Feature Attribution and Recourse via Probabilistic Contrastive Counterfactuals

Sainyam Galhotra [*1] **Romila Pradhan** [*2] **Babak Salimi** [2]

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 has focused on two main approaches: (1) Attribution of *responsibility* for an algorithm's decisions to its inputs, wherein responsibility is typically approached as a purely *associational* concept that can lead to misleading conclusions. (2) Generating counterfactual explanations and recourse, where these explanations are typically obtained by considering the smallest perturbation in an algorithm's input that can lead to the algorithm's desired outcome. However, these perturbations may not translate to real-world interventions. In this paper, we propose a principled and novel causality-based approach for explaining black-box decision-making systems that exploit *probabilistic contrastive counterfactuals* to provide a unifying framework to generate wide ranges of global, local and contextual explanations that provide insights into what causes an algorithm's decisions, and generate actionable recourse translatable into real-world interventions.

1. Introduction

Algorithmic decision-making systems are increasingly used to automate consequential decisions, such as lending, assessing job applications, 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) (Guidotti et al., 2018; Mittelstadt et al., 2019; Molnar, 2020) in which the goal is to generate *effective explanations* that help build trust by providing a mechanism for *normative evaluation* of an algorithmic system, ensuring different stakeholders that the system's decision rules are justifiable (Selbst & Barocas, 2018); and provide users with an *actionable recourse* to change the results of algorithms in the future (Berk, 2019; Wachter et al., 2017; Venkatasubramanian & Alfano, 2020; Karimi et al., 2020a).

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 (Friedman, 2001; Goldstein et al., 2015; Apley & Zhu, 2016; Hooker, 2004; Greenwell et al., 2018; Fisher et al., 2018; Lundberg & Lee, 2017; Lundberg et al., 2018; Datta et al., 2016); 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 (Ribeiro et al., 2016; 2018). 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 (Hooker & Mentch, 2019; Kumar et al., 2020; Frye et al., 2019; Guidotti et al., 2018; Molnar, 2020; Alvarez-Melis & Jaakkola, 2018). Furthermore, several recent works have argued for the use of *contrastive explanations* (a.k.a counterfactual explanations), sometimes called counterfactual explanation which are typically obtained by considering the smallest perturbation in an algorithm's input that can lead to the algorithm's desired outcome (Wachter et al., 2017; Laugel et al., 2017; Ustun et al., 2019; Mahajan et al., 2019; Mothilal et al., 2020b). 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 (Karimi et al., 2020a).

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 their limitations. Our system framework 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)

[*]Equal contribution [1]University of Chicago [2]University of California, San Diego. Correspondence to: Sainyam Galhotra <sainyam@uchicago.edu>, Romila Pradhan <rpradhan@ucsd.edu>, Babak Salimi <bsalimi@ucsd.edu>.

ICML (International Conference on Machine Learning) Workshop on Algorithmic Recourse.

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 (De Graaf & Malle, 2017; Gerstenberg et al., 2015; Pearl, 2009; Lipton, 1990; Woodward, 2005; Grynaviski, 2013; Morton, 2013). Their probabilistic interpretation has been formalized and studied extensively in AI, biostatistics, political science, epistemology, biology and legal reasoning (Greenland, 1999; Robins & Greenland, 1989; Greenland & Robins, 1999; Tian & Pearl, 2000; Greenland, 1999; Robertson, 1996; Cox Jr, 1984; Pearl, 2009; Grynaviski, 2013; Mandel, 2005). While their importance in achieving the objectives of XAI has been recognized in the literature (Miller, 2019), very few attempts have been made to operationalize causality-based contrastive counterfactuals for XAI.

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. (2) Showing that the problem of generating *actionable recourse* can be framed as an optimization problem that searches for a *minimal intervention* on a prespecified set of actionable variables that have a high *sufficiency score* for producing the algorithm's desired future outcome. (3) Establishing conditions under which the class of probabilistic contrastive counterfactuals we use can be bounded and estimated using observational data (Sec. 3).

2. Preliminaries

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 $\mathbf{x}$). Denote $Dom(X)$ the domain of a variable X. 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 (Pearl, 2009), 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 *endogenous* variables $\mathbf{V}$ and a set of *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 $F_X : Dom(\mathbf{Pa_V}(X)) \times Dom(\mathbf{Pa_U}(X)) \to Dom(X)$, 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.

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

$$\Pr(y_{\mathbf{X}\leftarrow\mathbf{x}} \mid \mathbf{k}) = \sum_{\mathbf{u}} \Pr(y_{\mathbf{X}\leftarrow\mathbf{x}}(\mathbf{u}))\ \Pr(\mathbf{u} \mid \mathbf{k}) \qquad (2)$$

Equation (2) readily suggests Pearl's three-step procedure for answering counterfactual queries (Pearl, 2009), which 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. Indeed, counterfactual queries are not identifiable in general if the underlying causal models is not fully specified.

For causal diagrams, Pearl defined the $\mathtt{do}$-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 $\Pr(\mathbf{y} \mid \mathtt{do}(\mathbf{x}))$. A sufficient condition for identification of intervetional queries 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 (Pearl, 2009):

$$\Pr(\mathbf{y} \mid \mathtt{do}(\mathbf{x})) = \sum_{\mathbf{c}\in Dom(\mathbf{C})} \Pr(\mathbf{y} \mid \mathbf{c}, \mathbf{x})\ \Pr(\mathbf{c}) \qquad (3)$$

In contrast to (2), the RHS of (3) is in terms of observed probabilities and can be estimated from observational data.

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. We then use these measures to generate recourse for an individual that received a negative outcome.

Explanation Scores. We are given a decision-making algorithm $f : Dom(\mathbf{I}) \to Dom(O)$, where $\mathbf{I}$ is set of input

features and O is a binary outcome, where $O = o$ denotes the positive decision and $O = o'$ denotes the negative decision. 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 (we implicitly assume an order $x > x'$).

Definition 1 (Explanation Scores)**.** *Given a PCM $\langle M, \Pr(\mathbf{u})\rangle$ and an algorithm $f : Dom(\mathbf{X}) \to Dom(O)$, 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*:*

$$\textsc{Nec}_x^{x'}(\mathbf{k}) \stackrel{def}{=} \Pr(o'_{X \leftarrow x'} \mid x, o, \mathbf{k}) \tag{4}$$

- *The* sufficiency score*:*

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

- *The* necessity and sufficiency score*:*

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

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

For simplicity of notation, we drop x' from $\textsc{Nec}_x^{x'}$, $\textsc{Suf}_x^{x'}$ and $\textsc{NeSuf}_x^{x'}$ whenever it is clear from the context. The sufficiency score in (5) 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'?" The necessity score is the dual of sufficiency score and the necessity and sufficiency score establishes a balance between necessary and sufficiency; it measures the probability that the algorithm responds in both ways.

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 explanation scores and counterfactual recourse, can be computed via Equation (2). However, in many applications, PCMs are not fully observed, and one must estimate explanation scores from data. First, we prove the following on explanation scores.

Proposition 3.1. Given a causal diagram G, if the decision-making algorithm $f : Dom(\mathbf{I}) \to Dom(O)$ is monotone relative to $\mathbf{x}, \mathbf{x}' \in Dom(\mathbf{X})$ and if there exist 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:

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

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

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

Proposition 3.1 facilitates 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. We further present general bounds to estimate these scores in the full version (Galhotra et al., 2021).

Counterfactual Recourse. For individuals for whom an algorithm's decision is negative, our framework generates explanations in terms of minimal interventions on a user-specified set of actionable variables that have a high *sufficiency score*, i.e., the intervention can produce a 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 to improve the outcome in the future.

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

$$\underset{\mathbf{a} \in Dom(\mathbf{A})}{\operatorname{argmin}} \ \textsc{Cost}(\mathbf{a}, \hat{\mathbf{a}}) \qquad \text{s.t. } \textsc{Suf}_{\hat{\mathbf{a}}}(\mathbf{v}) \geq \alpha \tag{7}$$

This optimization problem treats the decision-making algorithm as a black box. The solutions to this problem provide end-users with informative, feasible and actionable recourse by answering questions such as "What are the best courses of action that, if performed in the real world, would change the outcome for this individual with high probability?" We formulate the problem as a combinatorial optimization problem over the domain of actionable variables and express it as an integer programming (IP) problem of the form:

$$\underset{\hat{\mathbf{a}} \in Dom(\mathbf{A})}{\operatorname{argmin}} \ \sum_{A \in \mathbf{A}} \left(\phi_A \sum_{a \in Dom(A)} \delta_a \right) \tag{8}$$

$$\text{subject to} \qquad \textsc{Suf}_{\hat{\mathbf{a}}}(\mathbf{v}) \geq \alpha \tag{9}$$

$$\sum_{a \in Dom(A)} \delta_a \leq 1, \quad \forall A \in \mathbf{A} \tag{10}$$

$$\delta_a \in \{0, 1\}, \quad \forall a \in Dom(A), A \in \mathbf{A} \tag{11}$$

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 (9) 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 (10) 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. To compute the sufficiency score in (9) 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). Then, we can incorporate the *lower bound* obtained for the sufficiency score in the optimization problem, as follows:

$$\Pr(o \mid \hat{\mathbf{a}}, \mathbf{k}) \geq \Pr(o \mid \mathbf{a}, \mathbf{k}) + \alpha \Pr(o' \mid \mathbf{a}, \mathbf{k}) \qquad (12)$$

Since $\mathbf{k}, \mathbf{a}, \alpha$ are constant, the RHS of (12) 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 (12) as a linear inequality constraint for the IP in (8). *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).

4. Experiments

In this section we evaluate the recourse generated by our framework on the following datasets. We trained random forest classifiers and used them as black-box algorithms.

German Credit Data (Dua & Graff, 2017). This dataset consists of records of bank account holders with their personal, financial and demographic information. The prediction task classifies individuals as good/bad credit risks.

Adult Income Data (Dua & Graff, 2017) contains demographic information of individuals along with information on their occupation, working hours, etc. The task is to predict if the annual income of an individual exceeds 50K.

German-syn. We generate synthetic data following the causal graph similar to the `German` dataset with a difference that savings is assumed to affect credit history and other actionable attributes.

4.1. Solution Quality

Real data. We sampled individuals that achieved negative outcome and generated recourse with sufficiency threshold of 0.90. One of the examples consisted of an adult female applicant with low credit amount and savings but high credit history. In this case, our approach returned that credit history and savings need to be increased to achieve a positive outcome. Another example consisted of a senior male individual who had low savings and low credit history. In this case, our framework recommended to improve the credit history as the smallest change in actionable attributes to achieve positive outcome. We indeed performed this intervention and validated the correctness of the suggested recourse. We observed similar results for adult dataset. As an example, our framework suggested that increasing the hours to more than 42 would result in a high-income prediction for some of the individuals.

Synthetic Data – Correctness. We sampled 1000 random instances that received negative outcomes and generated recourse using our approach. Each unit change in attribute value was assigned unit cost. The output was evaluated with respect to the ground truth sufficiency and cost of returned actions. The lowest cost recourse suggested by our approach comprised of credit history as it further affected other features. On the other hand, a prior non-causal technique (Ustun et al., 2019) assumed independence between features and suggested to improve multiple features like savings, housing and credit limit to change the decision outcome. This experiment validates the optimality of the IP formulation in generating low cost and effective recourse that complies with the underlying causal effects.

To further test the scalability of our techniques, 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 it's scalability to larger inputs.

5. Related Work

Our work is related to a line of research that leverages counterfactuals to explain ML algorithm predictions. 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 (Ustun et al., 2019; Mothilal et al., 2020b; Joshi et al., 2019; Dhurandhar et al., 2018; Van Looveren & Klaise, 2019; Liu et al., 2019). However, it has been argued that feasibility is fundamentally a causal concept (Barocas et al., 2020; Mahajan et al., 2019; Karimi et al., 2019). 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 (Mahajan et al., 2019; Karimi et al., 2019; 2020b). Our framework extends this line of work by (1) formally defining feasibility in terms of probabilistic contrastive counterfactuals, and (2) 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. (2020a); Watson (1989) proposed a notion of necessity and sufficiency scores for quantifying feature importance but do not study recourse.

References

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Apley, D. W. and Zhu, J. Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*, 2016.

Barocas, S., Selbst, A. D., and Raghavan, M. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 80–89, 2020.

Berk, R. *Machine Learning Risk Assessments in Criminal Justice Settings*. Springer, 2019. ISBN 978-3-030-02271-6. doi: 10.1007/978-3-030-02272-3. URL https://doi.org/10.1007/978-3-030-02272-3.

Cox Jr, L. A. Probability of causation and the attributable proportion risk. *Risk Analysis*, 4(3):221–230, 1984.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617. IEEE, 2016.

De Graaf, M. M. and Malle, B. F. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pp. 592–603, 2018.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Fisher, A., Rudin, C., and Dominici, F. Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective. *arXiv preprint arXiv:1801.01489*, 68, 2018.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Frye, C., Feige, I., and Rowat, C. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

Galhotra, S., Pradhan, R., and Salimi, B. Explaining black-box algorithms using probabilistic contrastive counterfactuals. *arXiv preprint arXiv:2103.11972*, 2021.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., and Tenenbaum, J. B. How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*, 2015.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. 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.

Greenland, S. 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.

Greenland, S. and Robins, J. M. Epidemiology, justice, and the probability of causation. *Jurimetrics*, 40:321, 1999.

Greenwell, B. M., Boehmke, B. C., and McCarthy, A. J. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

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

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

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

Hooker, G. and Mentch, L. Please stop permuting features: An explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Karimi, A.-H., Barthe, G., Belle, B., and Valera, I. Model-agnostic counterfactual explanations for consequential decisions. *arXiv preprint arXiv:1905.11190*, 2019.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020a.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *arXiv preprint arXiv:2006.06831*, 2020b.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., and Detyniecki, M. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Lipton, P. Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27:247–266, 1990.

Liu, S., Kailkhura, B., Loveland, D., and Han, Y. Generative counterfactual introspection for explainable deep learning. *arXiv preprint arXiv:1907.03077*, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Mahajan, D., Tan, C., and Sharma, A. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Mandel, D. R. Counterfactual and causal explanation. *Routledge research international series in social psychology. The Psychology of Counterfactual Thinking*, pp. 11–27, 2005.

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

Mittelstadt, B., Russell, C., and Wachter, S. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

Molnar, C. *Interpretable Machine Learning*. Lulu. com, 2020.

Morton, A. Contrastive knowledge. *Contrastivism in philosophy*, pp. 101–115, 2013.

Mothilal, R. K., Mahajan, D., Tan, C., and Sharma, A. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv preprint arXiv:2011.04917*, 2020a.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020b.

Pearl, J. *Causality*. Cambridge university press, 2009.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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*, pp. 1135–1144, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pp. 1527–1535, 2018.

Robertson, D. W. Common sense of cause in fact. *Tex. L. Rev.*, 75:1765, 1996.

Robins, J. and Greenland, S. The probability of causation under a stochastic model for individual risk. *Biometrics*, pp. 1125–1138, 1989.

Selbst, A. D. and Barocas, S. The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87:1085, 2018.

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

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

Venkatasubramanian, S. and Alfano, M. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 284–293, 2020.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Watson, S. D. Reinvigorating title vi: Defending health care discrimination–it shouldn't be so easy. *Fordham L. Rev.*, 58:939, 1989.

Woodward, J. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.