Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals

Sainyam Galhotra*
University of Massachusetts Amherst
sainyam@cs.umass.edu

Romila Pradhan*
University of California San Diego
rpradhan@ucsd.edu

Babak Salimi
University of California San Diego
bsalimi@ucsd.edu

ABSTRACT

There has been a recent resurgence of interest in *explainable artificial intelligence* (XAI) that aims to reduce the opaqueness of AI-based decision-making systems, allowing humans to scrutinize and trust them. Prior work in this context has focused on the attribution of *responsibility* for an algorithm's decisions to its inputs wherein responsibility is typically approached as a purely *associational* concept. In this paper, we propose a principled causality-based approach for explaining black-box decision-making systems that addresses limitations of existing methods in XAI. At the core of our framework lies *probabilistic contrastive counterfactuals*, a concept that can be traced back to philosophical, cognitive, and social foundations of theories on how humans generate and select explanations. We show how such counterfactuals can quantify the *direct* and *indirect* influences of a variable on decisions made by an algorithm, and provide *actionable recourse* for individuals negatively affected by the algorithm's decision. Unlike prior work, our system, Lewis: (1) can compute provably effective explanations and recourse at local, global and contextual levels; (2) is designed to work with users with varying levels of background knowledge of the underlying causal model; and (3) makes no assumptions about the internals of an algorithmic system except for the availability of its input-output data. We empirically evaluate Lewis on four real-world datasets and show that it generates human-understandable explanations that improve upon state-of-the-art approaches in XAI, including the popular LIME and SHAP. Experiments on synthetic data further demonstrate the correctness of Lewis's explanations and the scalability of its recourse algorithm.

ACM Reference Format:
Sainyam Galhotra, Romila Pradhan, and Babak Salimi. 2021. Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3448016.3458455

*Both authors contributed equally to this research.

SIGMOD '21, June 20–25, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8343-1/21/06…$15.00
https://doi.org/10.1145/3448016.3458455

1 INTRODUCTION

Algorithmic decision-making systems are increasingly used to automate consequential decisions, such as lending, assessing job applications, informing release on parole, and prescribing life-altering medications. There is growing concern that the opacity of these systems can inflict harm to stakeholders distributed across different segments of society. These calls for transparency created a resurgence of interest in *explainable artificial intelligence* (XAI), which aims to provide human-understandable explanations of outcomes or processes of algorithmic decision-making systems (see [36, 64, 65] for recent surveys).

Effective explanations should serve the following purposes: (1) help to build trust by providing a mechanism for *normative evaluation* of an algorithmic system, ensuring different stakeholders that the system's decision rules are justifiable [86]; and (2) provide users with an *actionable recourse* to change the results of algorithms in the future [9, 45, 99, 101]. Existing methods in XAI can be broadly categorized based on whether explainability is achieved by design (*intrinsic*) or by post factum system analysis (*post hoc*), and whether the methods assume access to system internals (*model dependent*) or can be applied to any black-box algorithmic system (*model agnostic*).

In this work, we address post hoc and model-agnostic explanation methods that are applicable to any proprietary black-box algorithm. Prior work in this context has focused on the attribution of *responsibility* of an algorithm's decisions to its inputs. These approaches include methods for quantifying the *global* (population-level) or *local* (individual-level) *influence* of an algorithm's input on its output [5, 17, 24, 25, 31, 34, 38, 57, 58]; they also include methods based on *surrogate explainability*, which search for a simple and interpretable model (such as a decision tree or a linear model) that mimics the behaviour of a black-box algorithm [75, 76]. However, these methods can produce incorrect and misleading explanations primarily because they focus on the *correlation* between the input and output of algorithms as opposed to their *causal* relationship [4, 26, 36, 39, 48, 65]. Furthermore, several recent works have argued for the use of *counterfactual explanations*, which are typically obtained by considering the smallest perturbation in an algorithm's input that can lead to the algorithm's desired outcome [51, 60, 68, 96, 101]. However, due to the causal dependency between variables, these perturbations are not translatable into real-world interventions and therefore fail to generate insights that are actionable in the real world [8, 42, 44, 46, 60, 89].

This paper describes a **new causality-based framework** for generating post-hoc explanations for black-box decision-making algorithms **that unifies existing methods in XAI and addresses**

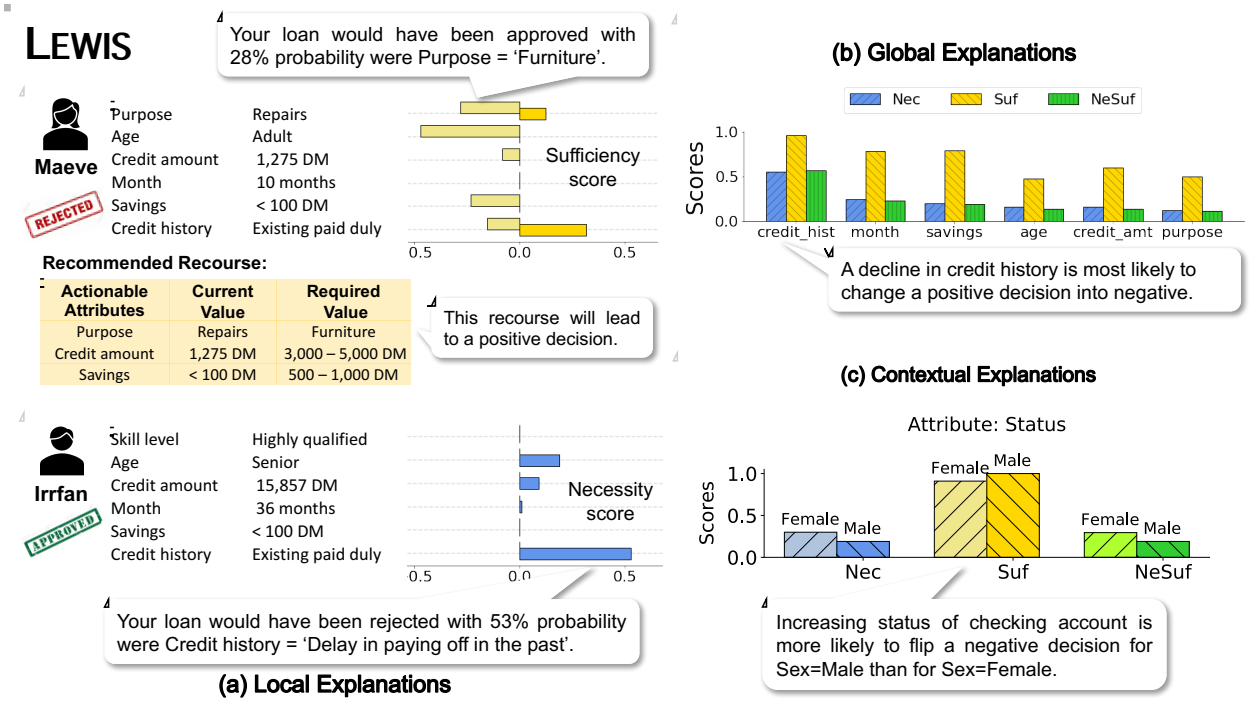


Figure 1: An overview of explanations generated by Lewis for a loan approval algorithm built using the UCI German credit dataset (see Section 5 for details). Given a black-box classification algorithm, Lewis generates: (a) Local explanations, that explain the algorithm's output for an individual; (b) Global explanations, that explain the algorithm's behavior across different attributes; and (c) Contextual explanations, that explain the algorithms's predictions for a sub-population of individiuals.

their limitations. Our system, Lewis, [1] reconciles the aforementioned objectives of XAI by: (1) providing insights into what *causes* an algorithm's decisions at the global, local and contextual (sub-population) levels, and (2) generating actionable recourse translatable into real-world interventions. At the heart of our proposal are *probabilistic contrastive counterfactuals* of the following form:

> "For individual(s) with attribute(s) <actual-value> for whom an algorithm made the decision <actual-outcome>, the decision would have been <foil-outcome> with *probability* <score> had the attribute been <counterfactual-value>." (1)

Contrastive counterfactuals are at the core of the philosophical, cognitive, and social foundations of theories that address how humans generate and select explanations [18, 29, 35, 55, 66, 71, 102]. Their probabilistic interpretation has been formalized and studied extensively in AI, biostatistics, political science, epistemology, biology and legal reasoning [14, 32, 32, 33, 35, 61, 71, 77, 78, 91]. While their importance in achieving the objectives of XAI has been recognized in the literature [63], very few attempts have been made to operationalize causality-based contrastive counterfactuals for XAI. The following example illustrates how Lewis employs contrastive counterfactuals to generate different types of explanations.

Example 1.1. Consider the black-box loan-approval algorithm in Figure 1 for which Lewis generates different kinds of explanations. For local explanations, Lewis ranks attributes in terms of their *causal* responsibility to the algorithm's decision. For individuals whose loans were rejected, the responsibility of an attribute is measured by its *sufficiency* score, defined as "the probability that the algorithm's decision would have been positive if that attribute had a counterfactual value". For Maeve, the sufficiency score of 28% for purpose of loan means that if purpose were 'Furniture', Maeve's loan would have been approved with a 28% probability. For individuals whose loans were approved, the responsibility of an attribute is measured by its *necessity* score, defined as "the probability that the algorithm's decision would have been negative if that attribute had a counterfactual value." For Irrfan, the necessity score of 53% for credit history means that had credit history been worse, Irrfan would have been denied the loan with a 53% probability. Furthermore, individuals with a negative decision, such as Maeve, would want to know the actions they could take that would likely change the algorithm's decision. For such users, Lewis suggests the minimal causal interventions on the set of actionable attributes that are sufficient, with high probability, to change the algorithm's decision in the future. Additionally, Lewis generates insights about the algorithm's *global* behavior with respect to each attribute by

[1] Our system is named after David Lewis (1941–2001), who made significant contributions to modern theories of causality and explanations in terms of counterfactuals. In his essay on causal explanation [52], Lewis argued that "to explain an event is to provide some information about its causal history." He further highlighted the role of counterfactual contrasts in explanations when he wrote, "One way to indicate what sort of explanatory information is wanted is through the use of contrastive why-questions . . . [where] information is requested about the difference between the actualized causal history of the explanandum and the unactualized causal histories of its unactualized alternatives [(termed as "foils" by Peter Lipton [55])]. Why did I visit Melbourne in 1979, rather than Oxford or Uppsala or Wellington?"

computing its necessity, sufficiency, and necessity and sufficiency scores at the population level. For instance, a higher necessity score for credit history indicates that a decline in its value is more likely to reverse a positive decision than a lower value of savings; a lower sufficiency score for age indicates that increasing it is less likely to overturn a negative decision compared to credit history or savings. By further customizing the scores for a *context* or sub-population of individuals that share some attributes, Lewis illuminates the *contextual* behavior of the algorithm in different sub-populations. In Figure 1, Lewis indicates that improving the `status` of checking account is more likely to reverse a negative decision for {`sex=Male`} than for {`sex=Female`}.

To compute these scores, Lewis relies on the ordinal importance of attribute values e.g., higher savings are more likely to be granted a loan than lower savings. In case the attribute values do not possess a natural ordering or the ordering is not known apriori, Lewis infers it from the output of the black-box algorithm (more in Section 4.1).

Our contributions. This paper proposes a principled approach for explaining black-box decision-making systems using probabilistic contrastive counterfactuals. Key contributions include:

(1) Adopting standard definitions of sufficient and necessary causation based on contrastive counterfactuals to propose **novel probabilistic measures, called *necessity scores* and *sufficiency scores***, which respectively quantify the extent to which an attribute is necessary and sufficient for an algorithm's decision (Section 3.1). We show that these measures play unique, complementary roles in generating effective explanations for algorithmic systems. While necessity score addresses the *attribution* of causal responsibility of an algorithm's decisions to an attribute, sufficiency score addresses the tendency of an attribute to *produce* the desired algorithmic outcome.

(2) Demonstrating that our newly proposed measures can generate a **wide range of explanations** for algorithmic systems that quantify the necessity and sufficiency of attributes that **implicitly** or **explicitly** influence an algorithm's decision-making process (Section 3.2). More importantly, Lewis generates **contextual explanations** at global or local levels and for a user-defined sub-population.

(3) Showing that the problem of generating **actionable recourse** can be framed as an optimization problem that searches for a **minimal intervention** on a pre-specified set of actionable variables that have a high **sufficiency score** for producing the algorithm's desired future outcome.

(4) Establishing conditions under which **the class of probabilistic contrastive counterfactuals we use can be bounded and estimated using historical data** (Section 4.1). Unlike previous attempts to generate actionable recourse using counterfactual reasoning, Lewis leverages established bounds and integer programming to generate reliable recourse under partial background knowledge on the underlying causal models (Section 4.2).

(5) Comparing Lewis to state-of-the-art methods in XAI (Sections 5 and 6). We present an **end-to-end experimental evaluation on both real and synthetic data**. On real datasets, we show that Lewis generates intuitive and actionable explanations that are consistent with insights from existing literature and surpass state-of-the-art methods in XAI. Evaluation on synthetic data demonstrates the accuracy and correctness of the explanation scores and actionable recourse that Lewis generates.

Symbol	Meaning
X, Y, Z	attributes (variables)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	sets of attributes
$Dom(X), Dom(\mathbf{X})$	their domains
$x \in Dom(X)$	an attribute value
$\mathbf{x} \in Dom(\mathbf{X})$	a tuple of attribute values
$\mathbf{k} \in Dom(\mathbf{K})$	a tuple of context attribute values
G	causal diagram
$\langle M, \Pr(\mathbf{u})\rangle$	probabilistic causal model
$O_{\mathbf{X}\leftarrow\mathbf{x}}$	potential outcome
$\Pr(\mathbf{V}=\mathbf{v}), \Pr(\mathbf{v})$	joint probability distribution
$\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}})$	abbreviates $\Pr(O_{\mathbf{X}\leftarrow\mathbf{x}}=o)$

Table 1: Notation used in this paper.

2 PRELIMINARIES

The notation we use in this paper is summarized in Table 1. We denote variables by uppercase letters, X, Y, Z, V; their values with lowercase letters, x, y, z, v; and sets of variables or values using boldface ($\mathbf{X}$ or $\mathbf{x}$). The domain of a variable X is $Dom(X)$, and the domain of a set of variables is $Dom(\mathbf{X}) = \prod_{X\in\mathbf{X}} Dom(X)$. All domains are discrete andfi nite; continuous domains are assumed to be binned. We use $\Pr(\mathbf{x})$ to represent a joint probability distribution $\Pr(\mathbf{X}=\mathbf{x})$. The basic semantic framework of our proposal rests on probabilistic causal models [71], which we review next.

Probabilistic causal models. A *probabilistic causal model* (PCM) is a tuple $\langle M, \Pr(\mathbf{u})\rangle$, where $M = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}\rangle$ is a *causal model* consisting of a set of *observable or endogenous* variables $\mathbf{V}$ and a set of *background or exogenous* variables $\mathbf{U}$ that are outside of the model, and $\mathbf{F} = (F_X)_{X\in\mathbf{V}}$ is a set of *structural equations* of the form $F_X : Dom(\mathbf{Pa_V}(X)) \times Dom(\mathbf{Pa_U}(X)) \rightarrow Dom(X)$, where $\mathbf{Pa_U}(X) \subseteq \mathbf{U}$ and $\mathbf{Pa_V}(X) \subseteq \mathbf{V} - \{X\}$ are called *exogenous parents* and *endogenous parents* of X, respectively. The values of $\mathbf{U}$ are drawn from the distribution $\Pr(\mathbf{u})$. A PCM $\langle M, \Pr(\mathbf{u})\rangle$ can be represented as a directed graph $G = \langle \mathbf{V}, \mathbf{E}\rangle$, called a *causal diagram*, where each node represents a variable, and there are directed edges from the elements of $\mathbf{Pa_U}(X) \cup \mathbf{Pa_V}(X)$ to X. We say a variable Z is a *descendant* of another variable X if Z is *caused* (either *directly* or *indirectly*) by X, i.e., if there is a directed edge or path from X to Z in G; otherwise, we say that Z is a *non-descendant* of X.

Interventions and potential outcomes. An *intervention* or an *action* on a set of variables $\mathbf{X} \subseteq \mathbf{V}$, denoted $\mathbf{X} \leftarrow \mathbf{x}$, is an operation that *modifies* the underlying causal model by replacing the structural equations associated with $\mathbf{X}$ with a constant $\mathbf{x} \in Dom(\mathbf{X})$. The *potential outcome* of a variable Y after the intervention $\mathbf{X} \leftarrow \mathbf{x}$ in a context $\mathbf{u} \in Dom(\mathbf{U})$, denoted $Y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u})$, is the *solution* to Y in the modified set of structural equations. Potential outcomes satisfy the following *consistency rule* used in derivations presented in Section 4.1.

$$\mathbf{X}(\mathbf{u}) = \mathbf{x} \implies Y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) = y \tag{2}$$

This rule states that in contexts where $\mathbf{X} = \mathbf{x}$, the outcome is invariant to the intervention $\mathbf{X} \leftarrow \mathbf{x}$. For example, changing the income level of applicants to high does not change the loan decisions for those who already had high income before the intervention.

The distribution $\Pr(\mathbf{u})$ induces a probability distribution over endogenous variables and potential outcomes. Using PCMs, one can express *counterfactual queries* of the form $\Pr(Y_{\mathbf{X}\leftarrow\mathbf{x}} = y \mid \mathbf{k})$, or simply $\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k})$; this reads as "For contexts with attributes $\mathbf{k}$, what is the probability that we would observe $Y = y$ had $\mathbf{X}$ been

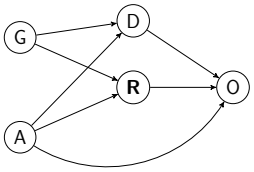


Figure 2: A causal diagram for Example 1.1.

$\mathbf{x}$?" and is given by the following expression:

$$\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k}) = \sum_{\mathbf{u}} \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}))\ \Pr(\mathbf{u} \mid \mathbf{k}) \tag{3}$$

Equation (3) readily suggests Pearl's three-step procedure for answering counterfactual queries [71][Chapter 7]: (1) update $\Pr(\mathbf{u})$ to obtain $\Pr(\mathbf{u} \mid \mathbf{k})$ (*abduction*), (2) modify the causal model to reflect the intervention $\mathbf{X} \leftarrow \mathbf{x}$ (*action*), and (3) evaluate the RHS of (3) using the index function $\Pr(Y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}) = y)$ (*prediction).* However, performing this procedure requires the underlying PCM to be fully observed, i.e, the distribution $\Pr(\mathbf{u})$ and the underlying structural equations must be known, which is an impractical requirement. In this paper, we assume that only background knowledge of the underlying causal diagram is available, but exogenous variables and structural equations are unknown.

The do-operator. For causal diagrams, Pearl defined the do-operator as a graphical operation that gives semantics to *interventional queries* of the form "What is the probability that we would observe $Y = y$ (at population-level) had $\mathbf{X}$ been $\mathbf{x}$?", denoted $\Pr(\mathbf{y} \mid \mathsf{do}(\mathbf{x}))$. Further, he proved a set of necessary and sufficient conditions under which interventional queries can be answered using historical data. A sufficient condition is the backdoor-criterion, which states that if there exists a set of variables $\mathbf{C}$ that satisfy a graphical condition relative to $\mathbf{X}$ and Y in the causal diagram G, the following holds (see [71][Chapter 3] for details):

$$\Pr(\mathbf{y} \mid \mathsf{do}(\mathbf{x})) = \sum_{\mathbf{c}\in Dom(\mathbf{C})} \Pr(\mathbf{y} \mid \mathbf{c}, \mathbf{x})\ \Pr(\mathbf{c}) \tag{4}$$

In contrast to (3), notice that the RHS of (4) is expressed in terms of observed probabilities and can be estimated from historical data using existing statistical and ML algorithms.

Counterfactuals vs. interventional queries. The do-operator is a population-level operator, meaning it can only express queries about the effect of an intervention at population level; in contrast, counterfactuals can express queries about the effect of an intervention on a sub-population or an individual. Therefore, every interventional query can be expressed in terms of counterfactuals, but not vice versa (see [74][Chapter 4],[6] for more details). For instance, $\Pr(y \mid \mathsf{do}(\mathbf{x})) = \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}})$; however, the counterfactual query $\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{x}', y')$, which asks about the effect of the intervention $\mathbf{X} \leftarrow \mathbf{x}$ on a sub-population with attributes $\mathbf{x}'$ and y', cannot be expressed in terms of the do-operator (see Example 2.1 below). Note that the probabilistic contrastive counterfactual statements in (1), used throughout this paper to explain a black-box decision-making system concerned with the effect of interventions at sub-population and individual levels, cannot be expressed using the do-operator and therefore cannot be assessed in general when the underlying probabilistic causal models are not fully observed. Nevertheless, in Section 4.1 we establish conditions under which these counterfactuals can be estimated or bounded using data.

Example 2.1. Continuing Example 1.1, Figure 2 represents a simple causal diagram for the loan application domain, where G corresponds to the attribute gender, A to age, D to the repayment duration in months, O to the decision of a loan application, and R compactly represents the rest of the attributes, e.g., status of checking account, employment, savings, etc. Note that the loan decision is binary: $O = 1$ and $O = 0$ indicate whether the loan has been approved or not, respectively. The interventional query $\Pr(O = 1 \mid \mathsf{do}(D = 24 \text{ months}))$ that is equivalent to the counterfactual $\Pr(O_{D\leftarrow 24 \text{ months}} = 1)$ reads as "What is the probability of loan approval at population-level had all applicants selected repayment duration of 24 months?" This query can be answered using data and the causal diagram (since $\{G, A\}$ satisfies the backdoor-criterion in the causal diagram in Figure 2). However, the counterfactual query $\Pr(O_{D\leftarrow 24 \text{ months}} = 1 \mid O = 0)$, which reads as 'What is the probability of loan approval for a group of applicants whose loan applications were denied had they selected a repayment duration of 24 months?", cannot be expressed using the do-operator.

3 EXPLANATIONS AND RECOURSE USING PROBABILISTIC COUNTERFACTUALS

In this section, we introduce three measures to quantify the influence of an attribute on decisions made by an algorithm (Section 3.1). We then use these measures to generate different types of explanations for algorithmic systems (Section 3.2).

3.1 Explanation Scores

We are given a decision-making algorithm $f : Dom(\mathbf{I}) \rightarrow Dom(O)$, where $\mathbf{I}$ is set of input attributes (a.k.a. features for ML algorithms) and O is a binary attribute, where $O = o$ denotes the positive decision (loan approved) and $O = o'$ denotes the negative decision (loan denied). Let us assume we are given a PCM $\langle M, \Pr(\mathbf{u})\rangle$ with a corresponding causal diagram G (this assumption will be relaxed in Section 4.1) such that $\mathbf{I} \subseteq \mathbf{V}$, i.e., the inputs of f are a subset of the observed attributes. Consider an attribute $X \in \mathbf{V}$ and a pair of attribute values $x, x' \in Dom(X)$. We quantify the influence of the attribute value x relative to a baseline x' on decisions made by an algorithm using the following scores, herein referred to as *explanation scores*; (we implicitly assume an order $x > x'$).

Definition 3.1 (Explanation Scores). Given a PCM $\langle M, \Pr(\mathbf{u})\rangle$ and an algorithm $f : Dom(\mathbf{X}) \rightarrow Dom(O)$, a variable $X \in \mathbf{V}$, and a pair of attribute values $x, x' \in Dom(X)$, we quantify the influence of x relative to x' on the algorithm's decisions in the context $\mathbf{k} \in Dom(\mathbf{K})$, where $\mathbf{K} \subseteq \mathbf{V} - \{X, O\}$, using the following measures:

- The *necessity score*:

$$\textsc{Nec}_x^{x'}(\mathbf{k}) \stackrel{\text{def}}{=} \Pr(o'_{X\leftarrow x'} \mid x, o, \mathbf{k}) \tag{5}$$

- The *sufficiency score*:

$$\textsc{Suf}_x^{x'}(\mathbf{k}) \stackrel{\text{def}}{=} \Pr(o_{X\leftarrow x} \mid x', o', \mathbf{k}) \tag{6}$$

- The *necessity and sufficiency score*:

$$\textsc{NeSuf}_x^{x'}(\mathbf{k}) \stackrel{\text{def}}{=} \Pr(o_{X\leftarrow x}, o'_{X\leftarrow x'} \mid \mathbf{k}), \tag{7}$$

where the distribution $\Pr(o_{X\leftarrow x})$ is well-defined and can be computed from the algorithm $f(\mathbf{I})$.[2]

For simplicity of notation, we drop x' from $\text{Nec}_x^{x'}$, $\text{Suf}_x^{x'}$ and $\text{NeSuf}_x^{x'}$ whenever it is clear from the context. The necessity score in (5) formalizes the probabilistic contrastive counterfactual in (1), where <actual-value> and <counterfactual-value> are respectively $\mathbf{k} \cup x$ and $\mathbf{k} \cup x'$, and <actual-decision> and <foil-decision> are respectively positive decision o and negative decision o'. This reads as "What is the probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *negative* instead of *positive* had X been x' instead of x?" In other words, $\text{Nec}_X(.)$ measures the algorithm's percentage of positive decisions that are *attributable to* or *due to* the attribute value x. The sufficiency score in (6) is the dual of the necessity score; it reads as "What would be the probability that for individuals with attributes $\mathbf{k}$, the algorithm's decision would be *positive* instead of *negative* had X been x instead of x'?" Finally, the necessity and sufficiency score in (7) establishes a balance between necessary and sufficiency; it measures the probability that the algorithm responds in both ways. Hence, it can be used to measure the general explanatory power of an attribute. In Section 4.1, we show that the necessary and sufficiency score is non-zero iff X causally influences the algorithm's decisions. (Note that the explanation scores are well-defined for a set of attributes.)

Remark. A major difference between our proposal and existing methods in XAI is the ability to account for the indirect influence of attributes that may not be explicitly used in an algorithm's decision making process, but implicitly influence its decisions via their proxies. The ability to account for such influences is particularly important in auditing algorithms for fairness, where typically sensitive attributes, such as race or gender, are not explicitly used as input to algorithms. For instance, in [97] Wall Street Journal investigators reported that a seemingly innocent online pricing algorithm that simply adjusts online prices based on users' proximity to competitors' stores is discriminative against lower-income individuals. In this case, the algorithm does not explicitly use income; however, it turns out that living further from competitors' stores is a proxy for low income. [3]

3.2 Lewis's Explanations

Based on the explanations scores proposed in Section 3.1, Lewis generates the following types of explanations.

Global, local and contextual explanations. To understand the influence of each variable $X \in \mathbf{V}$ on an algorithm's decision, Lewis computes the necessity score $\text{Nec}_X(\mathbf{k})$, sufficiency score $\text{Suf}_X(\mathbf{k})$, and necessity and sufficiency score $\text{NeSuf}_X(\mathbf{k})$ for each value $x \in Dom(X)$ in the following contexts: (1) $\mathbf{K} = \emptyset$: the scores measure the *global* influence of X on the algorithm's decision. (2) $\mathbf{K} = \mathbf{V}$: the scores measure the individual-level or *local* influence of X on the algorithm's decision. (3) A user-defined $\mathbf{K} = \mathbf{k}$ with $\emptyset \subsetneq \mathbf{K} \subsetneq \mathbf{V}$: the scores measure the *contextual* influence of X on the algorithm's decision. In the context $\mathbf{k}$, Lewis calculates the explanation scores for an attribute X by computing the maximum score over all pairs of attribute values $x, x' \in Dom(X)$. In addition to singleton variables, Lewis can calculate explanation scores for any user-defined set of attributes.

For a given individual, Lewis estimates the positive and negative contributions of a specific attribute value toward the outcome. Consider an individual with a negative outcome $O = o'$ having the attribute $X = x'$. The negative contribution of x' is characterized by the probability of getting a positive outcome on intervening $X \leftarrow x$, $\max_{x>x'} \text{Suf}_x^{x'}(\mathbf{k})$, and the positive contribution of x' for the individual is calculated as $\max_{x''<x'} \text{Suf}_{x'}^{x''}(\mathbf{k})$. Similarly, for an individual with a positive outcome $O = o$ attribute value $X = x'$, the positive contribution of x' is calculated by estimating the probability of o' if the attribute value was intervened to be smaller than x', $\max_{x''<x'} \text{Nec}_{x'}^{x''}(\mathbf{k})$ and the negative contribution of $X = x'$ is $\max_{x>x'} \text{Nec}_x^{x'}(\mathbf{k})$. Note that the negative contribution of attribute $X = x'$ is calculated by intervening on the individual at hand, but the positive contribution is estimated by intervening on individuals with $X = x''$ to satisfy the same context $\mathbf{k}$. In Figure 1, low credit amount contributes negatively to the outcome for Maeve as increasing credit amount improves their chances of getting the loan approved. Attributes like credit history contribute both positively and negatively: poor credit history worsens the chances of approval, but improving credit history furthers the chances of better credit.

Counterfactual recourse. For individuals for whom an algorithm's decision is negative, Lewis generates explanations in terms of minimal interventions on a user-specified set of actionable variables $\mathbf{A} \subseteq \mathbf{V}$ that have a high *sufficiency score*, i.e., the intervention can produce the positive decision with high probability. The explanations can be used either as justification in case the decision is challenged or as a feasible action that the individual may perform in order to improve the outcome in the future ("recourse"). For example, in Figure 1, the set of actionable items for Maeve may consist of her credit amount, loan duration, savings and purpose. Examples of specific actions include "increase the loan repayment duration" or "raise the amount in savings."

Given an individual with attributes $\mathbf{v}$, a set of actionable variables $\mathbf{A} \subseteq \mathbf{V}$, and a cost function $\text{Cost}(\mathbf{a}, \hat{\mathbf{a}})$ that determines the cost of an intervention that changes $\mathbf{A}$ from its current value $\mathbf{a}$ to $\hat{\mathbf{a}}$, for $\hat{\mathbf{a}} \in Dom(\mathbf{A})$, a *counterfactual recourse* can be computed using the following optimization problem:

$$\underset{\mathbf{a} \in Dom(\mathbf{A})}{\text{argmin}} \ \text{Cost}(\mathbf{a}, \hat{\mathbf{a}}) \qquad \text{s.t. } \text{Suf}_{\hat{\mathbf{a}}}(\mathbf{v}) \geq \alpha \tag{8}$$

The optimization problem in (8) treats the decision-making algorithm as a black box; hence, it can be solved merely using *historical data* (see Section 4.2). The solutions to this problem provide end-users with informative, feasible and actionable explanations and recourse by answering questions such as "What are the best courses of action that, if performed in the real world, would with high probability change the outcome for this individual?"

4 PROPERTIES AND ALGORITHMS

In this section, we study properties of the explanation scores in Section 3 and establish conditions under which they can be bounded or estimated from historical data (Section 4.1). We then develop

[2] For deterministic $f(\mathbf{I})$, $\Pr(o_{X\leftarrow x}) = \sum_{\mathbf{i} \in Dom(\mathbf{I})} \mathbb{1}_{\{f(\mathbf{i})=o\}} \Pr(\mathbf{I}_{X\leftarrow x} = \mathbf{i})$, where $\mathbb{1}_{\{f(\mathbf{i})=o\}}$ is an indicator function.

[3] In contrast to mediational analysis in causal inference that studies direct and indirect causal effects [70, 73], in this paper we are interested in quantifying the sufficiency and necessity scores of attributes explicitly and implicitly used by the algorithm.

an algorithm for solving the optimization problem for computing counterfactual recourse (Section 4.2).

4.1 Computing Explanation Scores

Recall from Section 2 that if the underlying PCM is fully specified, i.e., the structural equations and the exogenous variables are observed, then counterfactual queries, and hence the explanation scores, can be computed via Equation (3). However, in many applications, PCMs are not fully observed, and one must estimate explanation scores from data. First, we prove the following bounds on explanation scores, computed for a set of attributes $\mathbf{X}$.

Proposition4.1. Given a PCM $\langle M, \Pr(\mathbf{u})\rangle$ with a corresponding causal DAG G, an algorithm $f : Dom(\mathbf{I}) \rightarrow Dom(O)$, and a set of attributes $\mathbf{X} \subseteq \mathbf{V} - \{O\}$ with two sets of attribute values $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$, if $\mathbf{K}$ consists of non-descendants of $\mathbf{I}$ in G, then the explanation score can be bounded as follows:

$$\max\left(0, \frac{\Pr(o,\mathbf{x}\mid\mathbf{k})+\Pr(o,\mathbf{x}'\mid\mathbf{k}) - \Pr(o\mid \mathrm{do}(\mathbf{x}'),\mathbf{k})}{\Pr(o,\mathbf{x}\mid\mathbf{k})}\right) \leq \mathrm{NEC}_{\mathbf{x}}(\mathbf{k}) \leq \min\left(\frac{\Pr(o'\mid \mathrm{do}(\mathbf{x}'),\mathbf{k}) - \Pr(o',\mathbf{x}'\mid\mathbf{k})}{\Pr(o,\mathbf{x}\mid\mathbf{k})}, 1\right) \quad (9)$$

$$\max\left(0, \frac{\Pr(o',\mathbf{x}\mid\mathbf{k})+\Pr(o',\mathbf{x}'\mid\mathbf{k}) - \Pr(o'\mid \mathrm{do}(\mathbf{x}),\mathbf{k})}{\Pr(o',\mathbf{x}'\mid\mathbf{k})}\right) \leq \mathrm{SUF}_{\mathbf{x}}(\mathbf{k}) \leq \min\left(\frac{\Pr(o\mid \mathrm{do}(\mathbf{x}),\mathbf{k}) - \Pr(o,x\mid\mathbf{k})}{\Pr(o',\mathbf{x}'\mid\mathbf{k})}, 1\right) \quad (10)$$

$$\max\left(0, \Pr(o\mid \mathrm{do}(\mathbf{x}),\mathbf{k}) - \Pr(o\mid \mathrm{do}(\mathbf{x}'),\mathbf{k})\right) \leq \mathrm{NESUF}_{\mathbf{x}}(\mathbf{k}) \leq \min\left(\Pr(o\mid \mathrm{do}(\mathbf{x}),\mathbf{k}), \Pr(o'\mid \mathrm{do}(\mathbf{x}'),\mathbf{k})\right) \quad (11)$$

Proof. We prove the bounds for (9); (10) and (11) are proved similarly. The following equations are obtained from the law of total probability:

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) \quad (12)$$

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},o'_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},o_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) \quad (13)$$

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x}',\mathbf{k}) + \sum_{\mathbf{x}''\in Dom(\mathbf{X})-\{\mathbf{x},\mathbf{x}'\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x}'',\mathbf{k}) \quad (14)$$

By rearranging (12) and (13), we obtain the following equality:

$$\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},o_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) = \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) + \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) \quad (15)$$

The following bounds for the LHS of (15) are obtained from the Fréchet bound.[4]

$$LHS \geq \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) - \Pr(o',\mathbf{x}',\mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) - \sum_{\mathbf{x}''\in Dom(\mathbf{X})-\{\mathbf{x},\mathbf{x}'\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x}'',\mathbf{k})$$

(obtained from Eq. (14) and (2), lower bounding $\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k})$)

$$\geq \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) - \Pr(o',x,\mathbf{k}) - \Pr(o',x',\mathbf{k}) - \Pr(\mathbf{k}) + \Pr(x,\mathbf{k}) + \Pr(x',\mathbf{k})$$

$$= \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) + \Pr(o,x,\mathbf{k}) + \Pr(o,x',\mathbf{k}) - \Pr(\mathbf{k})$$

$$= \Pr(o,x,\mathbf{k}) + \Pr(o,x',\mathbf{k}) - \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) \quad (16)$$

$$LHS \leq \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) - \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},\mathbf{x},\mathbf{k}) + \Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k}) = \Pr(o,\mathbf{x},\mathbf{k}) \quad (17)$$

(obtained by upper bounding $\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k})$ in Eq. (15))

$$LHS \leq \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) - \Pr(o',\mathbf{x}',\mathbf{k}) - \sum_{\mathbf{x}''\in Dom(\mathbf{X})-\{\mathbf{x},\mathbf{x}'\}} \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x}'',\mathbf{k})$$

(obtained from Eq. (14) and (2), upper bounding $\Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}},o_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{x},\mathbf{k})$)

$$\leq \Pr(o'_{\mathbf{X}\leftarrow\mathbf{x}'},\mathbf{k}) - \Pr(o',\mathbf{x}',\mathbf{k}) \quad (18)$$

Equation (9) is obtained by dividing (16), (17) and (18) by $\Pr(o,\mathbf{x},\mathbf{k})$, applying the consistency rule (2), and considering the fact that since $\mathbf{K}$ consists of non-descendants of $\mathbf{X}$, the intervention $\mathbf{X} \leftarrow \mathbf{x}'$ does not change $\mathbf{K}$; hence, $\Pr(o_{\mathbf{X}\leftarrow\mathbf{x}'} \mid \mathbf{k}) = \Pr(o \mid \mathrm{do}(\mathbf{x}'),\mathbf{k})$.

□

[4] $\max\left(0, \sum_{x\in\mathbf{X}} \Pr(x) - (|\mathbf{x}|-1)\right) \leq \Pr(\mathbf{x}) \leq \min_{x\in\mathbf{x}} \Pr(x)$

Proposition 4.1 shows the explanation scores can be bounded whenever interventional queries of the form $\Pr(o \mid \mathrm{do}(\mathbf{x}),\mathbf{k})$ can be estimated from historical data using the underlying causal diagram G (cf. Section 2). The next proposition further shows that if the algorithm is *monotone* relative to $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$, i.e., if $\mathbf{x} > \mathbf{x}'$, then $O_{\mathbf{X}\leftarrow\mathbf{x}} \geq O_{\mathbf{X}\leftarrow\mathbf{x}'}$[5], and the exact value of the explanation scores can be computed from data. (In case the ordering between $\mathbf{x}$ and $\mathbf{x}'$ is not known apriori (e.g., for categorical values), we infer it by comparing the output of the algorithm for $\mathbf{x}$ and $\mathbf{x}'$.)

Proposition4.2. Given a causal diagram G, if the decision-making algorithm $f : Dom(\mathbf{I}) \rightarrow Dom(O)$ is monotone relative to $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$ and if there exists a set of variables $\mathbf{C} \subseteq \mathbf{V} - \{\mathbf{K} \cup \mathbf{X}\}$ such that $\mathbf{C} \cup \mathbf{K}$ satisfies the backdoor-criterion relative to $\mathbf{X}$ and $\mathbf{I}$ in G, the following holds:

$$\mathrm{NEC}_{\mathbf{x}}(\mathbf{k}) = \frac{\left(\sum_{\mathbf{c}\in Dom(\mathbf{C})} \Pr(o' \mid \mathbf{c},\mathbf{x}',\mathbf{k})\Pr(\mathbf{c}\mid\mathbf{x},\mathbf{k})\right) - \Pr(o'\mid\mathbf{x},\mathbf{k})}{\Pr(o\mid\mathbf{x},\mathbf{k})} \quad (19)$$

$$\mathrm{SUF}_{\mathbf{x}}(\mathbf{k}) = \frac{\left(\sum_{\mathbf{c}\in Dom(\mathbf{C})} \Pr(o\mid\mathbf{c},\mathbf{x},\mathbf{k})\Pr(\mathbf{c}\mid\mathbf{x}',\mathbf{k})\right) - \Pr(o\mid\mathbf{x}',\mathbf{k})}{\Pr(o'\mid\mathbf{x}',\mathbf{k})} \quad (20)$$

$$\mathrm{NESUF}_{\mathbf{x}}(\mathbf{k}) = \sum_{\mathbf{c}\in Dom(\mathbf{C})} \left(\Pr(o\mid\mathbf{x},\mathbf{k},\mathbf{c}) - \Pr(o\mid\mathbf{x}',\mathbf{c},\mathbf{k})\right)\Pr(\mathbf{c}\mid\mathbf{k}) \quad (21)$$

Proposition 4.2 facilitates bounding and estimating explanation scores from historical data when the underlying probabilistic causal models are not fully specified but background knowledge on the causal diagram is available. (See Section 6 for a discussion about the absence of causal diagrams). We establish the following connection between the explanation scores.

Proposition4.3. Explanation scores are related through the following inequality. For a binary X, the inequality becomes an equality.

$$\mathrm{NESUF}_{\mathbf{x}}(\mathbf{k}) \leq \Pr(o,\mathbf{x}\mid\mathbf{k})\,\mathrm{NEC}_{\mathbf{x}}(\mathbf{k}) + \Pr(o',\mathbf{x}'\mid\mathbf{k})\,\mathrm{SUF}_{\mathbf{x}}(\mathbf{k}) + 1 - \Pr(\mathbf{x}\mid\mathbf{k}) - \Pr(\mathbf{x}'\mid\mathbf{k}) \quad (22)$$

Therefore, for binary attributes, the necessary and sufficiency score can be seen as the weighted sum of necessary and sufficiency scores. Furthermore, the lower bound for the necessity and sufficiency score in Equation (11) is called the (conditional) *causal effect* of X on O [72]. Hence, if the causal effect of X on the algorithm's decision is non-zero, then so is the necessity and sufficiency score (for a binary X, it is implied from (22) that at least one of the sufficiency and necessity scores must also be non-zero). The following proposition shows the converse.

Proposition4.4. Given a PCM $\langle M, \Pr(\mathbf{u})\rangle$ with a corresponding causal DAG G, an algorithm $f : Dom(\mathbf{Z}) \rightarrow Dom(O)$ and an attribute $X \in \mathbf{V}$, if O is a non-descendant of X, i.e., there is no causal path from X to O, then for all $(x,x') \in Dom(X)$ and for all contexts $\mathbf{k} \in Dom(\mathbf{K})$, where $\mathbf{K} \subseteq \mathbf{V} - \{X, O\}$, it holds that $\mathrm{NEC}_{X}(\mathbf{k}) = \mathrm{SUF}_{X}(\mathbf{k}) = \mathrm{NESUF}_{X}(\mathbf{k}) = 0$.

Extensions to multi-class classification and regression. For multi-valued outcomes, i.e., $Dom(O) = \{o_1, \ldots, o_\gamma\}$, we assume an ordering of the values $o_1 > \ldots > o_\gamma$ such that $o_i > o_j$ implies that o_i is more desirable than o_j. This assumption holds in tasks where certain outcomes are favored over others and holds for real-valued outcomes that have a natural ordering of values. We partition $Dom(O)$ into sets $O^<$ and $O^\geq$ where $O^<$ denotes the set of values less than o and $O^\geq$ denotes the set of values greater than o. Note

[5] Monotonicity expresses the assumption that changing $\mathbf{X}$ from $\mathbf{x}'$ to $\mathbf{x}$ cannot change the algorithm's decision from positive to negative; increasing $\mathbf{X}$ always helps.

that we do not require a strict ordering of the values and can simply partition them as *favorable* and *unfavorable*. In these settings, we redefine the explanation scores with respect to each outcome value o. For example, necessity score is defined as the probability that the outcome O changes from a value greater than or equal to o to a value lower than o upon the intervention $X \leftarrow x'$:

$$\text{NEC}_x^{x'}(\mathbf{k}, o) \stackrel{\text{def}}{=} \Pr(O^{<}_{X \leftarrow x'} \mid x, O^{\geq}, \mathbf{k})$$

The other two scores can be extended in a similar fashion. Our propositions extend to these settings and can be directly used to evaluate the explanation scores using observational data.

4.2 Computing Counterfactual Recourse

Here, we describe the solution to the optimization problem discussed in (8) for providing an actionable recourse. We formulate our problem as a combinatorial optimization problem over the domain of actionable variables and express it as an integer programming (IP) problem of the form:

$$\underset{\hat{\mathbf{a}} \in Dom(\mathbf{A})}{\text{argmin}} \quad \sum_{A \in \mathbf{A}} \left(\phi_A \sum_{a \in Dom(A)} \delta_a \right) \tag{23}$$

$$\text{subject to} \qquad \text{SUF}_{\hat{\mathbf{a}}}(\mathbf{v}) \geq \alpha \tag{24}$$

$$\sum_{a \in Dom(A)} \delta_a \leq 1, \quad \forall A \in \mathbf{A} \tag{25}$$

$$\delta_a \in \{0, 1\}, \quad \forall a \in Dom(A), A \in \mathbf{A} \tag{26}$$

The objective function in the preceding IP is modeled as a linear function over the cost of actions over individual actionable variables. ϕ_A is a convex cost function that measures the cost of changing $A = a$ to $A = \hat{a}$, for each $A \in A$ ($\phi_A = 0$ when no action is taken on A) and can be predetermined as $\hat{a}$ deviates from a (e.g., the cost could increase linearly or quadratically with increasing deviation from $A = a$). Constraint (24) ensures that action $\hat{\mathbf{a}}$ will result in a sufficiency score greater than the user-defined threshold α. In other words, the intervention $\mathbf{A} \leftarrow \hat{\mathbf{a}}$ can lead to the positive outcome with a probability of at least α. Constraint (25) and indicator variables δ_a ensure that of all values in the domain of an actionable variable, only one is acted upon (or changed). Note that the IP formulation ensures that multiple actions can be taken at the same time. In particular, when $\delta_a = 0$, $\forall a \in Dom(A)$, $\forall A \in \mathbf{A}$, it implies no action is taken since (24) is already satisfied. To compute the sufficiency score in (24) from historical data, we rewrite it as $\text{SUF}_{\hat{\mathbf{a}}}(\mathbf{k} \cup \mathbf{a}) \geq \alpha$, where $\mathbf{K}$ consists of all non-descendants of $\mathbf{A}$ in the underlying causal diagram G, and we assume that $\mathbf{K}$ satisfies the backdoor-criterion relative to O and $\mathbf{A}$ (cf. Section 2). (See Section 6 for a discussion about violation of the assumptions.) Then, we can incorporate the *lower bound* obtained for the sufficiency score in Proposition 4.2 in the optimization problem, as follows:

$$\Pr(o \mid \hat{\mathbf{a}}, \mathbf{k}) \geq \Pr(o \mid \mathbf{a}, \mathbf{k}) + \alpha \Pr(o' \mid \mathbf{a}, \mathbf{k}) \tag{27}$$

Since $\mathbf{k}, \mathbf{a}, \alpha$ are constant, the RHS of (27) is also constant and can be pre-computed from data. We estimate the logit transformation of $\Pr(o \mid \hat{\mathbf{a}}, \mathbf{k})$ and model it as a linear regression equation. This allows us to express (27) as a linear inequality constraint for the IP in (23). If a solution to the IP is found, then an action is performed on each variable for which the indicator variable has a non-zero assignment. *The solution to this optimization problem can be seen as a recourse that can change the outcome of the algorithm with high probability for individuals with attributes* $\mathbf{k}$ *for which the algorithm made a negative decision.* Note that the number of constraints in this formulation grows linearly with the number of actionable variables (which is usually a much smaller subset of an individual's attributes).

Dataset	Att. [#]	Rows[#]	Global	Local	Recourse
Adult [53]	14	48k	7.5	4.2	3.7
German [21]	20	1k	0.75	0.42	2.24
COMPAS [1]	7	5.2k	2.03	1.34	-
Drug [21]	13	1886	1.25	0.84	-
German-syn	6	10k	1.35	1.01	1.65

Table 2: Runtime in seconds for experiments in Sec. 5.3.

5 EXPERIMENTS

This section presents experiments that evaluate the effectiveness of LEWIS. We answer the following questions. **Q1:** What is the end-to-end performance of LEWIS in terms of gaining insight into black-box machine learning algorithms? How does the performance of LEWIS change with varying machine learning algorithms? **Q2:** How does LEWIS compare to state-of-the-art methods in XAI? **Q3:** To what extent are the explanations and recourse options generated by LEWIS correct?

5.1 Datasets

We used the following ML benchmark datasets (also in Table 2):

German Credit Data (German) [21]. This dataset consists of records of bank account holders with their personal,fi nancial and demographic information. The prediction task classifies individuals as good/bad credit risks.

Adult Income Data (Adult) [21]. This dataset contains demographic information of individuals along with information on their level of education, occupation, working hours etc. The task is to predict whether the annual income of an individual exceeds 50K.

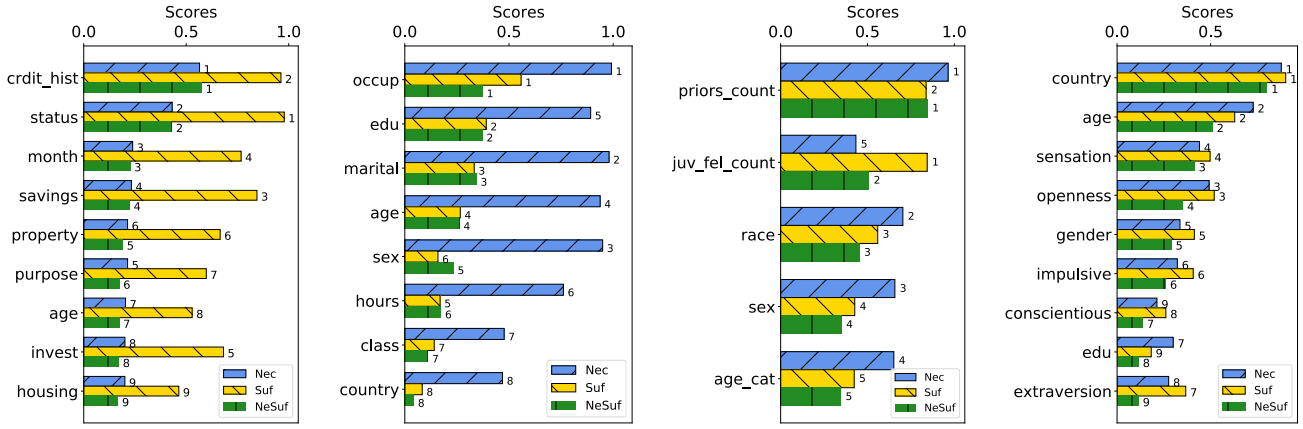
COMPAS [1]. This dataset contains information on offenders from Broward County, Florida. We consider the task of predicting whether an individual will recommit a crime within two years.

Drug Consumption (Drug) [21]. This dataset contains demographic information and personality traits (e.g., openness, sensation-seeking) of individuals. We consider a multi-class classification task of predicting when an individual consumed magic mushrooms: (i) never, (ii) more than a decade ago, and (iii) in the last decade.

German-syn. We generate synthetic data following the causal graph of the `German` dataset. The black-box algorithm runs random forest regressor to predict credit risk score within the range $[0, 1]$ where 1 denotes the best and 0 denotes the worst credit risk. We use this dataset to evaluate the correctness of LEWIS's scores compared to ground truth scores calculated using the structural equations.

5.2 Setup

We considered four black-box machine learning algorithms: random forest classifier [94], random forest regression [94], XGBoost [95], and a feed forward neural network [93]. We used causal diagrams presented in [12] for the Adult and German datasets and in [69] for COMPAS. For the Drug dataset, Country, Age, Gender and Ethnicity are considered root nodes that affect the outcome and other attributes; the outcome is also affected by all other attributes. We implemented our scores and the recourse algorithm in Python. We split each dataset into training and test data, learned a black-box algorithm (random forest classifier unless stated otherwise) over training data, and estimated conditional probabilities in (19)-(21) by regressing over test data predictions. We report explanation scores for each dataset under different scenarios. To present local explanations, we report the positive and negative contributions of an attribute value toward the current outcome (e.g., in Figure 5,



a) German b) Adult c) Compas – Software score d) Drug

Figure 3: Global explanations generated by Lewis, ranking attributes in terms of their necessity, sufficiency and necessity and sufficiency scores. The rankings are consistent with insights from existing literature.

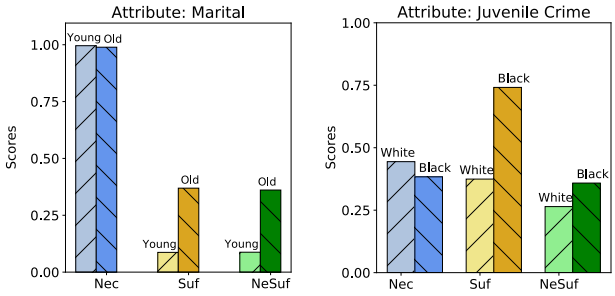
bars to the left (right) represent negative (positive) contributions of attribute values). To recommend recourse to individuals who receive a negative decision, we generate a set of actions with the minimal cost that, if taken, can change the algorithm's decision for them in the future with a user-defined probability threshold α.

5.3 End-to-End Performance

In the following experiments, we present the local, contextual and global explanations and recourse options generated by Lewis. In the absence of ground truth, we discuss the coherence of our results with intuitions from existing literature. Table 2 reports the running time of Lewis for computing explanations and recourse.

German. Consider the attributes `status` and `credit_history` in Figure 3a. Their near-perfect sufficiency scores indicate their high importance toward a positive outcome at the population level. For individuals for whom the algorithm generated a negative outcome, an increase in their credit history or maintaining above the recommended daily minimum in checking accounts (`status`) is more likely to result in a positive decision compared to other attributes such as housing or age. These scores and the low necessity scores of attributes are aligned with our intuition about good credit risks: (a) good credit history and continued good status of checking accounts add to the credibility of individuals in repaying their loans, and (b) multiple attributes favor good credit and a reduction in any single attribute is less likely to overturn the decision.

We report the local explanations generated by Lewis in Figures 8a and 8b. In the real world, younger individuals and individuals with inadequate employment experience or insufficient daily minimum amount in checking accounts are less likely to be considered good credit risks. This observation is evidenced in the negative contribution of status, age and employment for the negative outcome example. For the positive outcome example, current attribute values contribute toward the favorable outcome. Since increasing any of them is unlikely to further improve the outcome, the values do not have a negative contribution. Figure 1 presents

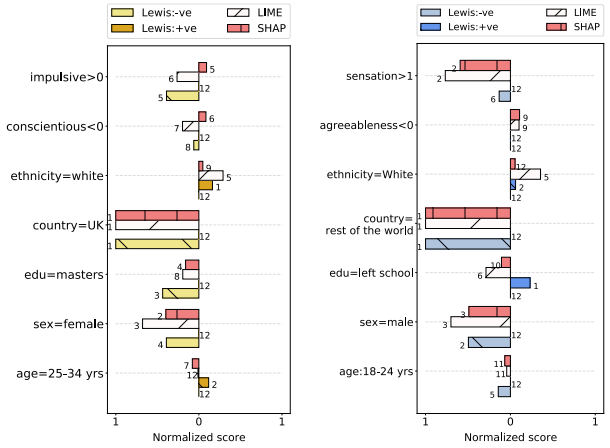


a) Effect of marital on different age groups. (Adult) b) Effect of juvenile crime on race. (COMPAS)

Figure 4: Lewis's contextual explanations show the effect of intervening on an attribute over different sub-populations.

an example actionable recourse scenario, suggesting an increase in savings, credit amount and purpose improves credit risk.

Adult. Several studies [92, 103] have analyzed the impact of gender and age in this dataset. The dataset has been shown to be inconsistent: income attributes for married individuals report household income, and there are more married males in the dataset indicating a favorable bias toward males [81]. We, therefore, expect age to be a necessary cause for higher income, but it may not be sufficient since increasing age does not imply that an individual is married. This intuition is substantiated by the high necessity and low sufficiency scores of age in Figure 3b. Furthermore, as shown in Figure 4a, changing marital status to a higher value has a greater effect on older than on younger individuals; this effect can be attributed to the fact that compared to early-career individuals, mid-career individuals typically contribute more to joint household income. Consequently, for an individual with a negative outcome (Figure 8c), marital status and age contribute toward the negative outcome. For an individual with a positive outcome (Figure 8d), changing any



a) Negative outcome example. **b) Positive outcome example.**

Figure 5: Lewis's local explanations. (Drug)

a) Adult + XGBoost **b) Adult + Neural Networks**

Figure 6: Generalizability of Lewis to black-box algorithms.

attribute value is less likely to improve the outcome. However, increasing working hours will further the favorable outcome with a higher probability. We calculated the recourse for the individual with negative outcome and identified that increasing the hours to more than 42 would result in a high-income prediction.

COMPAS. In the global explanation scores generated by Lewis for the COMPAS software used in courts (Figure 3c), the highest score of `priors_ct` validates the insights of previous studies [1, 84] that the number of prior crimes is one of the most important factors determining chances of recidivism. Figure 4b presents the effect of intervening on juvenile felonies count on the software score (for these explanations, we use prediction scores from the COMPAS software, not the classifier output). The higher sufficiency for Blacks compared to Whites indicates that an increase in juvenile crimes is more detrimental for the former. A reduction in the number, however, benefits the latter more, thereby validating the inherent bias in COMPAS scores. We did not perform recourse analysis as the attributes describe past crimes and, therefore, are not actionable.

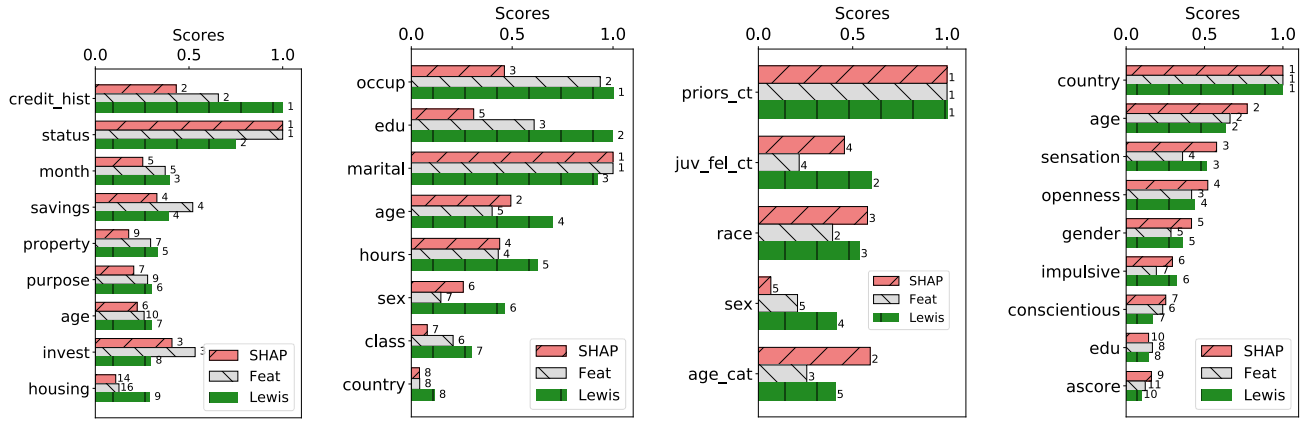
Drug. This dataset has been studied to understand the variation in drug patterns across demographics and the effectiveness of various sensation measurement features toward predicting drug usage. Figure 3d compares the global scores with respect to the outcome that the drug was used atleast once in lifetime. Previous studies [23] have found that consumption of the particular drug is common in certain countries, as substantiated by the high necessity and sufficiency scores of country. Furthermore, intuitively, individuals with a higher level of education are more likely to be aware of the effects of drug abuse and hence, less likely to indulge in its consumption. This intuition is supported by the observation in Figure 5a: a higher education level contributes toward the negative drug consumption outcome, and in Figure 5b: a lower education level contributes positively toward the drug consumption outcome. We observe similar conclusions for the explanations with respect to a different outcome such as drug used in the last decade.

Generalizability of Lewis to black-box algorithms. In Figure 6, we present the global explanations generated by Lewis for black-box algorithms that are harder to interpret and are likely to violate the monotonicity assumption, such as XGBoost and feed forward neural networks, and report the necessity and sufficiency score for each classifier. For ease in deploying neural networks, we conducted this set of experiments on `Adult` which is our largest dataset. We observed that different classifiers rank attributes differently depending upon the attributes they deem important. For example, the neural network learns class as the most important attribute. Since country and sex have a causal effect on class, Lewis ranks these three attributes higher than others (see Section 5.4 for a detailed interpretation of the results).

Key takeaways. (1) The explanations generated by Lewis capture causal dependencies between attributes, and are applicable to any black-box algorithm. (2) Lewis has proved effective in determining attributes causally responsible for a favorable outcome. (3) Its contextual explanations, that show the effect of particular interventions on sub-populations, are aligned with previous studies. (4) The local explanations offerfi ne-grained insights into the contribution of attribute values toward the outcome of an individual. (5) For individuals with an unfavorable outcome, whenever applicable, Lewis provides recourse in the form of actionable interventions.

5.4 Comparing Lewis to Other Approaches

We compared the global and local explanations generated by Lewis to existing approaches used for interpreting ML algorithms: `SHAP` [59], `LIME` [75] and feature importance (`Feat`) [11]. `SHAP` explains the difference between a prediction and the *global* average prediction, `LIME` explains the difference from a *local* average prediction, and `Feat` measures the increase in an algorithm's prediction error after permutating an attribute's values. `LIME` provides local explanations, `Feat` provides global explanations and `SHAP` generates both global and local explanations. `LIME` and `SHAP` provide marginal contribution of an attribute to classifier prediction and are not directly comparable to Lewis's probabilistic scores. However, since all the methods measure the importance of attributes in classifier predictions, we report the relative ranking of attributes generated on their normalized scores, and present a subset of attributes ranked high by any of the methods. We report the maximum NeSuf$_x$ score of an attribute obtained by Lewis on all of its value pairs. We also



a) German b) Adult c) Compas – Software score d) Drug

Figure 7: Comparing different global explanation methods: SHAP and Feat fail to account for causal relationships in the data that are effectively captured by Lewis.

compared the recourse generated by Lewis with LinearIP. (We contacted the authors of [46] but do not use it in evaluation since their technique does not work for categorical actionable variables). We used open-source implementations of the respective techniques.

German. In Figure 7a, note that housing is ranked higher by Lewis than by Feat and SHAP. The difference lies in the data: housing=own is highly correlated with a positive outcome. However, due to a skewed distribution (there are ∼ 10% of instances where housing=own), random permutations of housing do not generate new instances, and Feat is unable to identify it as an important attribute. Lewis uses the underlying causal graph to capture the causal relationship between the two attributes.

In Figures 8a and 8b, we report the rankings obtained by LIME, SHAP and Lewis on two instances that respectively have negative and positive predicted outcomes. Employment, age and account status have a high negative contribution toward the outcome in Figure 8a, indicating that increasing them is likely to reverse the decision. Intuitively, with age, continued employment and improved account status, individuals tend to have better savings, credit history, housing, etc., which, in turn, contribute toward a positive outcome. Lewis's ranking captures this causal dependency between the attributes, which is recorded by neither SHAP nor LIME.

To compare recourse generated by Lewis and LinearIP, we tested them on the example for Maeve in Figure 1. While both the methods identify the same solution for small thresholds, LinearIP did not return any solution for success threshold > 0.8. In contrast to Lewis that generalizes to black-box algorithms, LinearIP depends on linear classifiers and offers recommendations that do not account for the causal relationship between attributes.

Adult. In Figure 7b, the ranking of attributes generated by Lewis and Feat matches observations in prior literature that consider occupation, education and marital status to be the most important attributes. However, SHAP picks on the *correlation* of age with marital status and occupation (older individuals are more likely to be married and have better jobs), and ranks it higher. The rankings are similar for XGBoost (Figure 6a) and Random forest (Figure 7b) but different for the neural network (Figure 6b). We investigated the outputs and observed that the prediction of neural networks differs from that of random forest and XGBoost for more than 20% of the test samples, leading to varied ranking of attributes. Additionally, the class of an individual is ranked important by the classifier. Since country and sex have a causal impact on class, it justifies their high ranks as generated by Lewis. In Figure 6b, we do not report the scores for Feat as it does not support neural networks.

In Figures 8c and 8d, we compare Lewis with local explanation methods LIME and SHAP. Consistent with existing studies, Lewis recognizes the negative contribution of unmarried marital status and positive contribution of sex=male toward the negative outcome. For the positive outcome example, Lewis identifies that age, sex and country have a high positive contribution toward the outcome due to their causal impact on attributes such as occupation and marital status (ranked higher by SHAP and LIME). We also observed that the results of SHAP are not stable across different iterations.

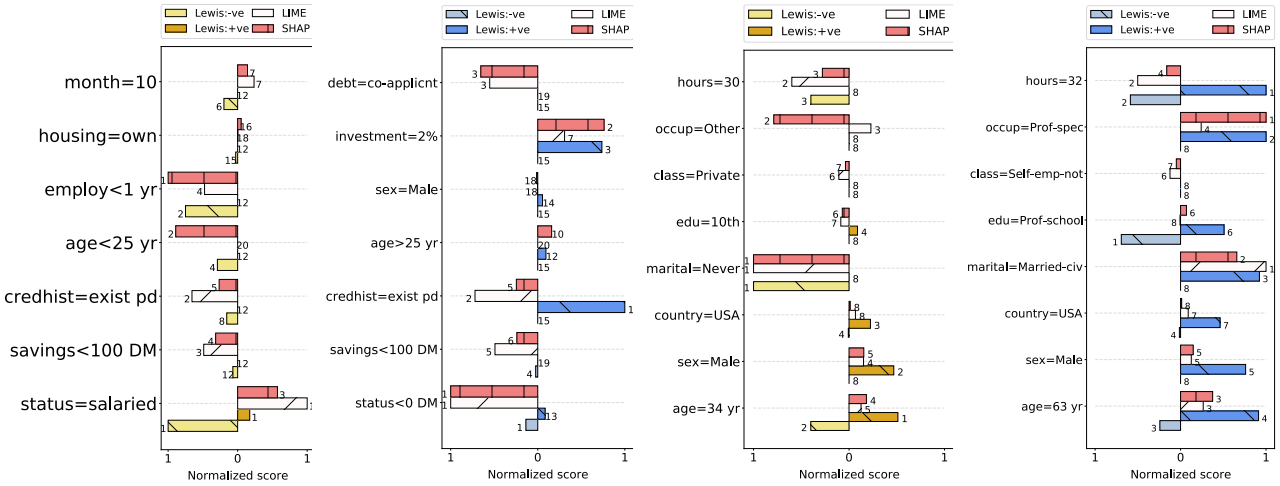
COMPAS. Since COMPAS scores were calculated based on criminal history and indications of juvenile delinquency [1], the higher ranking of juvenile crime history by Lewis is justified in Figure 7c. Bias penetrated into the system due to the correlation between demographic and non-demographic attributes. SHAP and Feat capture this correlation and rank age higher than juvenile crime history.

Drug. Figure 7d shows that all techniques have a similar ordering of attributes with country and age being most crucial for the desired outcome. Comparing the local explanations of Lewis with SHAP and LIME (Figure 5), we observe that Lewis correctly identifies the negative contribution of higher education toward negative drug consumption prediction and the positive contribution of a lower level of education toward a positive drug consumption prediction.

5.5 Correctness of Lewis's explanations

Since ground truth is not available in real-world data, we evaluate the correctness of Lewis on the German-Syn dataset.

Correctness of estimated scores. In Figure 9a, we compare the global explanation scores of different variables with ground truth necessity and sufficiency score estimated using Pearl's three-step



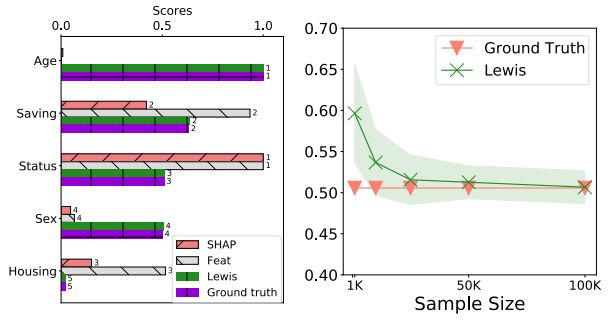
a) Negative outcome (German) **b) Positive outcome (German)** **c) Negative outcome (Adult)** **d) Positive outcome (Adult)**

Figure 8: Comparing different local explanation methods. SHAP and LIME explain the output of an instance in terms of its difference from the global or local average prediction; LEWIS explains it in terms of the underlying causal graph.

procedure discussed in equation (3) (Section 2). We present the comparison for a non-linear regression based black-box algorithm with respect to outcome $o = 0.5$. The average global explanation scores returned by LEWIS are consistently similar to ground truth estimates, thereby validating the correctness of Proposition 4.2. SHAP and Feat capture the correlation between the input and output attributes, and rank Status higher than Age and Sex which are assigned scores close to 0. These attributes do not directly impact the output but indirectly impact it through Status and Saving. This experiment validates the ability of LEWIS in capturing causal effects between different attributes, estimate explanation scores accurately and present actionable insights as compared to SHAP and Feat. To understand the effect of the number of samples on the scores estimated by LEWIS, we compare the NESUF scores of status for different sample sizes in Figure 9b. We observe that the variance in estimation is reduced with an increase in sample size and scores converge to ground truth estimates for larger samples.

Robustness to violation of monotonicity. To evaluate the impact of non-monotonicity on the explanation scores generated by LEWIS, we changed the structural equations for the causal graph of German-Syn to simulate non-monotonic effect of Age on the prediction attribute. This data was used to train random forest and XGBoost classifiers. We measured *monotonicity violation* as $\Lambda_{\texttt{viol}} = \Pr[o'_{X \leftarrow x} | o, x']$. Note that $\Lambda_{\texttt{viol}} = 0$ implies monotonicity and higher $\Lambda_{\texttt{viol}}$ denotes higher violation of monotonicity. We observed that the scores estimated by LEWIS differ from ground truth estimates by less than 5%, as long as the monotonicity violation is less than 0.25. Furthermore, the relative ranking of the attributes remains consistent with the ground truth ranking calculated using equation (3). This experiment demonstrates that the explanations generated by LEWIS are robust to slight violation in monotonicity.

Recourse analysis. We sampled 1000 random instances that received negative outcomes and generated recourse (sufficiency threshold $\alpha = 0.9$) using LEWIS. Each unit change in attribute value was



a) Quality of the estimates. **b) Effect of sample size on error.**

Figure 9: Comparing with ground truth.

assigned unit cost. The output was evaluated with respect to the ground truth sufficiency and cost of returned actions. In all instances, LEWIS's output achieved more than 0.9 sufficiency with the optimal cost. This experiment validates the optimality of the IP formulation in generating effective recourse. To further test the scalability of LEWIS, we considered a causal graph with 100 variables and increased the number of actionable variables from 5 to 100. The number of constraints grew linearly from 6 to 101 (one for each actionable variable and one for the sufficiency constraint), and the running time increased from 1.65 seconds to 8.35 seconds, demonstrating LEWIS's scalability to larger inputs.

6 RELATED WORK AND DISCUSSION

Our research is mainly related to XAI work in quantifying feature importance and counterfactual explanations.

Quantifying feature importance. Due to its strong axiomatic guarantees, methods based on Shapley values are emerging as the de facto approach for quantifying feature influence [2, 17, 26, 54, 57, 58, 62, 90]. However, several practical and epistemological issues have been identified with these methods. These issues arise

primarily because existing proposals for quantifying the marginal influence of an attribute do not have any causal interpretation in general and, therefore, can lead to incorrect and misleading explanations [26, 48, 62]. Another popular method for generating local explanations is LIME (Local Interpretable Model-agnostic Explanations [75], which trains an interpretable classifier (such as linear regression) on an instance obtained by perturbing that instance to be explained around its neighborhood. Several issues with LIME have also been identified in the literature, including its lack of human interpretability, its sensitivity to the choice of local perturbation, and its vulnerability to adversarial attacks [4, 58, 65, 88].

Unlike existing methods, our proposal offers the following advantages. (1) It is grounded in causality and counterfactual reasoning, captures insights from the theoretical foundation of explanations in philosophy, epistemology and social science, and can provide provably correct explanations. It has been argued that humans are selective about explanations and, depending on the context, certain contrasts are more meaningful than others [22, 63]. The notions of necessity and sufficiency have been shown to be strong criteria for preferred explanatory causes [14, 32, 33, 71, 77, 78, 91]. (2) It accounts for indirect influence of attributes on algorithm's decisions; the problem of quantifying indirect influence has received scant attention in XAI literature (see [3] for a non-causality-based approach). (3) It builds upon scores that are customizable and can therefore generate explanations at the global, contextual and local levels. (4) It can audit black-box algorithms merely by using historical data on its input and outputs.

Counterfactual explanations. Our work is related to a line of research that leverages counterfactuals to explain ML algorithm predictions [10, 44, 51, 60, 68, 96, 100, 101]. In this context, the biggest challenge is generating explanations that follow natural laws and are feasible and actionable in the real world. Recent work attempts to address feasibility use ad hoc constraints [20, 43, 56, 68, 96, 98]. However, it has been argued that feasibility is fundamentally a causal concept [8, 44, 60]. Few attempts have been made to develop a causality-based approach that can generate actionable recourse by relying on the strong assumption that the underlying probabilistic causal model is fully specified or can be learned from data [44, 46, 60]. Our framework extends this line of work by (1) formally defining feasibility in terms of probabilistic contrastive counterfactuals, and (2) providing a theoretical justification for taking a fully non-parametric approach for computing contrastive counterfactuals from historical data, thereby making no assumptions about the internals of the decision-making algorithm and the structural equations in the underlying probabilistic causal models. As an independent work, Mothilal et al. [67] proposed a notion of necessity and sufficiency scores for quantifying feature importance that appeals to the notion of actual causation.

Logic-based methods. Our work shares some similarities with recent work in XAI that employs tools from logic-based diagnosis and operates with the logical representations of ML algorithms [15, 41, 87]. In this context, the fundamental concepts of prime implicate/implicant are closely related to sufficiency and necessary causation when the underlying causal model is a *logical circuit* [16, 19, 37, 37, 40]. However, these methods can generate explanations only in terms of a set of attributes, are intractable in model-agnostic settings, fail to account for the causal interaction between attributes, and cannot go beyond deterministic algorithms.

Algorithmic fairness. The critical role of causality and background knowledge is recognized and acknowledged in the algorithmic fairness literature [27, 28, 47, 49, 69, 79, 80, 82, 83, 85]. In this context, contrastive counterfactuals have been used to capture individual-level fairness [13, 50]. It is easy to show that the notion of *counterfactual fairness* in [50] can be captured by the explanation scores introduced in this paper provided that an algorithm is counterfactually fair w.r.t. a protected attribute if the sufficiency score and necessity score of the sensitive attribute are *both* zero. Hence, Lewis is useful for reasoning about individual-level fairness and discrimination.

Probability of Causation. The metrics we introduce here for quantifying the necessity, sufficiency and necessity and sufficiency of an algorithm's input for its decisions are adopted from the literature on probability of causation [14, 32, 33, 71, 77, 78, 91]. The results developed in Section 4.1 generalize and subsume earlier results from [71, 91] and substantially simplify their proofs.

Assumptions and limitations. Our framework relies on two main assumptions to estimate and bound explanation scores, namely, the availability of (1) data that is a representative sample of the underlying population of interest, and (2) knowledge of the underlying causal diagram. Dealing with non-representative samples goes beyond the scope of this paper, but there are standard approaches that can be adopted (see, e.g., [7]). Furthermore, Lewis is designed to work with any level of user's background knowledge. If no background knowledge is provided, Lewis assumes no-confounding, i.e., $\Pr(o \mid \mathsf{do}(x), \mathbf{k}) = \Pr(o \mid x, \mathbf{k})$ and monotonicity. Under these assumptions, the necessity score and sufficiency score, respectively, become $\frac{\Pr(o'|\mathbf{x}',\mathbf{k})-\Pr(o'|\mathbf{x},\mathbf{k})}{\Pr(o|\mathbf{x},\mathbf{k})}$ and $\frac{\Pr(o|\mathbf{x},\mathbf{k})-\Pr(o|\mathbf{x}',\mathbf{k})}{\Pr(o'|\mathbf{x}',\mathbf{k})}$. When computed for individuals, these quantities can be interpreted as proportional to the difference between the ratio of positive/negative algorithmic decisions for individuals that are *similar* on all attributes except for X. In other words, the quantities measure the correlation between X and the algorithm's decisions across similar individuals. This correlation can be interpreted causally only under the no-confounding and monotonicity assumptions. *Nonetheless, quantifying the local influence of an attribute by measuring its correlation with an algorithm's decision across similar individuals underpins most existing methods for generating local explanations such as Shapley values based methods [2, 58], feature importance [90] , and LIME [75].* Approaches differ in terms of how they measure this correlation.

In principle, background knowledge on underlying causal models is required to generate effective and actionable explanations. *While this may be considered a limitation of our approach, we argue that all existing XAI methods either explicitly or implicitly make causal assumptions (such as those mentioned above in addition to feature independence and the possibility of simulating interventional distributions by perturbing data or using marginal distributions).* Our framework replaces assumptions that are unrealistic with assumptions about the underlying causal diagram that need not be perfect, can be validated using historical data and background knowledge [71], and can be learned from a mixture of historical and interventional data [30]. In the worst case, our assumptions about generating local explanations are similar to those of existing work.

REFERENCES

[1] Machine bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[2] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

[3] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

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

[5] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*, 2016.

[6] Elias Bareinboim, JD Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020.

[7] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108, 2012.

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

[9] Richard Berk. *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer, 2019.

[10] Leopoldo E. Bertossi, Jordan Li, Maximilian Schleich, Dan Suciu, and Zografoula Vagena. Causality-based explanation of classification outcomes. In *Proceedings of the Fourth Workshop on Data Management for End-To-End Machine Learning, In conjunction with the 2020 ACM SIGMOD/PODS Conference, DEEM@SIGMOD 2020, Portland, OR, USA, June 14, 2020*, pages 6:1–6:10, 2020.

[11] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[12] Silvia Chiappa. Path-specific counterfactual fairness. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7801–7808. AAAI Press, 2019.

[13] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[14] Louis Anthony Cox Jr. Probability of causation and the attributable proportion risk. *Risk Analysis*, 4(3):221–230, 1984.

[15] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. *arXiv preprint arXiv:2002.09284*, 2020.

[16] Adnan Darwiche and Judea Pearl. Symbolic causal networks. In *AAAI*, pages 238–244, 1994.

[17] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[18] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

[19] Johan De Kleer, Alan K Mackworth, and Raymond Reiter. Characterizing diagnoses and systems. *Artificial intelligence*, 56(2-3):197–222, 1992.

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

[21] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[22] Curt J Ducasse. On the nature and the observability of the causal relation. *The Journal of Philosophy*, 23(3):57–68, 1926.

[23] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk, 2017.

[24] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective. *arXiv preprint arXiv:1801.01489*, 68, 2018.

[25] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[26] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

[27] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

[28] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Fair data integration. *arXiv preprint arXiv:2006.06053*, 2020.

[29] Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*, 2015.

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

[31] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[32] Sander Greenland. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *American journal of public health*, 89(8):1166–1169, 1999.

[33] Sander Greenland and James M Robins. Epidemiology, justice, and the probability of causation. *Jurimetrics*, 40:321, 1999.

[34] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

[35] Eric Grynaviski. Contrasts, counterfactuals, and causes. *European Journal of International Relations*, 19(4):823–846, 2013.

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

[37] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 56(4):889–911, 2005.

[38] Giles Hooker. Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580, 2004.

[39] Giles Hooker and Lucas Mentch. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

[40] Mark Hopkins and Judea Pearl. Clarifying the usage of structural models for commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, pages 83–89. AAAI Press Menlo Park, CA, 2003.

[41] Alexey Ignatiev. Towards trustable explainable ai. In *29th International Joint Conference on Artificial Intelligence*, pages 5154–5158, 2020.

[42] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR, 2020.

[43] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[44] Amir-Hossein Karimi, Gilles Barthe, Borja Belle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

[45] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

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

[47] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[48] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[49] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[50] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pages 4069–4079, 2017.

[51] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

[52] David K Lewis. Causal explanation. 1986.

[53] M. Lichman. Uci machine learning repository, 2013.

[54] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[55] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

[56] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. *arXiv preprint arXiv:1907.03077*, 2019.

[57] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[58] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[59] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, pages 4765–4774, 2017.

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

[61] David R Mandel. Counterfactual and causal explanation. *Routledge research international series in social psychology. The Psychology of Counterfactual Thinking*, pages 11–27, 2005.

[62] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models with cooperative game theory. *arXiv preprint arXiv:1909.08128*, 2019.

[63] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[64] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[65] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[66] Adam Morton. Contrastive knowledge. *Contrastivism in philosophy*, pages 101–115, 2013.

[67] Ramaravind K Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv preprint arXiv:2011.04917*, 2020.

[68] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[69] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[70] Judea Pearl. Direct and indirect effects. In Jack S. Breese and Daphne Koller, editors, *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pages 411–420. Morgan Kaufmann, 2001.

[71] Judea Pearl. *Causality*. Cambridge university press, 2009.

[72] Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

[73] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

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

[75] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[76] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535, 2018.

[77] David W Robertson. Common sense of cause in fact. *Tex. L. Rev.*, 75:1765, 1996.

[78] James Robins and Sander Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, pages 1125–1138, 1989.

[79] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[80] Babak Salimi, Corey Cole, Peter Li, Johannes Gehrke, and Dan Suciu. Hypdb: a demonstration of detecting, explaining and resolving bias in olap queries. *Proceedings of the VLDB Endowment*, 11(12):2062–2065, 2018.

[81] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in OLAP queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1021–1035, 2018.

[82] Babak Salimi, Bill Howe, and Dan Suciu. Database repair meets algorithmic fairness. *ACM SIGMOD Record*, 49(1):34–41, 2020.

[83] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. Causal relational learning. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 241–256, 2020.

[84] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

[85] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810. ACM, 2019.

[86] Andrew D Selbst and Solon Barocas. The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085, 2018.

[87] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364*, 2018.

[88] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[89] Kacper Sokol and Peter A Flach. Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In *SafeAI@ AAAI*, 2019.

[90] ErikŠ trumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[91] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[92] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2017.

[93] https://docs.fast.ai/tabular.learner.htm. Fastai neural network.

[94] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. sklearn python library.

[95] https://xgboost.readthedocs.io/en/latest/. Xgboost.

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

[97] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. Websites vary prices, deals based on users' information. *Wall Street Journal*, 10:60–68, 2012.

[98] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

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

[100] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

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

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

[103] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, page 992–1001, USA, 2011. IEEE Computer Society.