A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

Amir-Hossein Karimi  
MPI for Intelligent Systems  
and ETH Zürich

Bernhard Schölkopf  
MPI for Intelligent Systems

Gilles Barthe  
MPI for Security and Privacy

Isabel Valera  
MPI for Intelligent Systems  
and Saarland University

ABSTRACT

Machine learning is increasingly used to inform decision-making in sensitive situations where decisions have consequential effects on individuals’ lives. In these settings, in addition to requiring models to be accurate and robust, socially relevant values such as fairness, privacy, accountability, and explainability play an important role for the adoption and impact of said technologies. In this work, we focus on algorithmic recourse, which is concerned with providing explanations and recommendations to individuals who are unfavourably treated by automated decision-making systems. We first perform an extensive literature review, and align the efforts of many authors by presenting unified definitions, formulations, and solutions to recourse. Then, we provide an overview of the prospective research directions towards which the community may engage, challenging existing assumptions and making explicit connections to other ethical challenges such as security, privacy, and fairness.

KEYWORDS

algorithmic recourse, counterfactuals, contrastive explanations, consequential recommendations, trustworthy machine learning

1 MOTIVATION

Consider the following setting: a 28-year-old female professional working as a software engineer seeks a mortgage to purchase a home. Consider further that the loan-granting institution (e.g., bank) uses a binary classifier and denies the individual the loan based on her attributes. Naturally, in this context, answering the following questions become relevant to the individual:

Q1. Why was I rejected the loan?
Q2. What can I do in order to get the loan in the future?

In the setting of the example above, unless the policy of the bank is relaxed, the individual must expend effort to change their situation to be favourably treated by the decision-making system. Examples such as the above are prevalent not only in finance [20, 133] but also in justice (e.g., pretrial bail) [10], healthcare (e.g., prescription of life-altering medication) [24, 27, 72], and other settings (e.g., hiring) [43, 134, 160] broadly classified as consequential decision-making settings [20, 34, 45, 91]. Given the rapid adoption of automated decision-making systems in these settings, designing models that not only have high objective accuracy but also afford the individual with explanations and recommendations to favourably change their situation is of paramount importance, and even argued to be a legal necessity (GDPR [176]). This is the concern of algorithmic recourse.

Contributions: Our review brings together the plethora of recent works on algorithmic recourse. In reviewing the literature, we identify opportunities to consolidate definitions, construct technical baselines for future comparison, and situate recourse in the broader ethical ML literature. The primary contribution is thus a unified overview of the definitions (§2), formulations (§3) and technical solutions (§4) for computing (nearest) contrastive explanations and (minimal) consequential recommendations, across a broad range of setups. A visual summary is curated and presented in Table 1. Clearly distinguishing between the two recourse offerings above and presenting their often overlooked different technical treatment is a common theme throughout this work. Additionally, we identify a number of prospective research directions which challenge the assumptions of existing setups, and present extensions to better situate recourse in the broader ethical ML literature (§5).

Who is this document for? This document aims to engage different audiences: practitioners (P), researchers (R), and legal scholars (L). §2 builds on the rich literature in the philosophy of science and presents a clear distinction between contrastive explanations and consequential recommendations based on the causal history of events (R, L). Table 1 presents an overview of 50+ technical papers that offer recourse in a broad range of setups (P, R). §3 formulates the recourse problem as constrained optimization and present the variety of constraints needed to model real-world settings (R). §4 presents the difficulty in solving for recourse in general settings, and summarizes existing approximate solutions that trade-off various desirable properties (P, R, L). Finally, §5 argues that recourse should be considered in relation to other ethical ML considerations (L), and outlines future research directions (R).

Scope: To adequately present the material, we provide brief introductions to various related topics, such as explainable machine learning (or XAI), causal and counterfactual inference, and optimization, among others. Our work is not meant to be a comprehensive review of these topics, and merely aims to situate algorithmic recourse in the context of these related fields. For the interested reader, an in-depth discussion on the social, philosophical, cognitive, and psychological aspects of explanations can be found at [21, 35, 125, 174]. For XAI, see [1, 2, 31, 55, 61, 67, 75, 111, 120, 128], for causal and counterfactual inference, see [63, 76, 131, 145, 146, 159, 183], and for optimization, see [52, 137, 166].
2 BACKGROUND

2.1 Recourse definitions

In its relatively young field, algorithmic recourse has been defined as, e.g., “an actionable set of changes a person can undertake in order to improve their outcome” [87]; “the ability of a person to obtain a desired outcome from a fixed model” [171]; or “the systematic process of reversing unfavorable decisions by algorithms and bureaucracies across a range of counterfactual scenarios” [174]. We submit that recourse can be achieved by an affected data-subject if they can understand [56, 87, 177] and accordingly act [92, 93] to alleviate an unfavorable situation, thus exercising temporally-extended agency [174]. Concretely, in the example of the previous section, recourse is offered when the loan applicant is given answers to both questions: provided explanation(s) as to why the loan was rejected (Q1); and offered recommendation(s) on how to obtain the loan in the future (Q2). Below, we describe the often overlooked difference between these questions and the different set of assumptions and tools needed to sufficiently answer each in general settings.

2.2 Recourse and causality

2.2.1 Contrastive Explanations. In an extensive survey of the social science literature, Miller [125] concluded that when people ask “Why P?”, questions they are typically asking “Why P rather than Q?”, where Q is often implicit in the context [81]. A response to such questions is commonly referred to as a contrastive explanation, and is appealing for two reasons. Firstly, contrastive questions provide a ‘window’ into the questioner’s mental model, identifying what they had expected (i.e., Q, the contrast case), and thus, the explanation can be better tuned towards the individual’s uncertainty and gap in understanding [124]. Secondly, providing contrastive explanations may be “simpler, more feasible, and cognitively less demanding” [124] than offering recommendations.

2.2.2 Consequential Recommendations. Providing an affected individual with recommendations amounts to suggesting a set of actions that should be performed to achieve a favourable outcome in the future. In this regard, several works have used contrastive explanations to directly infer actionable recommendations [91, 162, 171, 177], however, Karimi et al. [92] suggest that contrastive explanations may result in feasible or suboptimal actions in general settings. Instead, they suggest that actions should be interpreted as interventions in a causal model of world in which actions will take place [92]. Modelled in this manner, e.g., using a structural causal model (SCM) [144], the down-stream effects of interventions on other variables in the model (e.g., descendants of the intervened-upon variables) can directly be accounted for when recommending actions [21]. Thus a recommended set of actions for recourse, in a world governed by a SCM, are referred to as consequential recommendations [92, 93].

2.2.3 Clarifying terminology: contrastive, consequential, and counterfactual. Now that we understand that recourse explanations are sought contrastively, and that recourse recommendations can be considered as interventions on variables modelled using an SCM, we can rewrite the two recourse questions as:

Q1. What profile would have led to receiving the loan?
Q2. What actions would have led me to develop this profile?

In layman’s terms, recourse is offered when we “inform an individual where they need to get to, and how to get there” [92]. For instance, a diabetic patient may know that a lower blood pressure would reduce their risk of heart failure, but only a doctor with better understanding of the underlying processes could effectively suggest “interventions that will move the patient out of an at-risk group” [177]. Viewed in this manner, both contrastive explanations and consequential recommendations can be classified as a counterfactual [121], in that each considers the alteration of an entity in the history of the event P, where P is the undesired model output. Thus, responses to Q1 (resp. Q2) may also be called counterfactual explanations (resp. counterfactual recommendations), meaning what could have been (resp. what could have been done) [35].

To better illustrate the difference between contrastive explanations and consequential recommendations, we refer to Figure 1. According to Lewis [109, p. 217], “to explain an event is to provide some information about its causal history”. Lipton [110, p. 256] argues that in order to explain why P rather than Q, “we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q”. In the algorithmic recourse setting, because the model outputs are determined by its inputs (which temporally precede the prediction), the input features may be considered as the causes of the prediction. The determining factor in whether one is providing contrastive explanations as opposed to consequential recommendations is thus the level at which the causal history [156] is considered: while providing explanations only requires information on the relationship between the model inputs, \( \{X_i\}_{i=1}^D \), and predictions, \( \hat{Y} \), recommendations require information as far back as the causal relationships among inputs. The reliance on fewer assumptions [110, 125] thus explains why generating recourse explanations is easier than generating recourse recommendations.

In Table 1, we summarize the technical literature and specify the recourse questions that each paper answers (Q1: explanations, or Q2: recommendations). As we shall see, most of the literature focuses on offering explanations, primarily because of the additional assumptions needed to offer recommendations. Furthermore, offering recourse in diverse settings with desirable properties remains an open challenge, which we explore in the following sections.

Figure 1: Variables \( \{X_i\}_{i=1}^D \) capture observable characteristics of an individual which are fed into a blackbox (BB) decision-making system as the mechanism that yields the prediction, \( \hat{Y} \). Contrastive explanations are obtained via interventions on the BB inputs which do not affect other other variables, and can be seen as independent feature shifts. Conversely, consequential recommendations are interventions on a model of the world in which actions will take place and thus rely on accurate knowledge of the SCM or causal graph as a model of nature itself (both of which are non-identifiable from observational data alone [148]). As a result, recourse recommendations are more difficult to generate than recourse explanations.
A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

Table 1: An overview of recourse algorithms for consequential decision-making settings is presented. Ordered chronologically, we summarize the goal, formulation, solution, and properties of each algorithm. With the over-arching goal of providing recourse, each paper either primarily offers contrastive explanations, \( R \), consequential recommendations, \( \Gamma \), or both, depending on the context. Symbols are used to indicate supported settings in the experimental section of the paper (\( \Phi \)), settings that are natural extensions of the presented algorithm\(^1\) (\( \Psi \)), settings that are partially supported\(^2\) (\( \Omega \)), and settings that are not supported (\( \Theta \)). The models cover a broad range of tree-based (TB), kernel-based (KB), differentiable (DF), or other (OT) types. Actionability constraints (unconditional or conditional), plausibility constraints (domain-, density-, and prototypical-consistency), and additional constraints (diversity, sparsity) are also explored. While the primary datatypes used in consequential settings are tabular (involving a mix of numeric, binary, categorical, and ordinal variables), we also include additional works that generate recourse for non-tabular images and document datasets. Furthermore, papers that present analysis of such properties as optimality (opt.), coverage (cov.), and run-time complexity (rtm.) are specified in the table. Finally, we make note of those papers that provide open-source implementations of their algorithm.

Executive summary: the vast majority of the recourse literature has focused on generating contrastive explanations rather than consequential recommendations (c.f., §2). Differentiable models are the most widely supported class of models, and many constraints are only sparsely supported (c.f., §3). All tools generate solutions that to some extent trade-off desirable requirements, e.g., optimality, perfect coverage, efficient run-time, and access (c.f., §4), resulting in a lack of unifying comparison (c.f., §5). This table does not aim at serving to rank or be a qualitatively comparison of surveyed methods, and one has to exercise caution when comparing different setups. As a systematic organization of knowledge, we believe the table may be useful to practitioners looking for methods that satisfy certain properties, and useful for researchers who want to identify open problems and methods to further develop.

![Table](image)

\(^1\) E.g., a model-agnostic query-based algorithm supports all models, even if the experiments were only conducted on a subset of those presented in the table.

\(^2\) E.g., an algorithm may support numeric and binary variables, but not categorical.
3 FORMULATION

Given a fixed predictive model, commonly assumed to be a binary classifier, \( h : X \to \{0, 1\} \), with \( X = X_1 \times \cdots \times X_D \), we can define the set of contrastive explanations for a (factual) input \( \mathbf{x}^f \in X \) as:

\[
\mathcal{E} := \{ \mathbf{x}^{CF} \in \mathcal{P}(X) \mid h(\mathbf{x}^{CF}) \neq h(\mathbf{x}^f) \}.
\]

Here, \( \mathcal{P}(X) \subseteq X \) is a plausible subspace of \( X \), according to the distribution of training data (see §3.3.1). Descriptively, contrastive explanations identify alternative feature combinations (in nearby worlds [108]) that result in a favourable prediction from the fixed model. Assuming a notion of dissimilarity between instances, represented as \( \text{dist}((\cdot), \cdot) : X \times X \to \mathbb{R}_+ \), one can identify nearest contrastive explanations (a.k.a. counterfactual explanations) as follows:

\[
\begin{align*}
\mathbf{x}^{CF} & \in \underset{\mathbf{x}^{CF} \in X}{\arg\min} \quad \text{dist}(\mathbf{x}, \mathbf{x}^{CF}) \\
\text{s.t.} \quad h(\mathbf{x}^{CF}) & \neq h(\mathbf{x})
\end{align*}
\]

\( (1) \)

where \( \delta \) is the perturbation applied independently to the feature vector \( \mathbf{x} \) to obtain the counterfactual instance \( \mathbf{x}^{CF} \). As discussed in §2, although contrastive explanations identify the feature vectors that would achieve recourse, in general, the set of actions that would need to be performed to realize these features are not directly implied from the explanation [92]. Thus, a consequential recommendation for (factual) input \( \mathbf{x}^f \in X \) is defined as:

\[
\mathcal{R} := \{ \mathbf{a} \in \mathcal{A}(\mathbf{x}^f) \mid h(\mathbf{x}^{SCF}(\mathbf{a}; \mathbf{x}^f)) \neq h(\mathbf{x}^f) \}.
\]

Here, \( \mathcal{A}(\mathbf{x}^f) \) is the set of feasible actions that can be performed by the individual seeking recourse (see §3.3.2). Approaching the recourse problem from a causal perspective within the structural causal model (SCM) framework [92], actions are considered as interventions of the form \( \mathbf{a} = \text{do}(\{ \mathbf{x}_i := \mathbf{x}_i^f + \delta_i^{(f)} \}_{i \in f}) \in \mathcal{A}(\mathbf{x}^f) \), and \( \mathbf{x}^{SCF}(\mathbf{a}; \mathbf{x}^f) \) denotes the structural counterfactual of \( \mathbf{x}^f \) had action \( \mathbf{a} \) been performed, all else being equal [144]. Finally, given a notion of cost of actions, capturing the effort expended by the individual as \( \text{cost}(\cdot, \cdot) : \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+ \), one can identify minimal consequential recommendations as follows:

\[
\begin{align*}
\mathbf{a}^{*} \in \arg\min_{\mathbf{a} \in \mathcal{A}(\mathbf{x}^f)} \quad \text{cost}(\mathbf{a}; \mathbf{x}^f) \\
\text{s.t.} \quad h(\mathbf{x}^{CF}) & \neq h(\mathbf{x})
\end{align*}
\]

\( (2) \)

Importantly, the solution of (1) yields a nearest contrastive explanation (i.e., \( \mathbf{x}^{CF} = \mathbf{x}^f + \delta^f \)), with no direct mapping to a set of minimal consequential recommendations [92]. Conversely, solving (2) yields both a minimal consequential recommendation (i.e., \( \mathbf{a}^{*} \)) and a contrastive explanation (i.e., by construction \( \mathbf{x}^{CF} = \mathbf{x}^{SCF}(\mathbf{a}^{*}; \mathbf{x}^f) \)).

Our position is that, with the aim of providing recourse, the primary goal should be to provide minimal consequential recommendations that result in a (not necessarily nearest) contrastive explanation when acted upon. Offering nearest contrastive explanations that are not necessarily attainable through minimal effort is of secondary importance to the individual. In practice, however, due to the additional assumptions needed to solve (2) (specifically for computing \( \mathbf{x}^{SCF} \)), the literature often resorts to solving (1).

In the remainder of this section, we provide an overview of the objective function and constraints used in (1) and (2), followed by a description of the datatypes commonly used in recourse settings. Finally, we conclude with related formulations. Then in Section 4, we review the tools used to solve the formulations defined here.

3.1 Optimization objective

Generally, it is difficult to define dissimilarity (\( \text{dist} \)) between individuals, or cost functions for effort expended by individuals. Notably, this challenge was first discussed in the algorithmic fairness literature [60, 85], and later echoed throughout the algorithmic recourse community [21, 174]. In fact, "the law provides no formal guidance as to the proper metric for determining what reasons are most salient" [161]. In spite of this, existing works have presented various ad-hoc formulations with sensible intuitive justifications or practical allowance, which we review below.

3.1.1 On dist. Wachter et al. [177] define \( \text{dist}(\mathbf{x}, \mathbf{x}^f) \) as the Manhattan distance, weighted by the inverse median absolute deviation (MAD):

\[
\text{dist}(\mathbf{x}, \mathbf{x}^f) = \frac{\sum_{k \in [D]} |x_k - x_k^f|}{\text{MAD}_k}
\]

\( (3) \)

\( \text{MAD}_k = \text{median}_{j \in F}(|X_{jk} - \text{median}_{i \in F}(X_{ijk})) \)

This distance has several desirable properties, including accounting and correcting for the different ranges across features through the MAD heuristic, robustness to outliers with the use of the median absolute difference, and finally, favoring sparse solutions through the use of \( t_1 \) Manhattan distance.

Karimi et al. [91] proposed a weighted combination of \( t_p \) norms as a flexible measure across a variety of situations. The weights, \( \alpha, \beta, \gamma, \zeta \) as shown below, allow practitioners to balance between sparsity of changes (i.e., through the \( \ell_0 \) norm), an elastic measure of distance (i.e., through the \( t_1, t_2 \) norms) [52], and a maximum normalized change across all features (i.e., through the \( \ell_{\infty} \) norm):

\[
\begin{align*}
\text{dist}(\mathbf{x}, \mathbf{x}^f) = \alpha||\delta||_0 + \beta||\delta||_1 + \gamma||\delta||_2 + \zeta||\delta||_{\infty}
\end{align*}
\]

\( (4) \)

where \( \delta = [\delta_1, \cdots, \delta_D]^T \) and \( \delta_1 : X_1 \times X_2 \to [0, 1] \forall k \in [D] \). This measure accounts for the variability across heterogeneous features (see §3.5) by independently normalizing the change in each dimension according to its spread. Additional weights may also be used to relate emphasize changes in specific variables. Finally, other works aim to minimize dissimilarity on a graph manifold [149], in terms of Euclidean distance in a learned feature space [87, 143], or using a Riemannian metric in a latent space [14, 15].

3.1.2 On cost. Similar to [91], various works explore \( t_p \) norms to measure cost of actions. Ramakrishnan et al. [150] explore \( t_1, t_2 \) norm as well as constant cost if specific actions are undertaken; Karimi et al. [92, 93] minimize the \( t_2 \) norm between \( \mathbf{x}^f \) and the action \( \mathbf{a} \) assignment (i.e., \( ||\mathbf{a}||_2 \)); and Cui et al. [48] explore combinations of \( \ell_0, t_1, t_2 \) norms over a user-specified cost matrix. Encoding individual-dependent restrictions is critical, e.g., obtaining CREDIT is more difficult for an foreign students compared to local resident.

Beyond \( t_p \) norms, the work of Ustun et al. [171] propose the total- and maximum-log percentile shift measures, to automatically account for the distribution of points in the dataset, e.g.,

\[
\text{cost}(\mathbf{a}; \mathbf{x}^f) = \max_{k \in [D]} |Q_k(x_k^f + \delta_k) - Q_k(x_k^f)|
\]

\( (5) \)
where \( Q_L(\cdot) \) is the CDF of \( x_i \) in the target population. This type of metric naturally accounts for the relative difficulty of moving to unlikely (high or low percentile) regions of the data distribution. For instance, going from a 50 to 55th percentile in school grades is simpler than going from 90 to 95th percentile.

### 3.3 Actionability and plausibility constraints

**3.3.1 Plausibility.** Existing literature has formalized plausibility constraint as one of three categories: (i) domain-consistency; (ii) density-consistency; and (iii) prototypical-consistency. Whereas domain-consistency restricts the counterfactual instance to the range of admissible values for the domain of features [91], density-consistency focuses on likely states in the (empirical) distribution of features [52, 53, 87, 89, 104, 143] identifying instances close to the data manifold. A third class of plausibility constraints selects counterfactual instances that are either directly present in the dataset [149, 181], or close to a prototypical example of the dataset [11, 12, 97, 104, 173].

**3.3.2 Actionability (Feasibility).** The set of feasible actions, \( \mathcal{A}(x^f) \), is the set of interventions \( \text{do}\{X_i := x_i^f + \theta_i\}_{i \in I} \) that the individual, \( x^f \), is able to perform. To determine \( \mathcal{A}(x^f) \), we must identify the set of variables upon which interventions are possible, as well as the pre-/post-conditions that the intervention must satisfy. The actionability of each variable falls into three categories [92, 103]:

- I. actionable (and mutable), e.g., bank balance;
- II. mutable but non-actionable, e.g., credit score;
- III. immutable (and non-actionable), e.g., birthplace.

Intuitively, mutable but non-actionable variables are not directly actionable by the individual, but may change as a consequence of a change to its causal ancestors (e.g., regular debt payment).

Having identified the set of actionable variables, an intervention can change the value of a variable unconditionally (e.g., bank balance can either increase or decrease), or conditionally to a specific value [92] or in a specific direction [171].

### 3.4 Other types of constraints

**Domain Consistency.** A domain-consistency constraint restricts the counterfactual instance to the range of admissible values for the domain of features [91].

**Density Consistency.** A density-consistency constraint focuses on likely states in the (empirical) distribution of features [52, 53, 87, 89, 104, 143].

**Prototypical Consistency.** A prototypical-consistency constraint selects counterfactual instances that are either directly present in the dataset [149, 181], or close to a prototypical example of the dataset [11, 12, 97, 104, 173].

### 3.5 Other types of constraints

**Context-Dependence.** Context-dependence considerations arise in tasks such as gas rebates where the rebate offered to a Teddy Bear depends on whether or not the Teddy Bear is owned by a child.

**Actionability.** Actionability refers to the feasibility of an intervention. Where an intervention is feasible, the outcome of the intervention is predicted with higher accuracy.

**Plausibility.** Plausibility refers to the likelihood of an intervention being possible. Where an intervention is plausible, the outcome of the intervention is predicted with higher accuracy.

**Scope.** Scope refers to the set of variables that can be manipulated. A more inclusive set of variables results in a greater scope, which can enhance the feasibility and plausibility of interventions.
U.S. visa to attend a conference. While it is quite likely for there to be favorably treated foreign students from other countries with similar characteristics (plausible gender, field of study, academic record, etc.), it is impossible for our student to act on their birthplace for recourse (i.e., a plausible explanation but an infeasible recommendation). Conversely, an individual may perform a set of feasible actions that would put them in an implausible state (e.g., small $p(x_{\text{CF}})$; not in dataset) where the model fails to classify with high confidence. Thus, actionability and plausibility constraints may be used in conjunction depending on the recourse setting they describe.

3.4 Diversity and sparsity constraints

3.4.1 Diversity. Diverse recourse is often sought in the presence of uncertainty, e.g., unknown user preferences when defining definition and cost. Approaches seeking to generate diverse recourse generally fall in two categories: i) diversity through multiple runs of the same formulation; or ii) diversity via explicitly appending diversity constraints to prior formulations.

In the first camp, Wachter et al. [177] show that different runs of their gradient-based optimizer over a non-convex model (e.g., multilayer perceptron) results in different solutions as a result of different random seeds. Sharma et al. [162] show that multiple evolved instances of the genetic-based optimization approach can be used as diverse explanations, hence benefitting from not requiring multiple re-runs of the optimizer. Downs et al. [56], Mahajan et al. [117], Pawelczyk et al. [143] generate diverse counterfactuals by passing multiple samples from a latent space that is shared between factual and counterfactual instances through a decoder, and filtering those instances that correctly flip the prediction.

In the second camp, Russell [158] pursue a strategy whereby subsequent runs of the optimizer would prevent changing features in the same manner as prior explanations/recommendations. Karimi et al. [91] continue in this direction and suggest that subsequent recourse should not fall within an $\ell_1$-ball surrounding any of the earlier explanations/recommendations. Cheng et al. [42], Mothilal et al. [132] present a differentiable constraint that maximizes diversity among generated explanations by maximizing the determinant of a (kernel) matrix of the generated counterfactuals.

3.4.2 Sparsity. It is often argued that sparser solutions are desirable as they emphasize fewer changes (in explanations) or fewer variables to act upon (in recommendations) and are thus more interpretable for the individual [123]. While this is not generally accepted [142, 173], one can formulate this requirement as an additional constraint, whereby, e.g., $||\theta||_0 \leq s$, or $||\theta||_0 \leq s$. Formulating sparsity as an additional (hard) constraint, rather than optimizing for it in the objective, grants the flexibility to optimize for a different objective while ensuring that a solution would be sparse.

3.5 Datatypes and encoding

A common theme in consequential decision-making settings is the use of datatypes that refer to real-world attributes of individuals. As a result, datasets are often tabular $\mathbb{R}$, comprising of heterogeneous features, with a mix of numeric (real, integer), binary, categorical, or ordinal variables. Most commonly, the Adult [3], Australian Credit [58], German Credit [17], GiveMeCredit [187], COMPAS [100], HELOC [83], etc. are used, which are highly heterogeneous.

Different feature types obey different statistical properties, e.g., the integer-based heart rate, real-valued BMI, categorical blood type, and ordinal age group differ drastically in their range. Thus, heterogeneous data requires special handling in order to preserve their semantics. A common approach is to encode each variable according to a predetermined strategy, which pre-processes the data before model training and consequently during recourse generation. For instance, categorical and ordinal features may be encoded using one-hot encoding and thermometer encoding, respectively. To preserve the semantics of each variable during recourse generation, we must also ensure that the generated explanations/recommendations result in counterfactual instances that also satisfy the encoding constraints. For instance, Boolean and linear constraints of the form $\sum_j x_{i,j} = 1 \forall x_{i,j} \in \{0, 1\}$ are used to ensure that multiple categories are be simultaneously active, and thermometer-encoded ordinal variables are required to satisfy $x_{i,j} \geq x_{i,j+1} \forall x_{i,j} \in \{0, 1\}$. For a detailed overview, we refer to the work of Nazabal et al. [135].

In addition to tabular data, one may require contrastive explanations for image-based or text-based datasets, as summarized in Table 1. For image-based datasets, the algorithm may optionally operate on the raw data, or on super-pixel or other forms of extracted features, e.g., a hidden representation in a neural network. Text-based datasets are also commonly encoded as vectors representing GloVe [147] or bag-of-words embeddings [155].

3.6 Related formulations

The problem formulation for recourse generation, and specifically that of contrastive explanations, (1), is broadly related to several other problems in data mining and machine learning. For instance, cost-minimizing inverse classification problem [4, 101–103, 105, 118], aim to identify the “minimum required change to a case in order to reclassify it as a member of a different preferred class?” [118]. Actionable knowledge extraction is employed in data mining to suggest “behaviors which render a state of an instance into a preferred state” [57] according to a classifier [36–38, 90, 186]. Finally, adversarial perturbations are small imperceptible changes to the input of a classifier that would alter the output prediction to a false and highly confident region [39, 69, 129, 130, 136, 139, 140, 167]. An additional parallel shared by the above methods is in their assumption of a fixed underlying model. Extensions of the above, in which model designers anticipate and aim to prevent malicious behavior, exist in the strategic classification and adversarial robustness literature.

Whereas there exists strong parallels in their formulations, the differences arise in their intended use-cases and guarantees for the stakeholders involved. For example, as opposed to recourse which aims to build trust with affected individuals, the primary use-case cited in the actionable knowledge extraction literature is to deliver cost-effective actions to maximize profit or other business objectives. Furthermore, whereas a contrastive explanation aims to inform an individual about ways in which their situation would have led to a desirable outcome, an adversarial perturbation aims to fool the human by being imperceptible (e.g., by leaving the data distribution). In a sense, imperceptibility is the anti-thesis of explainability and trust. Finally, building on the presentation in §2, offering consequential recommendations relies on a causal modelling of the world, which is largely ignored by other approaches.
4 SOLUTION

By definition, recourse is offered when an individual is presented with contrastive explanations and consequential recommendations, which can be obtained by solving (1) and (2), respectively. Notably, the objective that is to be minimized (i.e., dist. or cost) may be non-linear, non-convex, or non-differentiable. Furthermore, without restricting the classifier family, the model constraint also need not be linear, monotonic, or convex. Finally, based on individual-/context-specific restrictions, the problem setting may require optimizing over a constrained set of plausible instances, \( P(X) \), or feasible actions, \( \mathcal{A}(x^f) \). Thus, a distance/cost-agnostic and model-agnostic solution with support for plausibility, feasibility, sparsity, and diversity constraints over heterogeneous datasets will in general require complex approaches, trading-off various desirable properties in the process. Below, we discuss the importance of these properties, and provide an overview of utilized solutions.

4.1 Properties

We remark that the optimizer and the resulting solutions should ideally satisfy some desirable properties, as detailed below. In practice, methods typically trade-off optimal guarantee \( \delta^* \), perfect coverage \( \Omega^* \), or efficient runtime \( \tau^* \), and may otherwise require prohibitive access to the underlying data or predictive model.

4.1.1 Optimality. Identified counterfactual instances should ideally be proximal to the factual instance, corresponding to a small change to the individual’s situation. When optimizing for minimal \( \text{dist} \) and cost in (1) and (2), the objective functions and constraints determine the existence and multiplicity of recourse. For factual instance \( x^f \), there may exist zero, one, or multiple\(^3\) optimal solutions and an ideal optimizer should thus identify (at least) one solution (explanation or recommendation, respectively) if one existed, or terminate and return N/A otherwise.

4.1.2 Perfect coverage. Coverage is defined as the number of individuals for which the algorithm can identify a plausible counterfactual instance (through either recourse type), if at least one solution existed \[91\]. Communicating the domain of applicability to users is critical for building trust \[127, 177\].

4.1.3 Efficient runtime. Because explanations/recommendations are likely to be offered in conversational settings \[81, 84, 125, 164, 180\], it is desirable to generate recourse in near-real-time. Thus, algorithms with efficient and interactive run-time are preferred.

4.1.4 Access. Different optimization approaches may require various levels of access to the underlying dataset or model. Access to the model may involve query access (where only the label is returned), gradient access (where the gradient of the output with respect to the input is requested), or class probabilities access (from which one can infer the confidence of the prediction), or complete white-box access (where all the model params are known).

Naturally, there are practical implications to how much access is permissible in each setting, which further restricts the choice of tools. Consider an organization that seeks to generate recourse for their clients. Unless these algorithms are ran in-house by said organization, it is unlikely that the organization would hand over training data, model parameters, or even a non-rate-limited API of their models to a third-party to generate recourse.

4.2 Tools

We consider the richly explored field of optimization \[32, 137, 166\] out of scope of this work and suffice to briefly review the tools used specifically for recourse generation, highlighting their domain of applicability, and relegating technical details to appropriate references. Not only is solving (1) and (2) difficult in general settings \[103\], it has even been shown to be NP-complete or NP-hard in restricted settings, e.g., solving for integer-based variables \[12\], solving for additive tree models \[16, 48, 169\] or neural networks \[94\], and solving for quadratic objectives and constraints \[12, 32, 141\]. Thus, except for exhaustive search over a potentially uncountable set of solutions, most works pursue approximate solutions in restricted settings, trading-off the desirable properties above (see Table 1).

Solutions can be broadly categorized as gradient-based-optimization, model-based, search-based, verification-based, and heuristics-based.

Under differentiability of the objective and constraints, gradient-optimization-based solutions such as FISTA \[26\] are employed \[52, 53, 173\] to find globally optimal solutions under convex Lagrangian, and first-order methods such as (L-)BFGS or projected gradient-descent may be used to identify local optima otherwise. Relatedly, rather than solving recourse for each individual independently, some works pursue a model-based approach, whereby a mapping from factual to counterfactual instances is learned through gradient optimization \[117, 143\]. These methods enjoy efficient runtimes at the cost of coverage loss and poor handling of heterogeneous data.

For non-differentiable settings, branch-and-bound-based \[107\] approaches split the search domain into smaller regions within which a solution may be easier to find. Under linearity of the objectives and constraints, integer linear programming (ILP) algorithms may be used when datatypes are discrete \[48, 171\], and mixed-integer linear programming (MILP) extensions are utilized when some variables are not discrete \[88, 158\]. (M)ILP formulations are solved using powerful off-the-shelf solvers such as CPLEX \[47\] and Gurobi \[138\]. One may also use a combination of iterative binary search and verification tools to obtain solutions to (1) and (2). Here, the problem is reformulated as a constrained satisfaction problem, where the constraint corresponding to the objective \( \text{dist} \) or cost \( \tau \) is updated in each iteration to reflect the bounds in which a solution is obtainable \[91, 92\]. As with (M)ILP, this approach benefits from the existence of off-the-shelf solvers such as Z3 \[50\], CVC4 \[23\], and pySMT \[62\]. The problem may also be cast and solved as program synthesis \[150\] or answer-set programming \[28\].

The methods above typically offer optimality and perfect coverage while relying on white-box access to the fixed model parameters.

A number of heuristics-based approaches are also explored, e.g., finding the shortest path (Dijkstra’s algorithm \[46\]) between \( x^f \) and potential \( x^\text{CF} \) ’s on an empirical graph where edges are placed between similar instances (according to, e.g., Gaussian kernel \[149\]). Finally, genetic-based approaches \[183, 190\] find solutions over different evolutions of candidate solutions according to various heuristics \[22, 49, 74, 97, 162\], and benefit from being model-/datatype-norm-agnostic via only requiring query access to the model.
5 PROSPECTS
In the previous sections we covered the definitions, formulations, and solutions of existing works aiming to offer algorithmic recourse. We showed that generating recourse explanations and recommendations required counterfactual reasoning based on different levels of causal knowledge. Counterfactual reasoning has roots not only in the philosophy of science [81, 82, 108–110, 184], but also in the psychology of human agents [35, 124, 125], and benefits from strong technical foundations [19, 78]. User studies have demonstrated that causal relationships are assessed by evaluating counterfactuals [121], and counterfactual simulation is used to predict future events [64]. Specifically, in the context of XAI, it has been shown that counterfactuals can “make the decisions of incriminable systems intelligible to developers and users” [35], and that people perform better at predicting model behavior when presented with counterfactual instances [98]. Organizations seek to deploy counterfactual-based explanations citing their easy-to-understand nature [29, 30] and GDPR-compliance [177]. Finally, from a practitioner’s standpoint, not only does algorithmic recourse benefit from the widely exercised practice of sharing open-source implementations (see Table 1), various graphical interfaces have also been developed to assist the on-boarding of non-technical stakeholders [42, 68, 181].

There are, however, a number of implicit assumptions made in existing setups, e.g., that the world dynamics are known and do not change, the predictive (supervised) model is fixed, and that changes only arise due to the actions of the individual seeking recourse. Moreover, in the multi-agent settings considered (with e.g., bank and loan seeker), agents are assumed to act truthfully with no gaming or false reporting of features, and agents are aligned in the aim to minimize an agreed-upon objective function. Below, we explore settings in which these assumptions do not hold, and offer potential solutions for extending to more realistic recourse settings.

5.1 Beyond deterministic recourse
In (2), we saw that minimal consequential recommendations are generated subject to the constraint that the counterfactual instance, \( x_{CF} \), is assigned to be the structural counterfactual of \( x^f \) under hypothetical actions \( a \), i.e., \( x_{CF} = x_{SCF}(a, x^f) \) [92]. Computing the structural counterfactual exactly, however, relies on strong assumptions (i.e., the true SCM is an additive noise model and is known). Karimi et al. [93] show that without complete knowledge of the true SCM, counterfactual analysis cannot be done exactly and thus recourse cannot be offered deterministically. Although they then present methods that offer recourse with high probability, they do so under specification of a causally sufficient graph. Future research in this direction may explore less strict settings, perhaps accounting for hidden confounders, or partially observed graphs [9, 44, 168], further adding to the uncertainty of recourse recommendations. Alternatively, sources of stochasticity may enter the recourse process via a non-deterministic decision-making system. For example, it has been demonstrated that for models trained on selective labels, fair and optimal decisions should be made stochastically [25, 96].

5.2 Beyond supervised recourse
In §3.2 we discussed how the standard binary classification setting could be extended to support multi-class classification and regression. Beyond these classic supervised learning settings, an individual may be subject to an automated decision maker that determines a matching of applicants to resources across a population, e.g., kindergarten assignment for children, housing for low-income families. Alternatively, one can expect to generate explanations in more interactive settings, such as for the actions and policies of a reinforcement learning agent [115, 116, 172] or for recommender systems [51, 65]. Finally, explanations may also be generated for time-series data [5, 16, 113], which can be extended to support online data streams and models that change over time [21, 142, 174].

5.3 Beyond individualized recourse
So far, the presented formulations aimed to offer recourse explanations pertaining to a single individual, and assumed that recourse recommendations would be undertaken by that individual. However, it is natural to extend the notion of recourse beyond the data subject in question, or beyond a single individual in the population. An example of the former setting is when the family member of a patient decides on a treatment on their behalf when the patient cannot directly exercise their agency due to incapacitation [174]. One may also consider common cases in judicial processes where a legal counsel represents and seeks recourse for their client which may then be exercised by another fiduciary. In such settings, the formulation of cost and feasibility of actions may need to be adjusted to account for restrictions on both the subject and the actor.

Alternatively, recourse may be achieved through the collective action of a group of people, rather than that of a single individual [92]. For instance, the efforts of social and political activists may culminate in a law change that offers better conditions for a group of individuals. In such settings, a (background) variable which is non-actionable (or incurs high cost) on an individual level may be rendered as actionable on a group level, which may in turn bring down the cost for all members of the group. This example also suggests that background variables may capture contextual information (e.g., economy) that are not characteristics of, but nonetheless affect, the individual. Furthermore, the individual may not have control over these macro variables that change over time and violate the stationarity assumption of the world. Modelling such considerations is an open problem and relegated to future work. Finally, the need to analyze recourse on a sub-population level may arise due to uncertainty in assumptions [93] or as an intentional study of other properties of the system, e.g., fairness [42, 77, 91, 153, 171], which we explore further below.

5.4 On the interplay of recourse and ethical ML
The research questions above have primarily focused on one stakeholder: the affected individual. However, giving the right of recourse to individuals should not be considered in a vacuum and independently of the effect that providing explanations/recommendations may have on other stakeholders (e.g., model deployer and regulators), or in relation to other desirable properties (e.g., fairness, security, privacy, robustness), broadly referred to as ethical ML. We explore this interplay below.

5.4.1 Recourse and fairness. Fairness in ML is a primary area of study for researchers concerned with uncovering and correcting
for potentially discriminatory behavior of machine learning models. In this regard, prior work has informally used the concept of fairness of recourse as a means to evaluate the fairness of predictions. For instance, Ustun et al. [171] look at comparable male/female individuals that were denied a loan and show that a disparity can be detected if the suggested recourse actions (namely, flipsets) require relatively more effort for individuals of a particular sub-group. Along these lines, Sharma et al. [162] evaluate group fairness via aggregating and comparing the cost of recourse (namely, burden) over individuals of different sub-populations. Karimi et al. [91] show that the addition of feasibility constraints (e.g., non-decreasing AGE) that results in an increase in the cost of recourse indicates a reliance of the fixed model on the sensitive attribute AGE, which is often considered as legally and socially unfair. Here we clarify that these notions are distinct and would benefit from a proper mathematical study of the relation between them.

The examples above suggest that evidence of discriminatory recourse (e.g., reliance on RACE) may be used to uncover unfair classification. We show, however, that the contrapositive statement does not hold: consider, for example, a 2-D dataset comprising of two sub-groups (i.e., s ∈ {0, 1}), where p(x|s) = N(0, 10^2). Consider a binary classifier, h : R × S → {0, 1}, where h(x, s) = sign(x). While the distribution of predictions satisfies demographic parity, the (average) recourse actions required of negatively predicted individuals in s = 1 is larger than those in s = 0. Thus, we observe unfair recourse even when the predictions are demographically-fair.

This contradiction (the conditional and contrapositive statement not holding simultaneously) can be resolved by considering a new notion of fairness, i.e., fairness of recourse, that does not imply or is not implied by the fairness of prediction. In this regard, Equalizing Recourse was recently presented by Gupta et al. [77] which offered the first recourse-based (and prediction-independent) notion of fairness. The authors demonstrate that one can directly calibrate for the average distance to the decision boundary to be equalized across different subgroups during the training of both linear and nonlinear classifiers. A natural extension would involve considering the cost of recourse actions, as opposed to the distance to the decision boundary, in flipping the prediction across subgroups. In summary, recourse may trigger new definitions of fairness to ensure non-discriminatory behavior of ML models.

5.4.2 Recourse and robustness. Robustness often refers to our expectation that model outputs should not change as a result of (small) changes to the input. In the context of recourse, we expect that similar individuals should receive similar explanations/recommendations, or that recourse suggestions for an individual should be to some extent invariant to the underlying decision-making system trained on the same dataset [157]. In practice, however, the stability of both gradient-based [8, 54, 66, 122] and counterfactual-based [104, 106] explanation systems has been called into question. Interestingly, it has been argued that it is possible for a model to have robust predictions but non-robust explanations [79], and vice versa [104] (similar to relation between fair predictions and fair recourse). Parallel studies argue that the sparsity of counterfactuals may contribute to non-robust recourse when evaluating explanations generated under different fixed models [142]. Finally, in the case of consequential recommendations, robustness will be affected by assumptions of the causal generative process (see Figure 1). Carefully reviewing assumptions and exploring such robustness issues in more detail is necessary to build trust in the recourse system, and in turn, in the algorithmic decision-making system.

5.4.3 Recourse, security, and privacy. Model extraction concerns have been raised in various settings for machine learning APIs [112, 154, 170, 178]. In such settings, an adversary aims to obtain a surrogate model, f̂, that is similar (e.g., in fidelity) to the target model, f:

\[ f = f̂ = \arg\min_{f̂ \in F} \mathbb{E}_{x \sim p(x)} |\mathcal{L}(f(x), f̂(x))| \]  

Here, an adversary may have access to various model outputs (e.g., classification label [112], class probabilities [170], etc.) under different query budgets (unlimited, rate-limited, etc. [41, 86]). Model extraction may be accelerated in presence of additional information, such as gradients of outputs w.r.t. inputs [126], or contrastive explanations [7]. Related to recourse, and of practical concern [21, 161, 165, 171], is a study of the ability of an adversary with access to a recourse API in extracting a model. Specifically, we consider a setting in which the adversary has access to a prediction API and a recourse API which given a factual instance, x̂, returns a nearest contrastive explanation, x̂CF, using a known distance function, \( \Delta : X \times X \times \mathbb{E}_F \). How many queries should be made to these API to perform a functionally equivalent extraction of \( f(\cdot) \)?

In a first attack strategy, one could learn a surrogate model on a dataset where factual instances and labels (form the training set or randomly sampled) are augmented with counterfactual instances and counterfactual labels. This idea was explored by [7] where they demonstrated that a high fidelity surrogate model can be extracted even under low query budgets. While easy to understand and implement, this attack strategy implicitly assumes that constructed dataset has i.i.d. data, and thus does not make use of the relations between factual and counterfactual pairs.

An alternative attack strategy considers that the model f can be fully represented by its decision boundaries, or the complementary decision regions \( \{R_0, \ldots, R_l\} \). Every contrastive explanation returned from the recourse API informs us that all instance

---

Figure 2: Here we illustrate the model stealing process in 2D and 3D using hypothetical non-linear decision boundaries. "How many optimal contrastive explanations are needed to extract the decision regions of a classifier?" can be formulated as "How many factual balls are needed to maximally pack all decision regions?"
surrounding the factual instance, up to a distance of $\Delta(x^F, x^{\text{CF}})$, share the same class label as $x^F$ according to $f$ (otherwise that instance would be the nearest contrastive explanation). More formally, $f(x^F) = f(x) \forall x \in B^A_x$, where $B^A_x$ is referred to as the $\Delta$-factual ball, centered at $x^F$ and with radius $\Delta(x^F, x^{\text{CF}})$. The model extraction problem can thus be formulated as the number of factual balls needed to maximally pack all decision regions (see Fig. 2):

$$\text{Pr}[\text{Vol}(R_i) - \bigcup_{i=1}^n B^A_{x^F} \subseteq R_i, \text{Vol}(B^A_{x^F}) \leq \varepsilon] \geq 1 - \delta \forall i$$

(7)

As in other extraction settings, $\hat{f}$ can then be used to infer private information on individuals in the training set, to uncover exploitable system vulnerabilities, or for free internal use. Understanding attack strategies may guide recourse policy and the design of defensive mechanisms to hinder the exploitation of such vulnerabilities.

Surprisingly, a model need not be extracted in the sense above to be revealing of sensitive information. Building on the intuition above, we note that a single contrastive explanation informs the data-subject that there are no instances in a certain vicinity (i.e., within $B^A_x$) such that their prediction is different. This information informs the data-subject about, e.g., whether their similar friend was also denied a loan, violating their predictive privacy. Even under partial knowledge of the friend’s attributes, an adversary may use the information about the shared predictions in $B^A_x$ to perform membership inference attacks [163] or infer missing attributes [59]. This problem is worsened when multiple diverse explanations are generated, and is an open problem.

5.4.4 Recourse and manipulation. Although a central goal of recourse is to foster trust between an individual and an automated system, it would be simplistic to assume that all parties will act truthfully in this process. For instance, having learned something about the decision-making process (perhaps through recommendations given to similar individuals), an individual may exaggerate some of their attributes for a better chance of favorable treatment [174]. Trust can also be violated by the recourse-offering party. As discussed earlier, the multiplicity of recourse explanations/recommendations (see §4.1.1) may allow for an organization to cherry-pick “the most socially acceptable explanation out of many equally plausible ones” [21, 79, 99] (see also, fairwashing [6]). In such cases of misaligned incentives, the oversight of a regulatory body, perhaps with random audits of either party, seems necessary. Another solution may involve mandating a minimum number of diverse recourse offerings, which would conflict with security considerations.

5.5 Towards unifying benchmarks

Table 1 presented an overview of the diverse settings in which recourse is sought. Despite the abundance of open-source implementations built on robust tools and working well in their respective settings, a comparative benchmark for recourse is lacking. This problem is exacerbated for consequential recommendations which further rely on assumptions about the causal generative process. In order to make objective progress, however, new contributions should be evaluated against existing methods. Thus, a next step for the community is the curation of an online challenge (e.g., using Kaggle) to benchmark the performance of existing and new methods. To broadly cover practical settings, we envision multiple tracks where authors can submit their generated explanations/recommendations given a fixed classifier, test dataset, and pre-determined dist/cost definition, and be evaluated using the metrics defined in §4.1. Authors may also submit results that satisfy additional actionability, plausibility, and diversity constraints, and be compared as such.

6 CONCLUSIONS

Our work started with a case study, of a 28-year-old female professional who was denied a loan by an automated decision-making system. We aimed to assist this individual in overcoming their difficult situation, i.e., to achieve algorithmic recourse, which was contingent on offering answers to two questions: why, and how? We studied the relation between these questions, and arrived at distinct responses, namely, contrastive explanations and consequential recommendations. Mindful of the goal of recourse, we emphasized minimal consequential recommendations over nearest contrastive explanations as the former directly optimizes for the least effort from the individual. Furthermore, we noted that offering recourse recommendations automatically implied recourse explanations (through simulation of the causal effect of undertaken actions), whereas the converse would not. In reviewing the literature, however, we observed an under-exploitation of consequential recommendations, which we attribute to its reliance on additional assumptions at the level of the causal generative process of the world in which actions take place.

In addition to unifying and precisely defining recourse, we present an overview of the many constraints (e.g., actionability, plausibility, diversity, sparsity) that are needed to model realistic recourse settings. With accompanying illustrative examples, we distinguish between the notions of dist vs. cost, and plausibility vs. actionability (feasibility), whose distinctions are largely ignored in the literature. Throughout, we reiterate that these notions are individual-/context-dependent, and that formulations cannot arise from a technical perspective alone. We summarize the technical literature in Table 1, as a guide for practitioners looking for methods that satisfy certain properties, and researchers that want to identify open problems and methods to further develop.

Finally, we identify a number of prospective research directions which challenge the assumptions of existing setups, and present extensions to better situate recourse in the broader ethical ML literature. The presented examples and discussion serve to illustrate the diversity of stakeholder needs and a tension between the desirable system properties (fairness, security, privacy, robustness) which we seek to offer in addition to recourse. Satisfyingly addressing these needs and navigating the entailed trade-offs may require new definitions and techniques, and relies on the cross-disciplinary expertise of a panel of technical and social scientists. We hope that the presented document may guide further discussion and progress in this direction.

ACKNOWLEDGMENTS

AHK sincerely thanks the senior authors for encouraging him to undertake the daunting task of writing a first draft, which eventually resulted in this manuscript. AHK is also appreciative of Julius von Kügelgen and Umang Bhatt for fruitful discussions on recourse and fairness, and Muhammad Waleed Gondal for helpful feedback throughout, and NSERC and CLS for generous funding support.
A survey of algorithmic recourse: definitions, formulations, solutions, and prospects

REFERENCES

[1] Ashraf Abdal, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI conference on human factors in computing systems. 1–18.

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52160.

[3] Adult data. 1996. https://archive.ics.uci.edu/ml/datasets/adult.

[4] Charu C Aggarwal, Chen Chen, and Jiawei Han. 2010. On Pearl’s hierarchy assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 337–347.

[5] Karthik Akshay, Kalyanmoy Deb, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In Advances in Neural Information Processing Systems. 592–603.

[6] Umang Bhatt, McKeane Andrus, Adrian Walker, and Alice Xiang. 2020. Machine Learning Explainability for External Stakeholders. arXiv preprint arXiv:2007.05408 (2020).

[7] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction for uncertainty quantification in machine-assisted medical decision making. IEEE Transactions on Knowledge and Data Engineering 22, 9 (2020), 1299–1312.

[8] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp). IEEE, 56–57.

[9] Armin Cohn, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In Human factors in computing systems. 2019. IJCAI. 327–333.

[10] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Sun. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In IJCAI. 6276–6282.

[11] Jeff Clute and David J. Leake. 2015. The effects of instruction and context on the clarity of algorithmic explanations. In Human Factors and Computing Systems. 2015. ACM, 1–10.

[12] Mathew Cutler, Stephen J. Dovers, and Andrew B. Paynter. 2018. The reconciliation of multiple models for counterfactual explanations. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency. 289–298.

[13] Longbing Cao, Yan Zhu, and Chengxi Zhang. 2015. Plausible Counterfactual Explanations for Machine Learning Models. arXiv preprint arXiv:2008.08530 (2020).

[14] Longbing Cao and Chengxi Zhang. 2006. Domain-driven actionable knowledge discovery. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 66 (2006), 1–18.
Karimi, Barthe, Schölkopf, Valera

[171] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable recourse in linear classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 10–19.

[172] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. 2018. Contrastive explanations with local foil trees. arXiv preprint arXiv:1806.07470 (2018).

[173] Arnaud Van Looveren and Janis Klaise. 2019. Interpretable Counterfactual Explanations Guided by Prototypes. arXiv preprint arXiv:1907.02584 (2019).

[174] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM.

[175] Tom Vermeire and David Martens. 2020. Explainable Image Classification with Evidence Counterfactual. arXiv preprint arXiv:2004.07511 (2020).

[176] Paul Vogt and Axel Von dem Bussche. [n.d.]. The EU General Data Protection Regulation (GDPR). [n. d.].

[177] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology 31, 2 (2017).

[178] Binxu Wang and Neil Zhenqiang Gong. 2018. Stealing hyperparameters in machine learning. In 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 36–52.

[179] Pei Wang and Nuno Vasconcelos. 2020. SCOUT: Self-aware Discriminant Counterfactual Explanations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8981–8990.

[180] Adrian Weller. 2017. Challenges for transparency. arXiv preprint arXiv:1708.02870 (2017).

[181] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. IEEE transactions on visualization and computer graphics 26, 1 (2019), 56–65.

[182] Adam White and Artur d’Avila Garcez. 2019. Measurable counterfactual local explanations for any classifier. arXiv preprint arXiv:1908.03020 (2019).

[183] Darrell Whitley. 1994. A genetic algorithm tutorial. Statistics and computing 4, 2 (1994), 65–85.

[184] James Woodward. 2005. Making things happen: A theory of causal explanation. Oxford university press.

[185] James Woodward. 2016. Causation and Manipulability. In The Stanford Encyclopedia of Philosophy (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[186] Qiang Yang, Jie Yin, Charles Ling, and Rong Pan. 2006. Extracting actionable knowledge from decision trees. IEEE Transactions on Knowledge and data Engineering 19, 1 (2006), 43–56.

[187] I-Cheng Yeh and Che-hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36, 2 (2009), 2473–2480.

[188] Xin Zhang, Armando Solar-Lezama, and Rishabh Singh. 2018. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In Advances in Neural Information Processing Systems. 4874–4885.

[189] Yunxia Zhao. 2020. Fast Real-time Counterfactual Explanations. arXiv preprint arXiv:2007.05684 (2020).

[190] Eckart Zitzler and Lothar Thiele. 1998. An evolutionary algorithm for multiobjective optimization: The strength pareto approach. TIK-report 43 (1998).