models that preclude access or promote manipulation.

The key challenge in recourse verification stems from the fact that we must test the sensitivity of a model's predictions with respect to actions rather than arbitrary changes in feature space. In a lending application, for example, actions on a feature such as `years_of_account_history` should set its value to a positive integer and should lead to a commensurate change in other temporal features like `age`. Such constraints are easy to specify for features that are semantically meaningful, but difficult to enforce in methods for recourse provision. To claim that a model assigns a fixed prediction to a point, we must prove that its predictions will not change under any possible action. In practice,

[†]Equal Contribution

Empirical Risk Minimization
most accurate classifier assigns
prediction without recourse to (1, 1)

reapplicant	age $\geq$ 60	Dataset		Best Model	
x_1	x_2	n^-	n^+	$\hat{f}$	$\hat{R}(\hat{f})$
0	0	10	25	+	10
0	1	11	25	+	11
1	0	12	25	+	12
1	1	27	15	−	15

Action Set
actionability constraints that we can
elicit and embed into optimization problems

$$A(x_1, x_2) = \left\{ \begin{array}{c} a_1 \geq 0 \\ a_2 \geq 0 \\ a_1 + x_1 \in \{0,1\} \\ a_2 + x_2 \in \{0,1\} \end{array} \right\}$$

Reachable Sets
sets of feature vectors that allow for
post-hoc recourse verification for **any model**

$R_A(0,0) = \{(0,0),(0,1),(1,0),(1,1)\}$
$R_A(0,1) = \{(0,1),(1,1)\}$
$R_A(1,0) = \{(1,0),(1,1)\}$
$R_A(1,1) = \{(1,1)\}$

Figure 1: Stylized classification task where the most accurate classifier on a dataset with $n^- = 60$ negative examples and $n^+ = 90$ positive examples assigns a prediction without recourse to individuals with $(x_1, x_2) = (1, 1)$. We predict $y = \texttt{repay_loan}$ using two binary features $(x_1, x_2) = (\texttt{reapplicant}, \texttt{age} \geq \texttt{60})$, which can only increase from 0 to 1. We denote the actions on each feature as (a_1, a_2) and show the constraints they must obey in the *action set*. Given any model, we certify the responsiveness of its outputs for (x_1, x_2) by checking its prediction for each point in the *reachable set* $R_A(x_1, x_2)$. In this case, 42 individuals with $(x_1, x_2) = (1, 1)$ are assigned a prediction without recourse as $f(x_1, x_2) = 0$ for all $(x_1, x_2) \in R_A(1, 1)$.

this requires an exhaustive search over a combinatorial subset of actionable feature space. This is a non-trivial computational task – especially for complex models – as we must certify the infeasibility of a combinatorial optimization problem that faithfully encodes a complex decision boundary.

Our main contributions include:

1. We present a model-agnostic approach for recourse verification by constructing a *reachable set* – i.e., a set of all points that a person can attain through their actions in feature space. Given a reachable set, we can certify the responsiveness of a model's predictions by simply querying its predictions over reachable points.

2. We develop fast methods to construct reachable sets for discrete feature spaces. Our methods can construct complete reachable sets for complex actionability constraints, and can support practical verification in model development and deployment.

3. We conduct a comprehensive empirical study on the infeasibility of recourse in consumer finance applications. Our results show how models can assign fixed predictions due to inherent actionability constraints, and demonstrate how existing methods to generate recourse actions and counterfactual explanations may inflict harm by failing to detect such instances.

4. We develop a Python package for recourse verification with reachable sets. Our package includes an API for practitioners to easily specify complex actionability constraints, and routines to test the actionability of recourse actions and counterfactual explanations.

Related Work We focus on a new direction for algorithmic recourse [57, 59, 29] – i.e., as a procedure to certify the responsiveness of a model's predictions with respect to the actions of its decision subjects. Although actionability is a defining characteristic of recourse [see 59], few works mention models may assign fixed predictions as a result of actionability constraints [57, 30, 10]. The lack of awareness stems, in part, from the fact that methods for recourse provision are typically designed and evaluated with simple actionability constraints such as immutability and monotonicity. As we show in Appendix C.5, however, infeasibility only arises once we start to consider actionability constraints that are difficult to handle in algorithm design.

We study recourse verification as a model auditing procedure to safeguard access in applications like lending. In such applications, verification is essential for reliable recourse provision – as it can flag a model that cannot provide recourse to consumers before it is deployed. To this end, our motivation aligns with a stream of work on the robustness of recourse provision with respect to distribution shifts [52, 18, 2, 47, 20], model updates [56, 49], and causal effects [40, 28, 35]. More broadly, recourse verification is a procedure to test the responsiveness of predictions over semantically meaningful features, which may be useful for stress testing for counterfactual invariance [58, 41, 50], certifying adversarial robustness on tabular datasets [39, 27, 23, 60, 42, 31], or designing models that incentivize improvement or deter strategic manipulation [19, 11, 38, 44, 15, 6, 21, 54, 32, 33, 1, 22].

2 RECOURSE VERIFICATION

We consider a standard classification task where we are given a model $f : \mathcal{X} \to \mathcal{Y}$ to predict a label $y \in \mathcal{Y} = \{0, 1\}$ from a vector of features $\boldsymbol{x} = [x_1, \ldots, x_d] \in \mathcal{X}$ in a bounded feature space $\mathcal{X}$. We assume each instance represents a person, and that $f(\boldsymbol{x}) = 1$ represents a desirable *target prediction* – e.g., an applicant with features $\boldsymbol{x}$ will repay a loan within 2 years.

Our goal is to test if each person can obtain a target prediction from the model by changing their features. We represent such changes in terms of actions. Formally, each *action* as a vector $\boldsymbol{a} = [a_1, \ldots, a_d] \in \mathbb{R}^d$ that shifts their features from $\boldsymbol{x}$ to $\boldsymbol{x} + \boldsymbol{a} = \boldsymbol{x}' \in \mathcal{X}$. We refer to the set of all actions from $\boldsymbol{x} \in \mathcal{X}$ as an *action set* $A(\boldsymbol{x})$, and assume that it contains a *null action* $\mathbf{0} \in A(\boldsymbol{x})$. In practice, an action set $A(\boldsymbol{x})$ is a collection of constraints. As shown in Table 1, we can express these constraints in natural language or as equations that we can embed into an optimization problem.

Semantically meaningful features admit hard actionability constraints. In the simplest cases, actionability constraints reflect the way that semantically meaningful features can only be altered in specific ways. For example, a model may use a feature that cannot change (e.g., `age`) or that can only be changed in specific ways (e.g., `has_phd`, which can only be changed from 0 to 1). More generally, constraints may require that changing one feature will induce changes on other features (e.g., changing `married` from 0 to 1 must set `single` from 1 to 0). Such downstream effects can be directional (e.g., changing `retired` from 1 to 0 will set `work_days_per_week` to 0, but not vice-versa), and may affect features that are not themselves actionable (e.g., changing `years_of_account_history` from 0 to 2 will increase `age` by 2 years).

Class	Separable	Discrete	Example	Features	Constraint
Immutability	✓	✗	`n_dependents` should not change	$x_j =$ `n_dependents`	$a_j = 0$
Monotonicity	✓	✗	`reapplicant` can only increase	$x_j =$ `reapplicant`	$a_j \geq 0$
Integrality	✓	✓	`n_accounts` must be positive integer ≤ 10	$x_j =$ `n_accounts`	$a_j \in \mathbb{Z} \cap [0 - x_j, 10 - x_j]$
Categorical Encoding	✗	✓	preserve one-hot encoding of `married`, `single`	$x_j =$ `married` $x_k =$ `single`	$a_l + x_l \in \{0, 1\} \quad x_k + a_k \in \{0, 1\}$ $a_j + x_j + a_k + x_k = 1$
Ordinal Encoding	✗	✓	preserve one-hot encoding of `max_degree_BS`, `max_degree_MS`	$x_j =$ `max_degree_BS` $x_k =$ `max_degree_MS`	$a_j + x_j \in \{0, 1\} \quad x_k + a_k \in \{0, 1\}$ $a_j + x_j + a_k + x_k = 1 \quad a_j + x_j \geq a_k + x_k$
Logical Implications	✗	✓	if `is_employed` $=$ `TRUE` then `work_hrs_per_week` ≥ 0 else `work_hrs_per_week` $= 0$	$x_j =$ `is_employed` $x_k =$ `work_hrs_per_week`	$a_j + x_j \in \{0, 1\}$ $a_k + x_k \in [0, 168]$ $a_j + x_j \leq 168(x_k + a_k)$
Causal Implications	✗	✓	if `years_of_account_history` increases then `age` will increase commensurately	$x_j =$ `years_at_residence` $x_k =$ `age`	$a_j \leq a_k$

Table 1: Examples of deterministic actionability constraints. We show how each constraint can be expressed in natural language and embedded into an optimization problem using standard techniques in mathematical programming [see e.g., 65]. We highlight constraints that are discrete and non-separable because they can only be enforced using special kinds of search algorithms.

Verification as a Feasibility Problem Given a model $f : \mathcal{X} \to \mathcal{Y}$ and a point $\boldsymbol{x} \in \mathcal{X}$ with action set $A(\boldsymbol{x})$, the *recourse provision* task seeks to find an action $\boldsymbol{a} \in A(\boldsymbol{x})$ that minimizes a cost function $\text{cost}(\boldsymbol{a} \mid \boldsymbol{x})$. This task requires finding an optimal solution to an optimization problem as in Eq. (1).

The *recourse verification* task seeks to determine if recourse is infeasible from $\boldsymbol{x}$ – i.e., if a model assigns the same prediction $f(\boldsymbol{x} + \boldsymbol{a}) = 0$ for all actions $\boldsymbol{a} \in A(\boldsymbol{x})$. This task only requires finding a feasible solution to Eq. (1), which can be cast as the optimization problem in Eq. (2).[1]

Recourse Provision

$$\begin{aligned} \min \quad & \text{cost}(\boldsymbol{a} \mid \boldsymbol{x}) \\ \text{s.t.} \quad & f(\boldsymbol{x} + \boldsymbol{a}) = 1 \\ & \boldsymbol{a} \in A(\boldsymbol{x}) \end{aligned} \tag{1}$$

Recourse Verification

$$\begin{aligned} \min \quad & 1 \\ \text{s.t.} \quad & f(\boldsymbol{x} + \boldsymbol{a}) = 1 \\ & \boldsymbol{a} \in A(\boldsymbol{x}) \end{aligned} \tag{2}$$

[1] We set the objective of Eq. (2) to a constant so that any algorithm that solves Eq. (2) will terminate as soon as it has found a feasible solution.

We can write the input and output of a recourse verification method as the function:

$$\mathsf{Recourse}(\boldsymbol{x}, f, A) = \begin{cases} \texttt{Yes}, & \text{if method returns an action } \boldsymbol{a} \in A(\boldsymbol{x}) \text{ such that } f(\boldsymbol{x}+\boldsymbol{a}) = 1 \\ \texttt{No}, & \text{if method proves that } f(\boldsymbol{x}+\boldsymbol{a}) = 0 \text{ for all actions } \boldsymbol{a} \in A(\boldsymbol{x}) \\ \perp, & \text{otherwise} \end{cases}$$

We say that a method for recourse verification *certifies feasibility* from $\boldsymbol{x}$ if it outputs `Yes` and that it *certifies infeasibility* from $\boldsymbol{x}$ if it outputs `No`. In practice, existing methods for recourse provision may return outputs that cannot support either of these claims. For example, they may fail to return an action without having searched exhaustively, or return an "action" that violates actionability constraints. In such cases, we say that the method *abstains* for $\boldsymbol{x}$ and denote its output as $\perp$.

Use Cases Recourse verification is a model auditing procedure to test the *responsiveness* of a model's predictions with respect to the actions of its decision subjects. We can apply this procedure to flag models that are unsafe in different consumer-facing applications by testing responsiveness by choosing an appropriate action set.

Detecting Preclusion. In applications where we would like to safeguard access (e.g., lending), we can flag that a model f precludes access by testing the responsiveness of predictions on points for which $f(\boldsymbol{x}) = 0$. In this case, we would specify an action set that captures indisputable constraints and applies to all individuals. We would claim that the model precludes access if $\mathsf{Recourse}(\boldsymbol{x}, f, A) = \texttt{No}$ for any point such that $f(\boldsymbol{x}) = 0$.

Ensuring Robustness. In applications where we would like to mitigate gaming (e.g., content moderation), we can certify that a model f is vulnerable to adversarial manipulation by testing the responsiveness of its predictions on points for which $f(\boldsymbol{x}) = 0$. In this case, we would specify an action set $A(\boldsymbol{x})$ that encodes a threat model [31] – i.e., actions that let individuals obtain a target prediction by changing spurious features [see 15, 43]. We would claim that the model is vulnerable to manipulation if $\mathsf{Recourse}(\boldsymbol{x}, f, A) = \texttt{Yes}$ for any point such that $f(\boldsymbol{x}) = 0$.

Since these audits apply over points in feature space, we can run verification at different stages of a model lifecycle to minimize the chances of inflicting harm. In model development, we would test if a model assigns fixed predictions to any point in the training data. In deployment, we would repeat this test for new points. In both cases, the procedure would establish that a model assigns fixed predictions, and could support further interventions to mitigate these effects (see Section 4).

Actionability can vary substantially between individuals [see 4, 59]. In principle, we can account for these variations by calling a recourse verification method with *personalized actionability constraints* that we elicit from each decision subject [via, e.g., an interface as in 62]. In practice, we can mitigate harm in consumer-facing applications without eliciting personalized constraints. This is because models may assign fixed predictions as a result of *inherent actionability constraints* – i.e., constraints that apply to all decision subjects and that practitioners could glean from a data dictionary (e.g., constraints that enforce physical limits or preserve a feature encoding). Seeing how inherent constraints represent a subset of personalized constraints, audits with inherent actionability constraints should be used to flag that a model inflicts harm rather than to certify that it is safe.

Algorithm Design Requirements and Pitfalls Methods for recourse verification should be designed to *certify infeasibility*. This is an essential requirement for verification – as it implies that a method can prove that a model's prediction will not change under any possible action. The vast majority of existing methods for recourse provision are ill-suited for verification because cannot certify infeasibility. In practice, these methods will return outputs that are inconclusive or incorrect for recourse verification tasks. We refer to these instances as *loopholes* and *blindspots* and define them below.

Definition 1. Given a recourse verification task for a model f for a point $\boldsymbol{x}$ with the action set $A(\boldsymbol{x})$, we say that a method returns a *loophole* if its output violates actionability constraints.

Methods for recourse provision return loopholes when they search for actions using an algorithm that cannot enforce all actionability constraints in a recourse verification task. For example, methods that search for recourse actions using gradient descent [45] will return loopholes when we must verify recourse with respect to an action set that includes the discrete actionability constraints in Table 1.

Definition 2. Given a recourse verification task for a model f for a point $\boldsymbol{x}$ with the action set $A(\boldsymbol{x})$, we say that a method exhibits a *blindspot* if it fails to find an action for a point where $\mathsf{Recourse}(\boldsymbol{x}, f, A) = \texttt{Yes}$.

Methods for recourse provision output blindspots when they cannot search exhaustively. Common algorithm design patterns that lead to blindspots include: (i) Searching for actions over observed data points [see e.g., 61, 46]; and enforcing actionability by post-hoc *filtering* – i.e., by generating a large collection of changes in feature space and filtering them to enforce actionability [see e.g., 37] – which exhibits blindspots when the generation step is guaranteed to generate all possible actions.

3 Verification with Reachable Sets

We introduce a model-agnostic approach for recourse verification. Our approach constructs *reachable sets* – i.e., sets of feature vectors that obey actionability constraints.

Definition 3. Given a point $\boldsymbol{x}$ and its action set $A(\boldsymbol{x})$, a *reachable set* contains all feature vectors that can be attained using the actions in $A(\boldsymbol{x})$: $R_A(\boldsymbol{x}) := \{\boldsymbol{x} + \boldsymbol{a} \mid \boldsymbol{a} \in A(\boldsymbol{x})\}$.

Given a reachable set $R_A(\boldsymbol{x})$, we can certify that a model f provides recourse to by querying its predictions on each point $\boldsymbol{x} \in R_A(\boldsymbol{x})$. Thus, we can write the verification function as:

$$\mathsf{Recourse}(\boldsymbol{x}, f, R) = \begin{cases} \texttt{Yes}, & \text{if there exists a reachable point } \boldsymbol{x}' \in R_A(\boldsymbol{x}) \text{ s.t. } f(\boldsymbol{x}') = 1 \\ \texttt{No}, & \text{if } f(\boldsymbol{x}') = 0 \text{ for all reachable points } \boldsymbol{x}' \in R_A(\boldsymbol{x}) \\ \perp, & \text{if } f(\boldsymbol{x}') = 0 \text{ for some reachable points } \boldsymbol{x}' \in R \subset R_A(\boldsymbol{x}) \end{cases} \tag{3}$$

Verification with reachable sets has three key benefits:

Model Agnostic Verification: We can use reachable sets to verify recourse for *any model class*. Model agnostic approaches are especially valuable for recourse verification because it is challenging, if not impossible, to develop a model-specific approach for complex model classes such as ensembles and deep neural networks. In particular, this stems from the fact that such method would have to certify the infeasibility of a combinatorial optimization problem that encodes both the model and the actionability constraints. In practice, such problems be prohibitively large to solve in an audit – as we would have to encode a complex decision boundary [see e.g., 55, 48].

Amortization: In a recourse verification task where we have access to a suitable method for recourse verification, we may still wish to verify recourse using a reachable set. This is because, once we have constructed reachable sets, we can use them to verify recourse for as many models as we wish.

Explicit Abstention: In settings where we cannot enumerate a complete reachable set, we can use the interior approximation of the reachable set $R \subset R_A(\boldsymbol{x})$. In this case, the procedure will certify recourse if it can find a feasible action. Otherwise, it will abstain – thus, flagging $\boldsymbol{x}$ as a *potential* prediction without recourse. We can exploit this property to speed up construction through a lazy initialization pattern. For example, rather than constructing a complete reachable set for every training example, we can construct an interior approximation $R \subset R_A(\boldsymbol{x})$. In this setup, we would use the interior approximations to certify feasibility, and only construct the full reachable sets $R = R_A(\boldsymbol{x})$ for points for which we would abstain $\mathsf{Recourse}(\boldsymbol{x}, f, R) = \perp$.

3.1 Construction

In Algorithm 1, we present a procedure to construct a reachable set for a given point by solving an optimization problem of the form:

$$\mathsf{FindAction}(\boldsymbol{x}, A) := \operatorname{argmin} \ \|\boldsymbol{a}\| \text{ s.t. } \boldsymbol{a} \in A(\boldsymbol{x}) \setminus \{\boldsymbol{0}\}.$$

We formulate $\mathsf{FindAction}(\boldsymbol{x}, A)$ as a mixed-integer program that we present in Appendix B. Our formulation can encode all actionability constraints in Table 1 and is designed to be solved in a way that is fast and reliable using an off-the-shelf solver [see e.g., 17, for a list].

Algorithm 1 GetReachableSet

Require: $\boldsymbol{x} \in \mathcal{X}$, feature vector
Require: $A(\boldsymbol{x})$, action set for $\boldsymbol{x}$
$R \leftarrow \{\boldsymbol{x}\}$
$A \leftarrow A(\boldsymbol{x})$
while $\mathsf{FindAction}(\boldsymbol{x}, A)$ is feasible **do**
  $\boldsymbol{a}^* \leftarrow \mathsf{FindAction}(\boldsymbol{x}, A)$
  $R \leftarrow R \cup \{\boldsymbol{x} + \boldsymbol{a}^*\}$
  $A \leftarrow A \setminus \{\boldsymbol{a}^*\}$
Output $R = R_A(\boldsymbol{x})$

Given a point $\boldsymbol{x}$, the procedure enumerates all reachable points by repeatedly solving this problem and removing prior solutions by adding a "no-good" constraint [see e.g., 53]. The procedure stops once $\mathsf{FindAction}(\boldsymbol{x}, A)$ is infeasible – at which point it has enumerated all possible actions and thus reachable points. In practice, the procedure can be stopped when a user-specified stopping condition is met, in which case it would return an interior reachable set $R \subset R_A(\boldsymbol{x})$ that can certify feasibility.

Decomposition Seeing how reachable sets grow exponentially with the number of features, Algorithm 1 may generate an incomplete reachable set that cannot certify infeasibility under reasonable time constraints. We overcome this issue through a decomposition – i.e., by applying Algorithm 1 to subsets of features that can be altered independently for all points $\boldsymbol{x} \in \mathcal{X}$.[2]

Given an action set over d features, we can identify subsets that can be altered independently by inspection. In this way, we can construct the *most granular partition* of features – i.e., a collection of $k \leq d$ feature subsets $\mathcal{M} := \{S_1, \ldots, S_k\}$ such that $A(\boldsymbol{x}) = \prod_{S \in \mathcal{M}} A_S(\boldsymbol{x}_S)$. Given the partition $\mathcal{M}$, we generate reachable sets for each feature subset R_S by calling Algorithm 1 for each $A_S(\boldsymbol{x}_S)$, and recover the full reachable set as $R = \prod_{S \in \mathcal{M}} R_S$.

Decomposition moderates the combinatorial explosion in our setting – making it viable to enumerate reachable sets in practice. This strategy leads to considerable improvement in runtime, as we construct reachable sets for each subset by solving smaller instances of $\mathsf{FindAction}()$, and can construct the reachable set for a single feature subsets without solving a MIP.

3.2 Auditing in Practice

We can verify recourse in model development by constructing a reachable set for each point in a dataset. Once we have constructed reachable sets for each point, we can call recourse verification for any model by querying its predictions on reachable points as per Eq. (3). In practice, the most time-consuming part of our approach stems from the construction of reachable sets. In our implementation, we can achieve a considerable speed up in construction through parallel computing and sharing reachable sets across points. In a task with immutable features, for example, we only need to construct and store a single reachable set for any points $\boldsymbol{x}$ and $\boldsymbol{x}'$ that only differ in terms of immutable feature values.

Given that our approach is designed to verify recourse with prediction queries, it may be time-consuming for models with a resource-intensive inference step. In such cases, we can minimize prediction queries through short-circuiting. In some settings, we can certify that a model provides recourse to a point analytically – i.e., without querying its predictions on reachable points – by applying the result in Theorem 4.

Theorem 4. *Suppose we have a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with n^+ positive examples, and a point $\boldsymbol{x}$ with the reachable set $R \subseteq R_A(\boldsymbol{x})$. In this case, every model $f : \mathcal{X} \to \mathcal{Y}$ will provide recourse to $\boldsymbol{x}$ so long as its false negative rate over $\mathcal{D}$ obeys:*

$$\mathsf{FNR}(f) < \frac{1}{n^+} \sum_{i=1}^{n} \mathbb{1}[\boldsymbol{x}_i \in R \ \wedge \ y_i = 1]$$

Theorem 4 highlights an alternative approach for recourse verification with reachable sets – i.e., we can certify that a model f must provide recourse to a point x so long as the false negative rate does not exceed the density of positive examples in its reachable set. The values can be computed on *any dataset* with labels – be it the training dataset or a separate dataset. In practice, this approach may be useful when working with model classes where prediction queries are time-consuming. In practice, the result requires a dataset that is "dense" enough so that a reachable set for a point contains other labeled examples. When this condition holds, we can certify that a model provides recourse to a point by comparing the false negative rate of f to the prevalence of positive examples in its reachable set.

[2] Formally, we say two subsets of features $S, T \subseteq [d]$ can be altered independently if the action set over $S \cup T$ can be expressed as a product of action sets over S and T for all points $\boldsymbol{x} \in \mathcal{X}$. For example, given the subsets S, T where $S \cup T = [d]$, we write $A(\boldsymbol{x}) = A_S(\boldsymbol{x}_S) \times A_T(\boldsymbol{x}_T)$ for all $\boldsymbol{x} = [\boldsymbol{x}_S, \boldsymbol{x}_T] \in \mathcal{X}$ where $\times$ denotes a Cartesian product.

3.3 Discussion and Extensions

Our methods are designed to construct reachable sets that can be used for recourse verification over discrete feature spaces. In principle, we can construct reachable sets for continuous feature spaces through sampling, but leave this as a topic for future work as it involves a probabilistic guarantee of infeasibility (see Section 5).

Our methods may be useful as a tool to enforce actionability over continuous feature spaces. In particular, we can extend our formulation for $\mathsf{FindAction}()$ as a routine to test the feasibility of changes from existing methods to generate recourse actions and counterfactual explanations. In a case where such methods would suggest that a person can change their prediction by altering their features from $\boldsymbol{x}$ to $\boldsymbol{x}'$, we can test the feasibility such changes with respect to actionability constraints by solving an optimization problem of the form:

$$\mathsf{IsReachable}(\boldsymbol{x}, \boldsymbol{x}', A) \quad := \quad \min \;\; 1 \quad \text{s.t.} \quad \boldsymbol{x} = \boldsymbol{x}' - \boldsymbol{a}, \; \boldsymbol{a} \in A(\boldsymbol{x}). \tag{4}$$

This routine can be used as a way to test for actionability in existing methods via post-hoc filtering. In such cases, the resulting procedure would allow practitioners to flag outputs that violate actionability constraints, and avoid the challenges of detecting loopholes. As we explain in Section 2, it would not be able to certify that recourse is infeasible.

4 Experiments

We present experiments showing how predictions without recourse arise under inherent actionability constraints and how existing methods can fail to detect these instances.

4.1 Setup

We work with three classification datasets from consumer finance, where models that assign fixed predictions would preclude credit access (see Table 2). We process each dataset by encoding categorical attributes and discretizing continuous features. We use the processed dataset to fit a classification model using one of the following model classes: *logistic regression* (LR), *XGBoost* (XGB), and *random forests* (RF). We train each model using an 80%/20% train/test split and tune hyperparameters using standard k-CV. We report the performance of each model in Appendix C.

We specify inherent actionability constraints for each dataset – focusing on identifying indisputable conditions that apply to all individuals (e.g., compliance with physical limits, preserving feature encoding, enforcing deterministic causal effects, and preventing changes to protected attributes). We list the constraints for each dataset in Appendix C. We note that the constraints for all datasets include a mix of separable constraints (e.g. immutability, integrality, monotonicity) as well as non-separable constraints (e.g., encoding presentation, deterministic causal effects).

We construct reachable sets for each point in the dataset using Algorithm 1. We use the reachable sets to identify individuals who are assigned a prediction without recourse by any one of the models. The results from reachable sets reflect the ground-truth feasibility of recourse for each point and each model class. We label our results as Reach and use them to benchmark the reliability of two salient methods to generate recourse actions and counterfactual explanations:

- AR [57], a *model-specific* method that can certify infeasibility for linear classifiers and handle separable actionability constraints,
- DiCE [45], a *model-agnostic* method that handles some separable actionability constraints.

4.2 Results and Discussion

On Predictions without Recourse We summarize our results for each dataset, method, and model class in Table 2. Our results show that models assign fixed predictions under inherent actionability constraints. In practice, individuals who are assigned predictions without recourse may vary drastically across models that perform equally well. Seeing how reachable sets do not change across models, these differences arise from the different decision boundaries of each model.

Dataset	Metrics	LR			XGB			RF		
		Reach	AR	DiCE	Reach	AR	DiCE	Reach	AR	DiCE
	Certifies No Recourse	22.2%	—	—	22.3%		—	31.3%		—
heloc	Outputs Action	77.8%	85.9%	57.6%	77.7%		57.3%	68.7%		49.3%
$n = 5{,}842$ $d = 43$	↳ Loopholes	**0.0%**	41.1%	34.4%	**0.0%**	NA	42.1%	**0.0%**	NA	29.5%
FICO [14]	Outputs No Action	22.2%	14.1%	42.4%	22.3%		42.7%	31.3%		50.7%
	↳ Blindspots	**0.0%**	0.0%	21.0%	**0.0%**		21.1%	**0.0%**		19.8%
	Certifies No Recourse	7.4%	—	—	7.1%		—	28.6%		—
german	Outputs Action	92.6%	91.7%	92.1%	92.9%		93.3%	71.4%		68.0%
$n = 1{,}000$ $d = 36$	↳ Loopholes	**0.0%**	2.2%	16.6%	**0.0%**	NA	23.1%	**0.0%**	NA	24.0%
Dua and Graff [12]	Outputs No Action	7.4%	8.3%	7.9%	7.1%		6.7%	28.6%		32.0%
	↳ Blindspots	**0.0%**	1.3%	0.9%	0.0%		0.0%	**0.0%**		3.4%
	Certifies No Recourse	15.6%	—	—	16.5%		—	0.2%		—
givemecredit	Outputs Action	84.4%	84.4%	79.7%	83.5%		78.5%	99.8%		97.7%
$n = 120{,}268$ $d = 23$	↳ Loopholes	**0.0%**	40.7%	34.6%	**0.0%**	NA	34.7%	**0.0%**	NA	57.7%
Kaggle [26]	Outputs No Action	15.6%	15.6%	20.3%	16.5%		21.5%	0.2%		2.3%
	↳ Blindspots	**0.0%**	0.0%	4.7%	**0.0%**		5.0%	**0.0%**		2.3%

Table 2: Overview of results for all datasets, model classes, and methods. For each dataset and model class, we use Reach to determine individuals who are assigned predictions without recourse. We use these results to reliability of AR and DiCE for recourse verification tasks. We evaluate each method in terms of the percentage of points where it: *certifies no recourse*; *outputs an action*; outputs a *loophole*, i.e., an action that violates actionability constraints; *outputs no action*; exhibits a *blindspot*, i.e., outputs no action when recourse exists. Here, each metric is expressed as a percentage of the points that are assigned a negative prediction by a model.

On Loopholes Our results in Table 2 show how methods to generate recourse actions may output *loopholes* – i.e., actions that violate actionability constraints. In particular, this failure mode affects between 2.2% to 57.7% of individuals across datasets, methods, and models. As we describe in Section 2, methods return loopholes when they cannot enforce *all* actionability constraints in a prediction task. In this case, we note that AR and DiCE can enforce separable actionability constraints. Thus, the loopholes arise from constraints that affect multiple features.

Loopholes reflect silent failures that undermine the benefits of recourse provision and may inflict harm. Consider a consumer finance application where we use AR or DiCE to provide consumers with actions that they can perform to qualify for a loan. In this case, loopholes that are left undetected would lead us to present consumers with recourse actions that are fundamentally impossible. On the heloc dataset, for example, we find that DiCE returns a loophole for 42.1% of individuals who are denied credit by an XGB model. Although some loopholes may be easy to spot through visual inspection or a basic immutability check, this is not always the case. In this case, $\approx 27\%$ of individuals receive an action that alters 5 or more features simultaneously – many of them can only be detected reliably through a programmatic approach that can test if they meet actionability constraints.

On the Illusion of Feasibility Our experiments show that recourse often *appears* to be feasible when methods are only able to enforce simple constraints. We study this effect in Appendix C through an ablation study where we audit models for the heloc dataset under special classes of actionability constraints. Our results show that methods return loopholes for individuals with fixed predictions under simple constraints such as immutability and monotonicity, and that infeasibility only arises once methods can enforce more complex constraints. In this case, we find that LR assigns a prediction without recourse to 22.2% of individuals. If we enforce monotonicity and integrality constraints, however, recourse *appears* to be feasible for $\approx 99\%$ points when using AR.

On Blindspots Our results show how existing methods for recourse provision may return results that are inconclusive or incorrect for verification. In Table 2, we highlight this failure mode by reporting the prevalence of *blindspots* – i.e., the proportion of instances where a method fails to return a recourse action for an individual who has recourse. On the heloc dataset, for example, we find that DiCE fails to find an action for 42.7% of individuals who are denied by the XGB model. In this case, DiCE returns an error message "no counterfactuals found for the given configuration, perhaps try with different parameters..." Our analysis shows that nearly half of these cases (21.1% of 42.7%) correspond to blindspots while the other half are predictions without recourse (21.6%).

Blindspots differ from loopholes in that they represent an "overt" failure mode that is unlikely to inflict harm. In practice, these failures are more likely to stump practitioners who find that a method

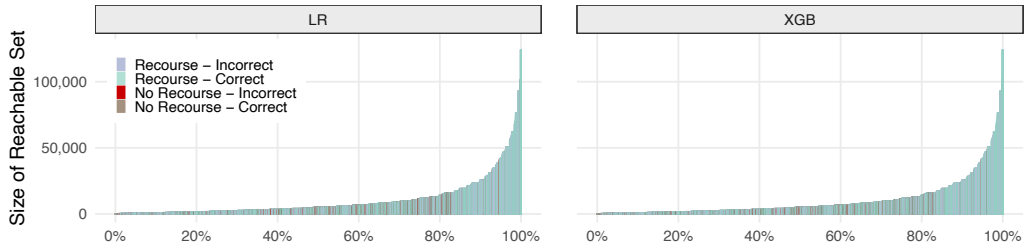

Figure 3: Reachable sets for each point in the `heloc` dataset, ordered from smallest to largest along the x-axis. We show predictions without recourse in red and highlight incorrectly predicted points in darker colors. As shown, predictions without recourse are prevalent across all reachable set sizes and can vary significantly between classifiers.

fails to find a recourse action. In such cases, the key issue is attribution, as practitioners cannot determine whether the failure is (1) a prediction without recourse; (2) the search algorithm used to search for actions; (3) a bug in the recourse provision package; (4) a bug in their code. More broadly, the results highlight the value of designing methods that can certify infeasibility – as methods that could certify infeasibility in case could point to such a result as evidence of preclusion.

On Interventions to Mitigate Preclusion Our results show how recourse verification can guide heuristic interventions that mitigate preclusion. At a minimum, we can use the results from an audit for model selection – i.e., to choose a model that minimizes preclusion among models that are almost equally accurate. On the `heloc` dataset, for example, we find that RF and XGB models have a test AUC ≈ 0.780 but an 11% difference in the predictions without recourse – thus, we can reduce preclusion without compromising performance by simply choosing to deploy an XGB model over an RF model. Seeing how reachable sets measure preclusion through prediction queries, we can apply this strategy in earlier stages of model development – e.g., we can search for model hyperparameters that minimize preclusion by defining a custom metric to compute the "preclusion rate." In this case, we note that parameters that control the decision boundary of the model can lead to substantial differences in preclusion without compromising training accuracy – as we only need to assign a target prediction to a reachable point rather than the current point. In general, we can mitigate preclusion by defining features to promote actionability or by dropping features that lead to fixed predictions. In contrast to the previous interventions, this may require constructing a new collection of reachable sets at each iteration.

5 Concluding Remarks and Limitations

Our paper highlights how machine learning models can assign fixed predictions as a result of actionability constraints, and describes how such predictions can lead to preclusion in consumer-facing applications such as lending and hiring. Our work proposes to address these failure modes by developing methods for a task called recourse verification.

Recourse verification broadly represents a new direction for research in algorithmic recourse – i.e., as a model auditing procedure to certify the responsiveness of predictions with respect to actions. The methods in this paper are designed for recourse verification over discrete feature spaces and deterministic actionability constraints, but should be extended to address the following limitations:

- Our methods are designed to certify infeasibility with respect to actions over discrete feature spaces – and cannot certify infeasibility with respect to actions on continuous features. In principle, it is possible to construct reachable sets that certify infeasibility over continuous feature spaces by sampling. We leave this topic for future work as it requires a different algorithm and returns a probabilistic guarantee of infeasibility.
- Our methods do not consider probabilistic causal effects – i.e., where actions on a feature *may* incite changes on downstream features in a probabilistic causal model [see, e.g., 30, 34, 10, 35]. Although our methods may be useful to generate actionable interventions in this setting, a reliable method for verification should return a probabilistic guarantee of infeasibility that accounts for potential misspecification in the causal model.

ETHICS STATEMENT

Our work highlights how machine learning models can assign fixed predictions as a result of actionability, and proposes a task called recourse verification to reliably detect such instances.

We study recourse verification as a model auditing procedure that practitioners and auditors can use to detect preclusion in consumer-facing applications such as lending and hiring. The normative basis for the right to access in such applications stems from principles such as equality of opportunity (e.g., in hiring) and universal access (e.g., for basic health insurance). In other words, these are applications where we would want to safeguard access – even if it comes at a cost – because it reflects the kind of society we want to build. In practice, ensuring access may not impose any cost. In lending, for example, lenders only collect labels for consumers who are approved for loans [36, 9, 63, due to selective labeling]. Thus, consumers assigned predictions without recourse cannot generate labels that would signal creditworthiness [7]. In the United States, such effects have cut off credit access for large consumer segments whose creditworthiness is unknown [64, see, e.g., 26M "credit invisibles"].

Our paper primarily studies auditing models in lending and hiring with respect to indisputable constraints that apply to all decision subjects – *inherent actionability constraints*. Our recommendation is based on the fact that such constraints can be gleaned from a data dictionary, that claims surrounding preclusion should be indisputable, and that models may lead to preclusion as a result of such constraints. Given that individuals in these applications will face additional actionability constraints, the results of such an audit should be used to flag models that preclude access rather than to certify that models are safe.

In applications where elicitation is possible, our proposed approach can support a number of practices to handle assumptions surrounding actionability in a way that promotes transparency, contestability, and participatory design. In particular, individuals can express their constraints in natural language – allowing stakeholders to scrutinize and contest them even without technical expertise in machine learning. In the event that stakeholders disagree on actionability constraints, we recommend determining if their disagreements affect claims of infeasibility through an ablation study. In such cases, we can run verification using the subset of "consensus constraints" that all stakeholders agree on. In this worst case, we may still find that models lead to preclusion since the "consensus constraints" will always contain inherent constraints.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) under grant IIS-2313105.

REFERENCES

[1] Alon, Tal, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1774–1781, 2020.

[2] Altmeyer, Patrick, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie Deursenvan , and Cynthia Liem. Endogenous macrodynamics in algorithmic recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning*.

[3] Arneson, Richard. Equality of Opportunity. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2015 edition, 2015.

[4] Barocas, Solon, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

[5] Bogen, Miranda and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December*, 7, 2018.

[6] Chen, Yatong, Jialu Wang, and Yang Liu. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355*, 2020.

[7] Chien, Jennifer, Margaret Roberts, and Berk Ustun. Learning through Recourse under Censoring. *NeurIPS Workshop on Learning and Decision-Making with Strategic Feedback*, 2021.

[8] Daniels, Norman. Equity of access to health care: some conceptual and ethical issues. *The Milbank Memorial Fund Quarterly. Health and Society*, pages 51–81, 1982.

[9] De-Arteaga, Maria, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *arXiv preprint arXiv:1807.00905*, 2018.

[10] Dominguez-Olmedo, Ricardo, Amir H Karimi, and Bernhard Schölkopf. On the adversarial robustness of causal algorithmic recourse. In *International Conference on Machine Learning*, pages 5324–5342. PMLR, 2022.

[11] Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70. ACM, 2018.

[12] Dua, Dheeru and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[13] Eubanks, Virginia. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[14] FICO. Explainable machine learning challenge, 2018. URL https://community.fico.com/s/explainable-machine-learning-challenge.

[15] Ghalme, Ganesh, Vineet Nair, Itay Eilat, Inbal Talgam-Cohen, and Nir Rosenfeld. Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR, 2021.

[16] Gilman, Michele E. Poverty lawgorithms: A poverty lawyer's guide to fighting automated decision-making harms on low-income communities. *Data & Society*, 2020.

[17] Gleixner, Ambros, Gregor Hendel, Gerald Gamrath, Tobias Achterberg, Michael Bastubbe, Timo Berthold, Philipp Christophel, Kati Jarck, Thorsten Koch, Jeff Linderoth, et al. Miplib 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3):443–490, 2021.

[18] Guo, Hangzhi, Feiran Jia, Jinghui Chen, Anna Squicciarini, and Amulya Yadav. Rocoursenet: Distributionally robust training of a prediction aware recourse model. *arXiv preprint arXiv:2206.00700*, 2022.

[19] Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. ACM, 2016.

[20] Hardt, Moritz, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. *arXiv preprint arXiv:2302.04262*, 2023.

[21] Harris, Keegan, Hoda Heidari, and Steven Z Wu. Stateful strategic regression. *Advances in Neural Information Processing Systems*, 34:28728–28741, 2021.

[22] Harris, Keegan, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Z Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022.

[23] Hein, Matthias and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NIPS*, 2017.

[24] Hurley, Mikella and Julius Adebayo. Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148, 2016.

[25] Jurgovsky, Johannes, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, and Olivier Caelen. Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, 100:234–245, 2018.

[26] Kaggle. Give Me Some Credit. http://www.kaggle.com/c/GiveMeSomeCredit/, 2011.

[27] Kantchelian, Alex, J. D. Tygar, and Anthony D. Joseph. Evasion and hardening of tree ensemble classifiers. In *ICML*, 2016.

[28] Karimi, Amir-Hossein, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[29] Karimi, Amir-Hossein, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. 2021.

[30] Karimi, Amir-Hossein, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[31] Kireev, Klim, Bogdan Kulynych, and Carmela Troncoso. Adversarial robustness for tabular data through cost and utility awareness. In *Network and Distributed System Security (NDSS) Symposium*, 2023.

[32] Kleinberg, Jon and Manish Raghavan. How Do Classifiers Induce Agents To Invest Effort Strategically? *ArXiv e-prints*, art. arXiv:1807.05307, July 2018.

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

[34] König, Gunnar, Timo Freiesleben, and Moritz Grosse-Wentrup. Improvement-focused causal recourse (ICR). *arXiv preprint arXiv:2210.15709*, 2022.

[35] Kügelgen, Julius von, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594, 2022.

[36] Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[37] Laugel, Thibault, Adulam Jeyasothy, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Achieving diversity in counterfactual explanations: a review and discussion. *arXiv preprint arXiv:2305.05840*, 2023.

[38] Levanon, Sagi and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pages 6243–6253. PMLR, 2021.

[39] Lowd, Daniel and Christopher Meek. Adversarial learning. In *ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647, 2005.

[40] Mahajan, Divyat, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[41] Makar, Maggie, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[42] Mathov, Yael, Eden Levy, Ziv Katzir, Asaf Shabtai, and Yuval Elovici. Not all datasets are born equal: On heterogeneous tabular data and adversarial examples. *Knowledge-Based Systems*, 242:108377, 2022.

[43] Miller, John, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[44] Milli, Smitha, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 230–239, 2019.

[45] Mothilal, Ramaravind K, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[46] Nguyen, Duy, Ngoc Bui, and Viet Anh Nguyen. Feasible recourse plan via diverse interpolation. In *International Conference on Artificial Intelligence and Statistics*, pages 4679–4698. PMLR, 2023.

[47] O'Brien, Andrew and Edward Kim. Toward multi-agent algorithmic recourse: Challenges from a game-theoretic perspective. In *The International FLAIRS Conference Proceedings*, volume 35, 2022.

[48] Parmentier, Axel and Thibaut Vidal. Optimal counterfactual explanations in tree ensembles. In *International Conference on Machine Learning*, pages 8422–8431. PMLR, 2021.

[49] Pawelczyk, Martin, Teresa Datta, Johannes Heuvelvan-den , Gjergji Kasneci, and Himabindu Lakkaraju. Algorithmic recourse in the face of noisy human responses. *arXiv preprint arXiv:2203.06768*, 2022.

[50] Quinzan, Francesco, Cecilia Casolo, Krikamol Muandet, Niki Kilbertus, and Yucen Luo. Learning counterfactually invariant predictors. *arXiv preprint arXiv:2207.09768*, 2022.

[51] Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[52] Rawal, Kaivalya, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts. *arXiv preprint arXiv:2012.11788*, 2020.

[53] Sandholm, Tuomas and Robert Shields. Nogood learning for mixed integer programming. In *Workshop on Hybrid Methods and Branching Rules in Combinatorial Optimization, Montréal*, volume 20, pages 21–22, 2006.

[54] Shavit, Yonadav, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. In *International Conference on Machine Learning*, pages 8676–8686. PMLR, 2020.

[55] Tjeng, Vincent and Russ Tedrake. Verifying neural networks with mixed integer programming. In *ICLR*, 2019.

[56] Upadhyay, Sohini, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *arXiv preprint arXiv:2102.13620*, 2021.

[57] Ustun, Berk, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. pages 10–19, 2019. doi: 10.1145/3287560.3287566.

[58] Veitch, Victor, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

[59] Venkatasubramanian, Suresh and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 284–293, 2020.

[60] Vos, Daniël and Sicco Verwer. Efficient training of robust decision trees against adversarial examples. In Meila, Marina and Tong Zhang, editors, *ICML*, 2021.

[61] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. 2017.

[62] Wang, Zijie J, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. Gam coach: Towards interactive and user-centered algorithmic recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[63] Wei, Dennis. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International Conference on Machine Learning*, pages 11035–11046. PMLR, 2021.

[64] Wilhelm, Colin. Big data and the credit gap. https://www.politico.com/agenda/story/2018/02/07/big-data-credit-gap-000630, February 2018.

[65] Wolsey, Laurence A. *Integer programming*. John Wiley & Sons, 2020.

[66] Wykstra, Stephanie. Government's Use of Algorithm Serves up False Fraud Charges. Undark, 6 January, 2020.