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

Amir-Hossein Karimi
MPI for Intelligent Systems
and ETH Zürich

Gilles Barthe
MPI for Security and Privacy

Bernhard Schölkopf
MPI for Intelligent Systems

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.[1] 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, 185], and for optimization, see [32, 137, 166].

[1] NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA).

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] show that contrastive explanations may result in infeasible 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?

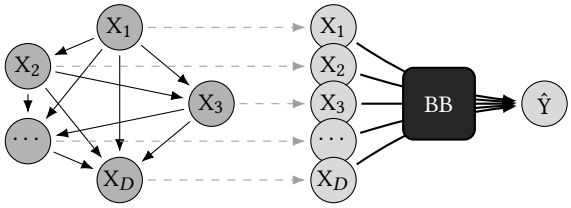

Figure 1: Variables $\{X_i\}_{i=1}^D$ capture observable characteristics of an individual which are fed into a blackbox (BB) decision-making system as the *mechanism* that yields the prediction, $\hat{Y}$. Contrastive explanations are obtained via interventions on the BB inputs which do not affect other 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.

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.

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, $\mathcal{E}$, or consequential recommendations, $\mathcal{R}$. Symbols are used to indicate supported settings in the experimental section of the paper (●), settings that are natural extensions of the presented algorithm[1] (◍), settings that are partially supported[2] (◐), and settings that are not supported (○). The models cover a broad range of tree-based (TB), kernel-based (KB), differentiable (DF), or other (OT) types. Actionability contraints (unconditional or conditional), plausibility constraints (domain-, density-, and prototypical-consistency), and additional constraints (diversity, sprasity) 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 that want to identify open problems and methods to further develop.

Algorithm	Formulation																Solution					
	Goal	Model				Actionability		Plausibility			Extra		Data types			Tools	Access	Properties			Code
		TB	KB	DF	OT	uncond.	cond.	dom.	dens.	proto.	diver.	spar.	▦	📷	🗎			opt.	cov.	rtm.	
(2014.03) SEDC [119]	$\mathcal{E}$	◍	●	◍	◍	○	○	○	○	○	○	○	○	○	●	heuristic	query	●	○	●	○
(2015.08) OAE [48]	$\mathcal{E}$	●	○	○	○	●	◍	○	○	○	○	○	●	○	○	ILP	white-box	●	○	●	○
(2016.05) HCLS [101, 103]	$\mathcal{E}$	●	●	●	◍	●	●	●	●	○	○	○	◐	○	○	grad opt/heuristic	gradient/query	●	○	●	●
(2017.06) Feature Tweaking [169]	$\mathcal{E}$	●	○	○	○	○	○	○	○	○	○	○	●	○	○	heuristic	white-box	○	○	●	●
(2017.11) CF Expl. [177]	$\mathcal{E}$	○	○	●	○	●	○	○	○	○	●	●	◐	○	○	grad opt	gradient	○	○	○	○
(2017.12) Growing Spheres [105]	$\mathcal{E}$	●	●	◍	◍	○	○	○	○	○	○	●	◐	●	○	heuristic	query	○	○	○	○
(2018.02) CEM [52]	$\mathcal{E}$	○	○	●	○	○	○	●	○	○	○	●	◐	●	○	FISTA	class prob.	○	○	○	●
(2018.02) POLARIS [188]	$\mathcal{E}$	○	○	●	○	●	○	○	○	○	○	○	●	●	○	heuristic	gradient	○	○	●	●
(2018.05) LORE [74]	$\mathcal{E}$	●	●	●	◍	○	○	●	○	○	○	○	●	○	○	gen alg + heuristic	query	●	○	●	●
(2018.06) Local Foil Trees [172]	$\mathcal{E}$	●	●	●	◍	○	○	○	○	○	○	○	●	○	○	heuristic	query	○	○	●	●
(2018.09) Actionable Recourse [171]	$\mathcal{E}$	○	○	●	○	●	●	●	○	○	○	○	◐	○	○	ILP	white-box	●	○	○	●
(2018.11) Weighted CFs [71]	$\mathcal{E}$	●	●	●	◍	○	○	○	○	○	○	●	◐	○	○	heuristic	query	○	○	○	○
(2019.01) Efficient Search [158]	$\mathcal{E}$	○	○	●	○	●	○	●	○	○	●	●	●	○	○	MILP	white-box	○	○	○	●
(2019.04) CF Visual Expl. [70]	$\mathcal{E}$	○	○	●	○	○	○	○	○	○	○	○	○	●	○	greedy search	white-box	●	○	●	○
(2019.05) MACE [91]	$\mathcal{E}$	●	◍	●	◍	●	●	●	○	○	●	●	●	○	○	SAT	white-box	●	●	●	●
(2019.05) DiCE [132]	$\mathcal{E}$	○	○	●	○	●	○	◍	○	○	●	○	●	○	○	grad opt	gradient	○	○	○	●
(2019.05) CERTIFAI [162]	$\mathcal{E}$	●	●	●	◍	●	○	●	○	○	●	○	●	●	○	gen alg	query	○	○	○	○
(2019.06) MACEM [53]	$\mathcal{E}$	●	◍	◍	◍	○	○	●	●	○	○	●	●	○	○	FISTA	query	○	○	○	○
(2019.06) Expl. using SHAP [152]	$\mathcal{E}$	●	●	●	●	○	○	○	○	○	○	○	◐	○	○	heuristic	query	○	○	○	●
(2019.07) Nearest Observable [181]	$\mathcal{E}$	◍	◍	●	◍	◍	◍	●	○	●	●	○	●	●	○	brute force	dataset	○	○	○	●
(2019.07) Guided Prototypes [173]	$\mathcal{E}$	◍	◍	●	◍	◍	○	●	●	●	○	●	●	●	○	grad opt/FISTA	gradient/query	○	○	●	●
(2019.07) REVISE [87]	$\mathcal{E}$	○	○	●	○	●	○	●	●	○	○	○	●	●	○	grad opt	gradient	○	○	○	○
(2019.08) CLEAR [182]	$\mathcal{E}$	◍	●	●	◍	○	○	●	●	○	○	○	●	○	○	heuristic	query	○	○	○	●
(2019.08) MC-BRP [113]	$\mathcal{E}$	◍	◍	●	◍	○	○	○	○	○	○	○	◐	○	○	heuristic	query	○	○	○	●
(2019.09) FACE [149]	$\mathcal{E}$	◍	◍	●	◍	●	●	◍	●	●	○	○	●	●	○	graph + heuristic	query	○	●	●	●
(2019.10) Action Sequences [150]	$\mathcal{E}$	○	○	●	○	●	●	○	○	○	○	○	●	●	○	program synthesis	class prob.	●	●	●	●
(2019.10) C-CHVAE [143]	$\mathcal{E}$	◍	●	●	◍	●	○	●	●	○	○	○	●	○	○	grad opt + heuristic	query + gradient	○	●	●	●
(2019.11) OCE [114]	$\mathcal{E}$	●	○	○	●	○	○	○	○	○	○	○	◐	○	○	grad opt + heuristic	white-box	●	●	○	●
(2019.12) Model-based CFs [117]	$\mathcal{E}$	○	○	●	○	●	●	○	●	○	●	○	●	●	○	grad opt	gradient	○	○	●	●
(2019.12) LIME-C/SHAP-C [151]	$\mathcal{E}$	◍	●	●	◍	●	○	○	○	○	○	○	◐	○	●	heuristic	query	●	●	●	●
(2019.12) EMAP [40]	$\mathcal{E}$	●	◍	◍	◍	○	○	○	○	○	○	○	●	●	○	grad opt	dataset/query	○	○	●	●
(2019.12) PRINCE [65]	$\mathcal{E}$	◍	◍	◍	●	○	○	○	○	○	○	○	●	○	○	graph + heuristic	query	●	●	●	●
(2019.12) LowProFool [18]	$\mathcal{E}$	○	○	●	○	○	○	●	○	○	○	○	◐	○	○	grad opt	gradient	●	○	○	○
(2020.01) ABELE [73]	$\mathcal{E}$	●	○	●	○	○	○	○	●	●	○	○	○	●	○	gen alg + heuristic	query + data	○	○	○	●
(2020.02) CEML [11–13]	$\mathcal{E}$	◍	◍	●	●	○	○	○	○	●	○	●	●	●	○	grad opt/heuristic	gradient/query	●	○	●	●
(2020.02) MINT [92]	$\mathcal{R}$	●	◍	●	◍	●	●	●	●	○	●	●	●	○	○	SAT	white-box	●	●	○	●
(2020.03) ViCE [68]	$\mathcal{E}$	◍	●	◍	◍	●	○	○	○	○	○	○	◐	○	○	heuristic	query	○	○	○	●
(2020.03) Plausible CFs [22]	$\mathcal{E}$	○	○	●	○	○	○	○	●	○	○	○	○	●	○	grad opt + gen alg	dataset	●	○	○	○
(2020.04) SEDC-T [175]	$\mathcal{E}$	◍	◍	●	◍	○	○	○	○	○	○	○	○	●	○	heuristic	query	○	○	●	○
(2020.04) MOC [49]	$\mathcal{E}$	●	●	●	●	●	●	●	●	○	●	●	●	○	○	gen alg	query	●	○	○	●
(2020.04) SCOUT [179]	$\mathcal{E}$	○	○	●	○	○	○	○	○	○	○	○	○	●	○	grad opt	gradient	○	○	●	○
(2020.04) ASP-based CFs [28]	$\mathcal{E}$	◍	◍	◍	◍	●	●	○	○	○	○	○	●	○	○	answer-set prog.	query	○	○	○	○
(2020.05) CBR-based CFs [95]	$\mathcal{E}$	◍	◍	●	◍	○	○	●	●	○	○	●	●	○	○	heuristic	query + data	○	○	○	○
(2020.06) Survival Model CFs [97]	$\mathcal{E}$	●	◍	◍	●	●	○	○	○	●	○	○	◐	○	○	gen alg	query	●	○	○	○
(2020.06) Probabilistic Recourse [93]	$\mathcal{R}$	●	◍	●	◍	●	○	●	●	○	◍	○	○	○	○	grad opt/brute force	gradient/query	●	○	○	●
(2020.06) C-CHVAE [142]	$\mathcal{E}$	●	○	●	○	●	○	○	●	○	○	○	◐	○	○	grad opt	gradient	○	○	○	●
(2020.07) FRACE [189]	$\mathcal{E}$	○	○	●	○	○	○	●	●	○	○	○	○	●	○	grad opt	gradient	○	○	●	○
(2020.07) DACE [88]	$\mathcal{E}$	●	○	●	○	●	◍	●	●	○	●	○	●	○	○	MILP	white-box	○	○	●	○
(2020.07) CRUDS [56]	$\mathcal{E}$	◍	◍	●	◍	●	●	○	●	○	●	○	◐	○	○	grad opt	gradient/data	○	●	○	○
(2020.07) Gradient Boosted-based CFs [5]	$\mathcal{E}$	●	○	○	○	○	○	○	○	○	●	○	●	○	○	heuristic	data	○	○	○	●
(2020.08) Gradual Construction [89]	$\mathcal{E}$	○	○	●	○	◍	○	◍	◍	○	○	○	◐	●	●	heuristic	class prob.	○	○	○	○
(2020.08) DECE [42]	$\mathcal{E}$	○	○	●	○	●	●	○	●	○	●	●	●	○	○	grad opt	gradient	○	○	○	●
(2020.08) Time Series CFs [16]	$\mathcal{E}$	◍	○	◍	○	○	○	○	○	○	○	○	◐	○	○	heuristic	query	○	●	●	○
(2020.08) PermuteAttack [80]	$\mathcal{E}$	●	◍	◍	◍	●	○	○	○	○	●	○	●	○	○	gen alg	query	○	○	○	○

[a] ◍ 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.
[b] ◐ 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, $h : \mathcal{X} \to \{0, 1\}$, with $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_D$, we can define the set of *contrastive explanations* for a (factual) input $\boldsymbol{x}^{\mathsf{F}} \in \mathcal{X}$ as $\mathcal{E} \coloneqq \{\boldsymbol{x}^{\mathsf{CF}} \in \mathcal{P}(\mathcal{X}) \mid h(\boldsymbol{x}^{\mathsf{CF}}) \neq h(\boldsymbol{x}^{\mathsf{F}})\}$. Here, $\mathcal{P}(\mathcal{X}) \subseteq \mathcal{X}$ is a *plausible* subspace of $\mathcal{X}$, 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 $\texttt{dist}(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, one can identify *nearest contrastive explanations* (a.k.a. counterfactual explanations) as follows:

$$\begin{aligned} \boldsymbol{x}^{*\mathsf{CF}} \in \underset{\boldsymbol{x}^{\mathsf{CF}} \in \mathcal{X}(\mathcal{P})}{\mathit{argmin}} \quad & \texttt{dist}(\boldsymbol{x}, \boldsymbol{x}^{\mathsf{F}}) \\ \text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{CF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\ & \boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta} \end{aligned} \tag{1}$$

where $\boldsymbol{\delta}$ is the perturbation applied independently to the feature vector $\boldsymbol{x}^{\mathsf{F}}$ to obtain the counterfactual instance $\boldsymbol{x}^{\mathsf{CFE}}$. 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 $\boldsymbol{x}^{\mathsf{F}} \in \mathcal{X}$ is defined as $\mathcal{R} \coloneqq \{a \in \mathcal{A}(\boldsymbol{x}^{\mathsf{F}}) \mid h\big(\boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}})\big) \neq h(\boldsymbol{x}^{\mathsf{F}})\}$. Here, $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$ 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 $\boldsymbol{a} = \text{do}(\{\text{X}_i \coloneqq x_i^{\mathsf{F}} + \theta_i\}_{i \in \mathcal{I}}) \in \mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$, and $\boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}})$ denotes the structural counterfactual of $\boldsymbol{x}^{\mathsf{F}}$ had action $\boldsymbol{a}$ been performed, all else being equal [144]. Finally, given a notion of cost of actions, capturing the effort expended by the individual as $\texttt{cost}(\cdot; \cdot) : \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+$, one can identify *minimal consequential recommendations* as follows:

$$\begin{aligned} \boldsymbol{a}^{*} \in \underset{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{x}^{\mathsf{F}})}{\mathit{argmin}} \quad & \texttt{cost}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}}) \\ \text{s.t.} \quad & h(\boldsymbol{x}^{\mathsf{CF}}) \neq h(\boldsymbol{x}^{\mathsf{F}}) \\ & \boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}}) \end{aligned} \tag{2}$$

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

[2] Relatedly, the counterfactual instance that results from performing optimal actions, $\boldsymbol{a}^{*}$, *need not* correspond to the counterfactual instance resulting from optimally and independently shifting features according to $\boldsymbol{\delta}^{*}$; see [92, prop. 4.1] and [21, Fig. 1]. This discrepancy may arise due to, e.g., minimal recommendations suggesting that actions be performed on an ancestor of those variables that are input to the model.

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 (`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 `dist` as the Manhattan distance, weighted by the inverse median absolute deviation (MAD):

$$\begin{aligned} \texttt{dist}(\boldsymbol{x}, \boldsymbol{x}^{\mathsf{F}}) &= \sum_{k \in [D]} \frac{|\boldsymbol{x}_k - \boldsymbol{x}_k^{\mathsf{F}}|}{\text{MAD}_k} \\ \text{MAD}_k &= \text{median}_{j \in P}(|X_{j,k} - \text{median}_{l \in P}(X_{l,k})|) \end{aligned} \tag{3}$$

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 ℓ_1 Manhattan distance.

Karimi et al. [91] propose a weighted combination of ℓ_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 ℓ_0 norm), an elastic measure of distance (i.e., through the ℓ_1, ℓ_2 norms) [52], and a maximum normalized change across all features (i.e., through the ℓ_∞ norm):

$$\texttt{dist}(\boldsymbol{x}; \boldsymbol{x}^{\mathsf{F}}) = \alpha||\boldsymbol{\delta}||_0 + \beta||\boldsymbol{\delta}||_1 + \gamma||\boldsymbol{\delta}||_2 + \zeta||\boldsymbol{\delta}||_\infty \tag{4}$$

where $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_D]^T$ and $\delta_k : \mathcal{X}_k \times \mathcal{X}_k \to [0, 1] \; \forall \; k \in [D]$. 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 relative 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 ℓ_p norms to measure cost of actions. Ramakrishnan et al. [150] explore ℓ_1, ℓ_2 norm as well as constant cost if specific actions are undertaken; Karimi et al. [92, 93] minimize the ℓ_2 norm between $\boldsymbol{x}^{\mathsf{F}}$ and the action $\boldsymbol{a}$ assignment (i.e., $||\boldsymbol{\theta}||_2$); and Cui et al. [48] explore combinations of ℓ_0, ℓ_1, ℓ_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 ℓ_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.,

$$\texttt{cost}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}}) = \max_{k \in [D]} |Q_k(\boldsymbol{x}_k^{\mathsf{F}} + \theta_k) - Q_k(\boldsymbol{x}_k^{\mathsf{F}})| \tag{5}$$

where $Q_k(\cdot)$ is the CDF of x_k 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.

3.1.3 ***On the relation between*** `dist` ***and*** `cost`. In a world in which changing one variable does not affect others, one can see a parallel between the counterfactual instance of (1), i.e., $\boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}$, and that of (2), i.e., $\boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}}) = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\theta}$. This mirroring form suggests that definitions of dissimilarity between individuals (i.e., `dist`) and effort expended by an individual (i.e., `cost`) *may* be used interchangeably. Following [92], however, we do not consider a general 1-1 mapping between `dist` and `cost`. For instance, in a two-variable system with MEDICATION as the parent of HEADACHE, an individual that consumes more than the recommended amount of medication may not recover from the headache, i.e., higher cost but smaller *symptomatic* distance relative to another individual who consumed the correct amount. Furthermore, while dissimilarity is often considered to be symmetric (i.e., $\texttt{dist}(\boldsymbol{x}^{\mathsf{F}}_A, \boldsymbol{x}^{\mathsf{F}}_B) = \texttt{dist}(\boldsymbol{x}^{\mathsf{F}}_B, \boldsymbol{x}^{\mathsf{F}}_A)$), the effort needed to go from one profile to another need not satisfy symmetry, e.g., spending money is easier than saving money (i.e., $\texttt{cost}(\boldsymbol{a} = \text{do}(\text{X}_\$:= \boldsymbol{x}^{\mathsf{F}}_{A,\$} - \$500); \boldsymbol{x}^{\mathsf{F}}_A) \leq \texttt{cost}(\boldsymbol{a} = \text{do}(\text{X}_\$:= \boldsymbol{x}^{\mathsf{F}}_{B,\$} + \$500); \boldsymbol{x}^{\mathsf{F}}_B)$. These example illustrate that the interdisciplinary community must continue to engage to define the distinct notions of `dist` and `cost`, and such definitions cannot arise from a technical perspective alone.

3.2 Model and counterfactual constraints

3.2.1 ***Model.*** A variety of fixed models have been explored in the literature for which recourse is to be generated. As summarized in Table 1, we broadly divide them in four categories: i) tree-based (TB); ii) kernel-based (KB); iii) differentiable (DF); and iv) other (OT) types (e.g., generalized linear models, Naive Bayes, k-Nearest Neighbors). While the literature on recourse has primarily focused on binary classification settings, most formulations can easily be extended to multi-class classification or regression settings. Extensions to such settings are straightforward, where the model constraint is replaced with $h(\boldsymbol{x}^{\mathsf{CF}}) = k$ for a target class, or $h(\boldsymbol{x}^{\mathsf{CF}}) \in [a, b]$ for a desired regression interval, respectively. Alternatively, soft predictions may be used in place of hard predictions, where the goal may be, e.g., to increase the prediction gap between the highest-predicted and second-highest-predicted class, i.e., $\text{Pred}(\boldsymbol{x}^{\mathsf{CF}})_i - \text{Pred}(\boldsymbol{x}^{\mathsf{CF}})_j$ where $i = argmax_{k \in K} \text{Pred}(\boldsymbol{x}^{\mathsf{CF}})_k$, $j = argmax_{k \in K \setminus i} \text{Pred}(\boldsymbol{x}^{\mathsf{CF}})_k$. In the end, the model constraint representing the change in prediction may be arbitrarily non-linear, non-differentiable, and non-monotone [21], which may limit the applicability of solutions (c.f. §4).

3.2.2 ***Counterfactual.*** The counterfactual constraint depends on the type of recourse offered. Whereas $\boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{F}} + \boldsymbol{\delta}$ in (1) is a linear constraint, computing $\boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}})$ in (2) involves performing the three step abduction-action-prediction process of Pearl et al. [146] and may thus be non-parametric and arbitrary involved. A closed-form expression for deterministically computing the counterfactual in additive noise models is presented in [92], and probabilistic derivations for more general SCMs are presented in [93].

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 conterfactual 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, $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$, is the set of interventions $\text{do}(\{\text{X}_i := x^{\mathsf{F}}_i + \theta_i\}_{i \in \mathcal{I}})$ that the individual, $\boldsymbol{x}^{\mathsf{F}}$, is able to perform. To determine $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$, 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]. Karimi et al. [92] present the following examples to show that the actionable feasibility of an intervention on X_i may be contingent on any number of conditions:

i. pre-intervention value of the intervened variable (i.e., x^{F}_i); e.g., an individual's AGE can only increase, i.e., $[x^{\mathsf{SCF}}_{\text{AGE}} \geq x^{\mathsf{F}}_{\text{AGE}}]$;
ii. pre-intervention value of other variables (i.e., $\{x^{\mathsf{F}}_j\}_{j \subset [d] \setminus i}$); e.g., an individual cannot apply for CREDIT on a temporary VISA, i.e., $[x^{\mathsf{F}}_{\text{VISA}} = \texttt{PERMANENT}] \geq [x^{\mathsf{SCF}}_{\text{CREDIT}} = \texttt{TRUE}]$;
iii. post-intervention value of the intervened variable (i.e., x^{SCF}_i); e.g., an individual may undergo heart surgery (an additive intervention) only if they won't remiss due to sustained SMOKING HABITS, i.e., $[x^{\mathsf{SCF}}_{\text{HEART}} \neq \texttt{REMISSION}]$
iv. post-intervention value of other variables (i.e., $\{x^{\mathsf{SCF}}_j\}_{j \subset [d] \setminus i}$); e.g., an individual may undergo heart surgery only *after* their blood pressure (BP) is regularized due to medicinal intervention, i.e., $[x^{\mathsf{SCF}}_{\text{BP}} = \texttt{O.K.}] \geq [x^{\mathsf{SCF}}_{\text{HEART}} = \texttt{SURGERY}]$

All such feasibility conditions can easily be encoded as Boolean/-logical constraint into $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$ and jointly solved for in the constrained optimization formulations (1), (2). An important side-note to consider is that $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$ is *not* restricted by the SCM assumptions, but instead, by individual-/context-dependent consideration that determine the *form*, *feasibility*, and *scope* of actions [92].

3.3.3 ***On the relation between actionability & plausibility.*** While seemingly overlapping, *actionability* (i.e., $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$) and *plausibility* (i.e., $\mathcal{P}(\mathcal{X})$) are two distinct concepts: whereas the former restrict actions to those that are *possible to do*, the latter require that the resulting counterfactual instance be *possibly true, believable, or realistic*. Consider a Middle Eastern PhD student who is denied a

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 $p(\boldsymbol{x}^{\mathsf{CF}})$; 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 `dist` 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 benefiting 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 ℓ_p-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., $||\boldsymbol{\delta}||_0 \leq s$, or $||\boldsymbol{\theta}||_0 \leq s$. Formulating sparsity as an additional (hard) constraint, rather than optimizing for it in the objective, grants the flexibility to optimize for a different object 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 ▦, 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 preprocesses 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 $\sum_j \boldsymbol{x}_{i,j} = 1 \; \forall \; \boldsymbol{x}_{i,j} \in \{0, 1\}$ are used to ensure that multiple categories are be simultaneously active, and thermometer-encoded ordinal variables are required to satisfy $\boldsymbol{x}_{i,j} \geq \boldsymbol{x}_{i,j+1} \; \forall \; \boldsymbol{x}_{i,j} \in \{0, 1\}$. 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 mailicious 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 imperceptable (e.g., by leaving the data distribution). In a sense, imperceptability 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, $\mathcal{P}(\mathcal{X})$, or feasible actions, $\mathcal{A}(\boldsymbol{x}^{\mathsf{F}})$.[3] 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 δ^*, perfect coverage Ω^*, or efficient runtime τ^*, 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 `dist` and `cost` in (1) and (2), the objective functions and constraints determine the *existence* and *multiplicity* of recourse. For factual instance $\boldsymbol{x}^{\mathsf{F}}$, there may exist zero, one, or multiple[4] 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 formualations 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 (`dist` or `cost`) 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 $\boldsymbol{x}^{\mathsf{F}}$ and potential $\boldsymbol{x}^{\mathsf{CF}}$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.

[3] Optimization terminology refers to both of these constraint sets as *feasibility* sets.
[4] The existence of multiple equally costly recourse actions is commonly referred to as the Rashoman effect [33].

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 inscrutable 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, $\boldsymbol{x}^{\mathsf{CF}}$, is assigned to be the structural counterfactual of $\boldsymbol{x}^{\mathsf{F}}$ under hypothetical actions $\boldsymbol{a}$, i.e., $\boldsymbol{x}^{\mathsf{CF}} = \boldsymbol{x}^{\mathsf{SCF}}(\boldsymbol{a}; \boldsymbol{x}^{\mathsf{F}})$ [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 uncertainy 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., $s \in \{0, 1\}$), where $p(x|s) = \mathcal{N}(0, 10^s)$. Consider a binary classifier, $h : \mathbb{R} \times \mathbb{S} \rightarrow \{0, 1\}$, where $h(x, s) = \text{sign}(x)$. 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 not holding simultaneously) can be resolved by considering a new and distinct 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.

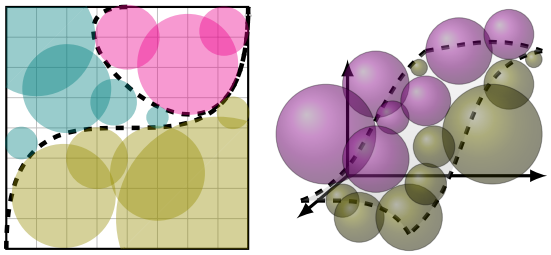

Figure 2: Here we illustrate the model stealing process in 2D and 3D using hypothetical non-linear decision boundaries. "How many optimal contrastive explanations are needed to extract the decision regions of a classifier?" can be formulated as "How many factual balls are needed to maximally pack all decision regions?"

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, $\hat{f}$, that is similar (e.g., in fidelity) to the target model, f:

$$f \approx \hat{f} = \arg\min_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\boldsymbol{x} \sim P(\boldsymbol{x})} \Big[\mathcal{L}\big(f(\boldsymbol{x}), \hat{f}(\boldsymbol{x})\big) \Big]. \tag{6}$$

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[5] [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, $\boldsymbol{x}^{\text{F}}$, returns a nearest contrastive explanation, $\boldsymbol{x}^{*\text{CF}}$, using a known distance function, $\Delta : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$.[6] 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* $\{\mathcal{R}_1, \cdots, \mathcal{R}_l\}$. Every contrastive explanation returned from the recourse API informs us that all instance

[5] A large class of explanation methods rely on the gradients to offer saliency/attribution maps, especially in the image domain.

[6] Explanation models such as MACE [91] provide optimal solutions, $\boldsymbol{x}_\epsilon^{\text{CF}}$, where $f(\boldsymbol{x}^{\text{F}}) \neq f(\boldsymbol{x}_\epsilon^{\text{CF}})$, $\Delta(\boldsymbol{x}^{\text{F}}, \boldsymbol{x}_\epsilon^{\text{CF}}) \leq \Delta(\boldsymbol{x}^{\text{F}}, \boldsymbol{x}^{*\text{CF}}) + \epsilon$, where $\boldsymbol{x}^{*\text{CF}}$ is the optimal nearest contrastive explanation. In practice, $\epsilon = 1e-5$ which in turn results in $\boldsymbol{x}_\epsilon^{\text{CF}} \approx \boldsymbol{x}^{*\text{CF}}$.

surrounding the factual instance, up to a distance of $\Delta(\boldsymbol{x}^{\mathsf{F}}, \boldsymbol{x}^{*\mathsf{CF}})$, share the same class label as $\boldsymbol{x}^{\mathsf{F}}$ according to f (otherwise that instance would be the nearest contrastive explanation). More formally, $f(\boldsymbol{x}^{\mathsf{F}}) = f(\boldsymbol{x}) \; \forall \; \boldsymbol{x} \in \mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}}$, where $\mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}}$ is referred to as the Δ-*factual ball*, centered at $\boldsymbol{x}^{\mathsf{F}}$ and with radius $\Delta(\boldsymbol{x}^{\mathsf{F}}, \boldsymbol{x}^{*\mathsf{CF}})$. The model extraction problem can thus be formulated as the number of factual balls needed to maximally pack all decision regions (see Fig. 2):

$$\Pr[\text{Vol}(\mathcal{R}_l) - \cup^{n}_{i=1, \mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}\, i} \subseteq \mathcal{R}_l} \text{Vol}(\mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}\, i}) \leq \epsilon] \geq 1 - \delta \; \forall \; l \quad (7)$$

As in other extraction settings, $\hat{f}$ can then 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 $\mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}}$) 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 $\mathcal{B}^{\Delta}_{\boldsymbol{x}^{\mathsf{F}}}$ 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-exploration 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.

REFERENCES

[1] Ashraf Abdul, 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. The inverse classification problem. *Journal of Computer Science and Technology* 25, 3 (2010), 458–468.

[5] Carlos Aguilar-Palacios, Sergio Muñoz-Romero, and José Luis Rojo-Álvarez. 2020. Cold-Start Promotional Sales Forecasting through Gradient Boosted-based Contrastive Explanations. *IEEE Access* (2020).

[6] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. *arXiv preprint arXiv:1901.09749* (2019).

[7] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884* (2020).

[8] David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

[9] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91, 434 (1996), 444–455.

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[11] André Artelt and Barbara Hammer. 2019. Efficient computation of counterfactual explanations of LVQ models. *arXiv preprint arXiv:1908.00735* (2019).

[12] André Artelt and Barbara Hammer. 2019. On the computation of counterfactual explanations–A survey. *arXiv preprint arXiv:1911.07749* (2019).

[13] André Artelt and Barbara Hammer. 2020. Convex Density Constraints for Computing Plausible Counterfactual Explanations. *arXiv preprint arXiv:2002.04862* (2020).

[14] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. 2017. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379* (2017).

[15] Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. 2020. Geometrically Enriched Latent Spaces. *arXiv preprint arXiv:2008.00565* (2020).

[16] Emre Ates, Burak Aksar, Vitus J Leung, and Ayse K Coskun. 2020. Counterfactual Explanations for Machine Learning on Multivariate Time Series Data. *arXiv preprint arXiv:2008.10781* (2020).

[17] Kevin Bache and Moshe Lichman. 2013. UCI machine learning repository.

[18] Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. 2019. Imperceptible Adversarial Attacks on Tabular Data. *arXiv preprint arXiv:1911.03274* (2019).

[19] E Bareinboim, JD Correa, D Ibeling, and T Icard. 2020. On Pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)* (2020).

[20] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS Tutorial* 1 (2017).

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

[22] Alejandro Barredo-Arrieta and Javier Del Ser. 2020. Plausible Counterfactuals: Auditing Deep Learning Classifiers with Realistic Adversarial Examples. *arXiv preprint arXiv:2003.11323* (2020).

[23] Clark Barrett, Christopher L Conway, Morgan Deters, Liana Hadarean, Dejan Jovanović, Tim King, Andrew Reynolds, and Cesare Tinelli. 2011. Cvc4. In *International Conference on Computer Aided Verification*. Springer, 171–177.

[24] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773* (2017).

[25] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. 2019. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*. 8974–8984.

[26] Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2, 1 (2009), 183–202.

[27] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1, 1 (2019), 20–23.

[28] Leopoldo Bertossi. 2020. An ASP-Based Approach to Counterfactual Explanations for Classification. *arXiv preprint arXiv:2004.13237* (2020).

[29] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. 2020. Machine Learning Explainability for External Stakeholders. *arXiv preprint arXiv:2007.05408* (2020).

[30] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[31] Or Biran and Courtenay Cotton. [n.d.]. Explanation and justification in machine learning: A survey.

[32] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

[33] Leo Breiman et al. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[34] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[35] Ruth MJ Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.. In *IJCAI*. 6276–6282.

[36] Longbing Cao, Dan Luo, and Chengqi Zhang. 2007. Knowledge actionability: satisfying technical and business interestingness. *International Journal of Business Intelligence and Data Mining* 2, 4 (2007), 496–514.

[37] Longbing Cao and Chengqi Zhang. 2006. Domain-driven actionable knowledge discovery in the real world. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 821–830.

[38] Longbing Cao, Yanchang Zhao, Huaifeng Zhang, Dan Luo, Chengqi Zhang, and Eun Kyo Park. 2009. Flexible frameworks for actionable knowledge discovery. *IEEE Transactions on Knowledge and Data Engineering* 22, 9 (2009), 1299–1312.

[39] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[40] Matt Chapman-Rounds, Marc-Andre Schulz, Erik Pazos, and Konstantinos Georgatzis. 2019. EMAP: Explanation by Minimal Adversarial Perturbation. *arXiv preprint arXiv:1912.00872* (2019).

[41] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. 2020. Boosting decision-based black-box adversarial attacks with random sign flip. In *Proceedings of the European Conference on Computer Vision*.

[42] Furui Cheng, Yao Ming, and Huamin Qu. 2020. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *arXiv preprint arXiv:2008.08353* (2020).

[43] Lee Cohen, Zachary C Lipton, and Yishay Mansour. 2019. Efficient candidate screening under multiple tests and implications for fairness. *arXiv preprint arXiv:1905.11361* (2019).

[44] Gregory F Cooper and Changwon Yoo. 1999. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. 116–125.

[45] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[46] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. 2009. *Introduction to algorithms*. MIT press.

[47] IBM ILOG Cplex. 2009. V12. 1: User's Manual for CPLEX. *International Business Machines Corporation* 46, 53 (2009), 157.

[48] Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. 2015. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 179–188.

[49] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-Objective Counterfactual Explanations. *arXiv preprint arXiv:2004.11165* (2020).

[50] Leonardo De Moura and Nikolaj Bjørner. 2008. Z3: An efficient SMT solver. In *International conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 337–340.

[51] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and user agency: the reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 436–445.

[52] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[53] Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, and Ruchir Puri. 2019. Model agnostic contrastive explanations for structured data. *arXiv preprint arXiv:1906.00117* (2019).

[54] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983* (2019).

[55] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[56] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. [n.d.]. CRUDS: Counterfactual Recourse Using Disentangled Subspaces. ([n. d.]).

[57] Jianfeng Du, Yong Hu, Charles X Ling, Ming Fan, and Mei Liu. 2011. Efficient action extraction with many-to-many relationship between actions and features.

In *International Workshop on Logic, Rationality and Interaction*. Springer, 384–385.

[58] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[59] Cynthia Dwork and Vitaly Feldman. 2018. Privacy-preserving prediction. *arXiv preprint arXiv:1803.10266* (2018).

[60] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

[61] Alex A Freitas. 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter* 15, 1 (2014), 1–10.

[62] Marco Gario and Andrea Micheli. 2015. PySMT: a solver-agnostic library for fast prototyping of SMT-based algorithms. In *SMT workshop*, Vol. 2015.

[63] Andrew Gelman. 2011. Causality and statistical learning.

[64] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. 2017. Eye-tracking causality. *Psychological science* 28, 12 (2017), 1731–1744.

[65] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

[66] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.

[67] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[68] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 531–535.

[69] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[70] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451* (2019).

[71] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245* (2018).

[72] Thomas Grote and Philipp Berens. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics* 46, 3 (2020), 205–211.

[73] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. 2019. Black Box Explanation by Learning Image Exemplars in the Latent Feature Space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 189–205.

[74] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[75] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[76] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2018. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337* (2018).

[77] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. *arXiv preprint arXiv:1909.03166* (2019).

[78] Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science* 56, 4 (2005), 843–887.

[79] Leif Hancox-Li. 2020. Robustness in machine learning explanations: does it matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 640–647.

[80] Masoud Hashemi and Ali Fathi. 2020. PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards. *arXiv preprint arXiv:2008.10138* (2020).

[81] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[82] Denis J Hilton and Ben R Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review* 93, 1 (1986), 75.

[83] Steffen Holter, Oscar Gomez, and Enrico Bertini. [n.d.]. FICO Explainable Machine Learning Challenge. ([n. d.]). https://community.fico.com/s/explainable-machine-learning-challenge

[84] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8779–8788.

[85] Christina Ilvento. 2019. Metric Learning for Individual Fairness. *arXiv preprint arXiv:1906.00250* (2019).

[86] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598* (2018).

[87] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. REVISE: Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *arXiv preprint arXiv:1907.09615* (2019).

[88] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. [n.d.]. DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization. ([n. d.]).

[89] Sin-Han Kang, Hong-Gyu Jung, Dong-Ok Won, and Seong-Whan Lee. 2020. Counterfactual Explanation Based on Gradual Construction for Deep Networks. *arXiv preprint arXiv:2008.01897* (2020).

[90] Masud Karim and Rashedur M Rahman. 2013. Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing. (2013).

[91] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. 895–905.

[92] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic Recourse: from Counterfactual Explanations to Interventions. *arXiv preprint arXiv:2002.06278* (2020).

[93] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv preprint arXiv:2006.06831* (2020).

[94] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. 2017. Reluplex: An efficient SMT solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*. Springer, 97–117.

[95] Mark T Keane and Barry Smyth. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). *arXiv preprint arXiv:2005.13997* (2020).

[96] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*. 277–287.

[97] Maxim S Kovalev and Lev V Utkin. 2020. Counterfactual explanation of machine learning survival models. *arXiv preprint arXiv:2006.16793* (2020).

[98] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[99] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[100] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. https://github.com/propublica/compas-analysis.

[101] Michael T Lash, Qihang Lin, Nick Street, Jennifer G Robinson, and Jeffrey Ohlmann. 2017. Generalized inverse classification. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 162–170.

[102] Michael T Lash, Qihang Lin, and W Nick Street. 2018. Prophit: Causal inverse classification for multiple continuously valued treatment policies. *arXiv preprint arXiv:1802.04918* (2018).

[103] Michael T Lash, Qihang Lin, W Nick Street, and Jennifer G Robinson. 2017. A budget-constrained inverse classification framework for smooth classifiers. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1184–1193.

[104] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2019. Issues with post-hoc counterfactual explanations: a discussion. *arXiv preprint arXiv:1906.04774* (2019).

[105] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse Classification for Comparison-based Interpretability in Machine Learning. *arXiv preprint arXiv:1712.08443* (2017).

[106] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294* (2019).

[107] Eugene L Lawler and David E Wood. 1966. Branch-and-bound methods: A survey. *Operations research* 14, 4 (1966), 699–719.

[108] David K Lewis. 1973. *Counterfactuals*. Harvard University Press.

[109] David K Lewis. 1986. Causal explanation. (1986).

[110] Peter Lipton. 1990. Contrastive explanation. (1990).

[111] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[112] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 641–647.

[113] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 90–98.

[114] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. 2019. Actionable Interpretability through Optimizable Counterfactual Explanations for

Tree Ensembles. *arXiv preprint arXiv:1911.12199* (2019).
[115] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. Explainable reinforcement learning through a causal lens. *arXiv preprint arXiv:1905.10958* (2019).
[116] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Distal Explanations for Explainable Reinforcement Learning Agents. *arXiv preprint arXiv:2001.10284* (2020).
[117] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *arXiv preprint arXiv:1912.03277* (2019).
[118] Michael V Mannino and Murlidhar V Koushik. 2000. The cost-minimizing inverse classification problem: a genetic algorithm approach. *Decision Support Systems* 29, 3 (2000), 283–300.
[119] David Martens and Foster Provost. 2014. Explaining data-driven document classifications. *Mis Quarterly* 38, 1 (2014), 73–100.
[120] Kenneth McGarry. [n.d.]. A survey of interestingness measures for knowledge discovery. ([n. d.]).
[121] Ann L McGill and Jill G Klein. 1993. Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology* 64, 6 (1993), 897.
[122] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*. 7775–7784.
[123] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.
[124] Tim Miller. 2018. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163* (2018).
[125] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[126] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. 2019. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 1–9.
[127] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.
[128] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15.
[129] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1765–1773.
[130] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.
[131] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal Interpretability for Machine Learning-Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.
[132] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2019. DiCE: Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *arXiv preprint arXiv:1905.07697* (2019).
[133] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research* 9, 5 (2002), 583–597.
[134] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.
[135] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition* (2020), 107501.
[136] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 427–436.
[137] Jorge Nocedal and Stephen Wright. 2006. *Numerical optimization*. Springer Science & Business Media.
[138] GUROBI OPTIMIZATION. 2014. INC. Gurobi optimizer reference manual, 2015. *URL: http://www. gurobi. com* (2014), 29.
[139] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 506–519.
[140] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.
[141] Jaehyun Park and Stephen Boyd. 2017. General heuristics for nonconvex quadratically constrained quadratic programming. *arXiv preprint arXiv:1703.07870* (2017).
[142] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. *arXiv preprint arXiv:2006.13132* (2020).
[143] Martin Pawelczyk, Johannes Haug, Klaus Broelemann, and Gjergji Kasneci. 2019. Towards User Empowerment. *arXiv preprint arXiv:1910.09398* (2019).
[144] Judea Pearl. 2000. *Causality: models, reasoning and inference*. Vol. 29. Springer.
[145] Judea Pearl. 2010. The foundations of causal inference. *Sociological Methodology* 40, 1 (2010), 75–149.
[146] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
[147] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[148] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference*. The MIT Press.
[149] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2019. FACE: Feasible and Actionable Counterfactual Explanations. *arXiv preprint arXiv:1909.09369* (2019).
[150] Goutham Ramakrishnan, Yun Chan Lee, and Aws Albargouthi. 2019. Synthesizing Action Sequences for Modifying Model Decisions. *arXiv preprint arXiv:1910.00057* (2019).
[151] Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. 2019. Counterfactual explanation algorithms for behavioral and textual data. *arXiv preprint arXiv:1912.01819* (2019).
[152] Shubham Rathi. 2019. Generating counterfactual and contrastive explanations using SHAP. *arXiv preprint arXiv:1906.09293* (2019).
[153] Kaivalya Rawal and Himabindu Lakkaraju. 2020. Interpretable and Interactive Summaries ofActionable Recourses. *arXiv preprint arXiv:2009.07165* (2020).
[154] Robert Nikolai Reith, Thomas Schneider, and Oleksandr Tkachenko. 2019. Efficiently stealing your machine learning models. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*. 198–210.
[155] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[156] David-Hillel Ruben. 2015. *Explaining explanation*. Routledge.
[157] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[158] Chris Russell. 2019. Efficient Search for Diverse Coherent Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, 20–28. https://doi.org/10.1145/3287560.3287569
[159] Bernhard Schölkopf. 2019. Causality for machine learning. *arXiv preprint arXiv:1911.10500* (2019).
[160] Candice Schumann, Jeffrey S Foster, Nicholas Mattei, and John P Dickerson. 2020. We Need Fairness and Explainability in Algorithmic Hiring. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 1716–1720.
[161] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.
[162] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2019. CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models. *arXiv preprint arXiv:1905.07857* (2019).
[163] Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy risks of explaining machine learning models. *arXiv preprint arXiv:1907.00164* (2019).
[164] Kacper Sokol and Peter A Flach. 2018. Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements.. In *IJCAI*. 5785–5786.
[165] Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In *SafeAI@ AAAI*.
[166] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. 2012. *Optimization for machine learning*. Mit Press.
[167] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[168] Jin Tian and Judea Pearl. 2001. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. 512–521.
[169] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 465–474.
[170] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*. 601–618.

[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 Voigt 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] Binghui 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:1708.01870* (2017).

[181] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, 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).