Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice

David S. Watson[1] **Limor Gultchin**[2,3] **Ankur Taly**[4] **Luciano Floridi**[5,3]

[1]Department of Statistical Science, University College London, London, UK
[2]Department of Computer Science, University of Oxford, Oxford, UK
[3]The Alan Turing Institute, London, UK
[4]Google Inc., Mountain View, USA
[5]Oxford Internet Institute, University of Oxford, Oxford, UK

Abstract

Necessity and sufficiency are the building blocks of all successful explanations. Yet despite their importance, these notions have been conceptually underdeveloped and inconsistently applied in explainable artificial intelligence (XAI), a fast-growing research area that is so far lacking in firm theoretical foundations. Building on work in logic, probability, and causality, we establish the central role of necessity and sufficiency in XAI, unifying seemingly disparate methods in a single formal framework. We provide a sound and complete algorithm for computing explanatory factors with respect to a given context, and demonstrate its flexibility and competitive performance against state of the art alternatives on various tasks.

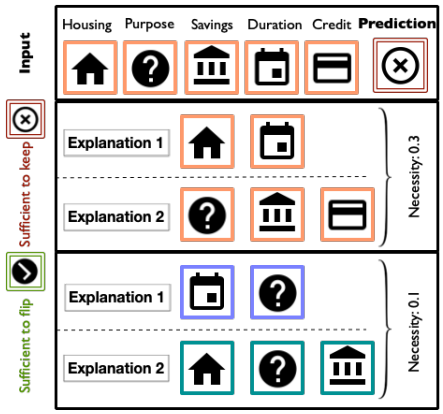

Figure 1: We describe minimal sufficient factors (here, sets of features) for a given input (top row), with the aim of preserving or flipping the original prediction. We report a sufficiency score for each set and a cumulative necessity score for all sets, indicating the proportion of paths towards the outcome that are covered by the explanation. Feature colors indicate source of feature values (input or reference).

1 INTRODUCTION

Machine learning algorithms are increasingly used in a variety of high-stakes domains, from credit scoring to medical diagnosis. However, many such methods are *opaque*, in that humans cannot understand the reasoning behind particular predictions. Post-hoc, model-agnostic local explanation tools (e.g., feature attributions, rule lists, and counterfactuals) are at the forefront of a fast-growing area of research variously referred to as *interpretable machine learning* or *explainable artificial intelligence* (XAI).

Many authors have pointed out the inconsistencies between popular XAI tools, raising questions as to which method is more reliable in particular cases [Mothilal et al., 2020a; Ramon et al., 2020; Fernández-Loría et al., 2020]. Theoretical foundations have proven elusive in this area, perhaps due to the perceived subjectivity inherent to notions such as "intelligible" and "relevant" [Watson and Floridi, 2020]. Practitioners often seek refuge in the axiomatic guarantees of Shapley values, which have become the de facto standard in many XAI applications, due in no small part to their attractive theoretical properties [Bhatt et al., 2020]. However, ambiguities regarding the underlying assumptions of the method [Kumar et al., 2020] and the recent proliferation of mutually incompatible implementations [Sundararajan and Najmi, 2019; Merrick and Taly, 2020] have complicated this picture. Despite the abundance of alternative XAI tools [Molnar, 2021], a dearth of theory persists. This has led some to conclude that the goals of XAI are underspecified [Lipton, 2018], and even that post-hoc methods do more harm than good [Rudin, 2019].

We argue that this lacuna at the heart of XAI should be filled by a return to fundamentals – specifically, to *necessity* and *sufficiency*. As the building blocks of all successful explanations, these dual concepts deserve a privileged position

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1382–1392.

in the theory and practice of XAI. Following a review of related work (Sect. 2), we operationalize this insight with a unified framework (Sect. 3) that reveals unexpected affinities between various XAI tools and probabilities of causation (Sect. 4). We proceed to implement a novel procedure for computing model explanations that improves upon the state of the art in various quantitative and qualitative comparisons (Sect. 5). Following a brief discussion (Sect. 6), we conclude with a summary and directions for future work (Sect. 7).

We make three main contributions. (1) We present a formal framework for XAI that unifies several popular approaches, including feature attributions, rule lists, and counterfactuals. (2) We introduce novel measures of necessity and sufficiency that can be computed for any feature subset. The method enables users to incorporate domain knowledge, search various subspaces, and select a utility-maximizing explanation. (3) We present a sound and complete algorithm for identifying explanatory factors, and illustrate its performance on a range of tasks.

2 NECESSITY AND SUFFICIENCY

Necessity and sufficiency have a long philosophical tradition [Mackie, 1965; Lewis, 1973; Halpern and Pearl, 2005], spanning logical, probabilistic, and causal variants. In propositional logic, we say that x is a sufficient condition for y iff $x \rightarrow y$, and x is a necessary condition for y iff $y \rightarrow x$. So stated, necessity and sufficiency are logically *converse*. However, by the law of contraposition, both definitions admit alternative formulations, whereby sufficiency may be rewritten as $\neg y \rightarrow \neg x$ and necessity as $\neg x \rightarrow \neg y$. By pairing the original definition of sufficiency with the latter definition of necessity (and vice versa), we find that the two concepts are also logically *inverse*. These formulae suggest probabilistic relaxations, measuring x's sufficiency for y by $P(y|x)$ and x's necessity for y by $P(x|y)$. Because there is no probabilistic law of contraposition, these quantities are generally uninformative w.r.t. $P(\neg x|\neg y)$ and $P(\neg y|\neg x)$, which may be of independent interest. Thus, while necessity is both the converse and inverse of sufficiency in propositional logic, the two formulations come apart in probability calculus. We revisit the distinction between probabilistic conversion and inversion in Rmk. 1 and Sect. 4.

These definitions struggle to track our intuitions when we consider causal explanations [Pearl, 2000; Tian and Pearl, 2000]. It may make sense to say in logic that if x is a necessary condition for y, then y is a sufficient condition for x; it does not follow that if x is a necessary *cause* of y, then y is a sufficient *cause* of x. We may amend both concepts using *counterfactual probabilities* – e.g., the probability that Alice would still have a headache if she had not taken an aspirin, given that she does not have a headache and did take an aspirin. Let $P(y_x|x', y')$ denote such a quantity, to be read as "the probability that Y would equal y under an intervention that sets X to x, given that we observe $X = x'$ and $Y = y'$." Then, according to Pearl [2000, Ch. 9], the probability that x is a sufficient cause of y is given by $\texttt{suf}(x, y) := P(y_x|x', y')$, and the probability that x is a necessary cause of y is given by $\texttt{nec}(x, y) := P(y'_{x'}|x, y)$.

Analysis becomes more difficult in higher dimensions, where variables may interact to block or unblock causal pathways. VanderWeele and Robins [2008] analyze sufficient causal interactions in the potential outcomes framework, refining notions of synergism without monotonicity constraints. In a subsequent paper, VanderWeele and Richardson [2012] study the irreducibility and singularity of interactions in sufficient-component cause models. Halpern [2016] devotes an entire monograph to the subject, providing various criteria to distinguish between subtly different notions of "actual causality", as well as "but-for" (similar to necessary) and sufficient causes. These authors generally limit their analyses to Boolean systems with convenient structural properties, e.g. conditional ignorability and the stable unit treatment value assumption. Operationalizing their theories in a practical method without such restrictions is one of our primary contributions.

Necessity and sufficiency have begun to receive explicit attention in the XAI literature. Ribeiro et al. [2018a] propose a bandit procedure for identifying a minimal set of Boolean conditions that entails a predictive outcome (more on this in Sect. 4). Dhurandhar et al. [2018] propose an autoencoder for learning pertinent negatives and positives, i.e. features whose presence or absence is decisive for a given label, while Zhang et al. [2018] develop a technique for generating symbolic corrections to alter model outputs. Both methods are optimized for neural networks, unlike the model-agnostic approach we develop here.

Another strand of research in this area is rooted in logic programming. Several authors have sought to reframe XAI as either a SAT [Ignatiev et al., 2019; Narodytska et al., 2019] or a set cover problem [Lakkaraju et al., 2019; Grover et al., 2019], typically deriving approximate solutions on a prespecified subspace to ensure computability in polynomial time. We adopt a different strategy that prioritizes completeness over efficiency, an approach we show to be feasible in moderate dimensions (see Sect. 6 for a discussion). Mothilal et al. [2020a] build on Halpern [2016]'s definitions of necessity and sufficiency to critique popular XAI tools, proposing a new feature attribution measure with some purported advantages. Their method relies on the strong assumption that predictors are mutually independent. Galhotra et al. [2021] adapt Pearl [2000]'s probabilities of causation for XAI under a more inclusive range of data generating processes. They derive analytic bounds on multidimensional extensions of `nec` and `suf`, as well as an algorithm for point identification when graphical structure permits. Oddly, they claim that non-causal applications of necessity and sufficiency are somehow "incorrect and misleading" (p. 2), a norma-

tive judgment that is inconsistent with many common uses of these concepts. Rather than insisting on any particular interpretation of necessity and sufficiency, we propose a general framework that admits logical, probabilistic, and causal interpretations as special cases. Whereas previous works evaluate individual predictors, we focus on feature *subsets*, allowing us to detect and quantify interaction effects. Our formal results clarify the relationship between existing XAI methods and probabilities of causation, while our empirical results demonstrate their applicability to a wide array of tasks and datasets.

3 A UNIFYING FRAMEWORK

We propose a unifying framework that highlights the role of necessity and sufficiency in XAI. Its constituent elements are described below.

Target function. Post-hoc explainability methods assume access to a target function $f : \mathcal{X} \mapsto \mathcal{Y}$, i.e. the model whose prediction(s) we seek to explain. For simplicity, we restrict attention to the binary setting, with $Y \in \{0, 1\}$. Multi-class extensions are straightforward, while continuous outcomes may be accommodated via discretization. Though this inevitably involves some information loss, we follow authors in the contrastivist tradition in arguing that, even for continuous outcomes, explanations always involve a juxtaposition (perhaps implicit) of "fact and foil" [Lipton, 1990]. For instance, a loan applicant is probably less interested in knowing why her credit score is precisely y than she is in discovering why it is below some threshold (say, 700). Of course, binary outcomes can approximate continuous values with arbitrary precision over repeated trials.

Context. The context $\mathcal{D}$ is a probability distribution over which we quantify sufficiency and necessity. Contexts may be constructed in various ways but always consist of at least some input (point or space) and reference (point or space). For instance, we may want to compare $\boldsymbol{x}_i$ with all other samples, or else just those perturbed along one or two axes, perhaps based on some conditioning event(s).

In addition to predictors and outcomes, we optionally include information exogenous to f. For instance, if any events were conditioned upon to generate a given reference sample, this information may be recorded among a set of auxiliary variables $\boldsymbol{W}$. Other examples of potential auxiliaries include metadata or engineered features such as those learned via neural embeddings. This augmentation allows us to evaluate the necessity and sufficiency of factors beyond those found in $\boldsymbol{X}$. Contextual data take the form $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{W}) \sim \mathcal{D}$. The distribution may or may not encode dependencies between (elements of) $\boldsymbol{X}$ and (elements of) $\boldsymbol{W}$. We extend the target function to augmented inputs by defining $f(\boldsymbol{z}) := f(\boldsymbol{x})$.

Factors. Factors pick out the properties whose necessity and sufficiency we wish to quantify. Formally, a factor $c : \mathcal{Z} \mapsto \{0, 1\}$ indicates whether its argument satisfies some criteria with respect to predictors or auxiliaries. For instance, if $\boldsymbol{x}$ is an input to a credit lending model, and $\boldsymbol{w}$ contains information about the subspace from which data were sampled, then a factor could be $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} = \text{"female"}] \wedge \boldsymbol{w}[do(\mathsf{income} > \$50\text{k})]]$, i.e. checking if $\boldsymbol{z}$ is female and drawn from a context in which an intervention fixes income at greater than $50k. We use the term "factor" as opposed to "condition" or "cause" to suggest an inclusive set of criteria that may apply to predictors $\boldsymbol{x}$ and/or auxiliaries $\boldsymbol{w}$. Such criteria are always observational w.r.t. $\boldsymbol{z}$ but may be interventional or counterfactual w.r.t. $\boldsymbol{x}$. We assume a finite space of factors $\mathcal{C}$.

Partial order. When multiple factors pass a given necessity or sufficiency threshold, users will tend to prefer some over others. For instance, factors with fewer conditions are often preferable to those with more, all else being equal; factors that change a variable by one unit as opposed to two are preferable, and so on. Rather than formalize this preference in terms of a distance metric, which unnecessarily constrains the solution space, we treat the partial ordering as primitive and require only that it be complete and transitive. This covers not just distance-based measures but also more idiosyncratic orderings that are unique to individual agents. Ordinal preferences may be represented by cardinal utility functions under reasonable assumptions (see, e.g., [von Neumann and Morgenstern, 1944]).

We are now ready to formally specify our framework.

Definition 1 (Basis)**.** A *basis* for computing necessary and sufficient factors for model predictions is a tuple $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, where f is a target function, $\mathcal{D}$ is a context, $\mathcal{C}$ is a set of factors, and $\preceq$ is a partial ordering on $\mathcal{C}$.

3.1 EXPLANATORY MEASURES

For some fixed basis $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle$, we define the following measures of sufficiency and necessity, with probability taken over $\mathcal{D}$.

Definition 2 (Probability of Sufficiency)**.** The probability that c is a sufficient factor for outcome y is given by:

$$PS(c, y) := P(f(\boldsymbol{z}) = y \mid c(\boldsymbol{z}) = 1).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is sufficient for y is given by:

$$PS(C, y) := P(f(\boldsymbol{z}) = y \mid \sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1).$$

Definition 3 (Probability of Necessity)**.** The probability that c is a necessary factor for outcome y is given by:

$$PN(c, y) := P(c(\boldsymbol{z}) = 1 \mid f(\boldsymbol{z}) = y).$$

The probability that factor set $C = \{c_1, \ldots, c_k\}$ is necessary for y is given by:

$$PN(C, y) := P(\sum_{i=1}^{k} c_i(\boldsymbol{z}) \geq 1 \mid f(\boldsymbol{z}) = y).$$

Remark 1. These probabilities can be likened to the "precision" (positive predictive value) and "recall" (true positive rate) of a (hypothetical) classifier that predicts whether $f(\boldsymbol{z}) = y$ based on whether $c(\boldsymbol{z}) = 1$. By examining the confusion matrix of this classifier, one can define other related quantities, e.g. the true negative rate $P(c(\boldsymbol{z}) = 0 | f(\boldsymbol{z}) \neq y)$ and the negative predictive value $P(f(\boldsymbol{z}) \neq y | c(\boldsymbol{z}) = 0)$, which are contrapositive transformations of our proposed measures. We can recover these values exactly via $PS(1 - c, 1 - y)$ and $PN(1 - c, 1 - y)$, respectively. When necessity and sufficiency are defined as probabilistic inversions (rather than conversions), such transformations are impossible.

3.2 MINIMAL SUFFICIENT FACTORS

We introduce Local Explanations via Necessity and Sufficiency (LENS), a procedure for computing explanatory factors with respect to a given basis $\mathcal{B}$ and threshold parameter τ (see Alg. 1). First, we calculate a factor's probability of sufficiency (see probSuff) by drawing n samples from $\mathcal{D}$ and taking the maximum likelihood estimate $\hat{PS}(c, y)$. Next, we sort the space of factors w.r.t. $\preceq$ in search of those that are τ-minimal.

Definition 4 (τ-minimality)**.** We say that c is τ-minimal iff (i) $PS(c, y) \geq \tau$ and (ii) there exists no factor c' such that $PS(c', y) \geq \tau$ and $c' \prec c$.

Since a factor is necessary to the extent that it covers all possible pathways towards a given outcome, our next step is to span the τ-minimal factors and compute their cumulative PN (see probNec). As a minimal factor c stands for all c' such that $c \preceq c'$, in reporting probability of necessity, we expand C to its upward closure.

Thms. 1 and 2 state that this procedure is *optimal* in a sense that depends on whether we assume access to oracle or sample estimates of PS (see Appendix A for all proofs).

Theorem 1. With oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is sound and complete. That is, for any C returned by Alg. 1 and all $c \in \mathcal{C}$, c is τ-minimal iff $c \in C$.

Population proportions may be obtained if data fully saturate the space $\mathcal{D}$, a plausible prospect for categorical variables of low to moderate dimensionality. Otherwise, proportions will need to be estimated.

Theorem 2. With sample estimates $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, Alg. 1 is uniformly most powerful. That is, Alg. 1 identifies the most τ-minimal factors of any method with fixed type I error α.

Multiple testing adjustments can easily be accommodated, in which case modified optimality criteria apply [Storey, 2007].

Remark 2. We take it that the main quantity of interest in most applications is sufficiency, be it for the original or alternative outcome, and therefore define τ-minimality w.r.t. sufficient (rather than necessary) factors. However, necessity serves an important role in tuning τ, as there is an inherent trade-off between the parameters. More factors are excluded at higher values of τ, thereby inducing lower cumulative PN; more factors are included at lower values of τ, thereby inducing higher cumulative PN. See Appendix B.

4 ENCODING EXISTING MEASURES

Explanatory measures can be shown to play a central role in many seemingly unrelated XAI tools, albeit under different assumptions about the basis tuple $\mathcal{B}$. In this section, we relate our framework to a number of existing methods.

Feature attributions. Several popular feature attribution algorithms are based on Shapley values [Shapley, 1953], which decompose the predictions of any target function as a sum of weights over d input features:

$$f(\boldsymbol{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_j, \tag{1}$$

where ϕ_0 represents a baseline expectation and ϕ_j the weight assigned to X_j at point $\boldsymbol{x}_i$. Let $v : 2^d \mapsto \mathbb{R}$ be a value function such that $v(S)$ is the payoff associated with feature subset $S \subseteq [d]$ and $v(\{\emptyset\}) = 0$. Define the complement $R = [d] \backslash S$ such that we may rewrite any $\boldsymbol{x}_i$ as a pair of subvectors, $(\boldsymbol{x}_i^S, \boldsymbol{x}_i^R)$. Payoffs are given by:

$$v(S) = \mathbb{E}[f(\boldsymbol{x}_i^S, \boldsymbol{X}^R)], \tag{2}$$

although this introduces some ambiguity regarding the reference distribution for $\boldsymbol{X}^R$ (more on this below). The Shapley value ϕ_j is then j's average marginal contribution to all subsets that exclude it:

$$\phi_j = \sum_{S \subseteq [d] \backslash \{j\}} \frac{|S|!(d - |S| - 1)!}{d!} v(S \cup \{j\}) - v(S). \tag{3}$$

It can be shown that this is the unique solution to the attribution problem that satisfies certain desirable properties, including efficiency, linearity, sensitivity, and symmetry.

Algorithm 1 LENS

Input: $\mathcal{B} = \langle f, \mathcal{D}, \mathcal{C}, \preceq \rangle, \tau$
Output: Factor set C, $(\forall c \in C)\ PS(c, y), PN(C, y)$

Sample $\hat{D} = \{\boldsymbol{z}_i\}_{i=1}^{n} \sim \mathcal{D}$

function probSuff(c, y)
    n(c&y) = $\sum_{i=1}^{n} \mathbb{1}[c(\boldsymbol{z}_i) = 1 \wedge f(\boldsymbol{z}_i) = y]$
    n(c) = $\sum_{i=1}^{n} c(\boldsymbol{z}_i)$
    return n(c&y) / n(c)

function probNec(C, y, upward_closure_flag)
    if upward_closure_flag **then**
        $C = \{c \mid c \in \mathcal{C} \wedge \exists\ c' \in C : c' \preceq c\}$
    end if
    n(C&y) = $\sum_{i=1}^{n} \mathbb{1}[\sum_{j=1}^{k} c_j(\boldsymbol{z}_i) \geq 1 \wedge f(\boldsymbol{z}_i) = y]$
    n(y) = $\sum_{i=1}^{n} \mathbb{1}[f(\boldsymbol{z}_i) = y]$
    return n(C&y) / n(y)

function minimalSuffFactors(y, τ, sample_flag, α)
    sorted_factors = topological_sort($\mathcal{C}, \preceq$)
    cands = []
    for c in sorted_factors **do**
        if $\exists(c', \_) \in$ cands $: c' \preceq c$ **then**
            continue
        end if
        ps = probSuff(c, y)
        if sample_flag **then**
            p = binom.test(n(c&y), n(c), τ, alt = $>$)
            if p $\leq \alpha$ **then**
                cands.append(c, ps)
            end if
        else if ps $\geq \tau$ **then**
            cands.append(c, ps)
        end if
    end for
    cum_pn = probNec($\{c \mid (c, \_) \in$ cands$\}, y$, TRUE)
    return cands, cum_pn

Reformulating this in our framework, we find that the value function v is a sufficiency measure. To see this, let each $\boldsymbol{z} \sim \mathcal{D}$ be a sample in which a random subset of variables S are held at their original values, while remaining features R are drawn from a fixed distribution $\mathcal{D}(\cdot|S)$.[1]

Proposition 1. Let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x} \subseteq \boldsymbol{z}$ was constructed by holding $\boldsymbol{x}^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\cdot|S)$. Then $v(S) = PS(c_S, y)$.

Thus, the Shapley value ϕ_j measures X_j's average marginal increase to the sufficiency of a random feature subset. The advantage of our method is that, by focusing on particular subsets instead of weighting them all equally, we disregard irrelevant permutations and home in on just those that meet a τ-minimality criterion. Kumar et al. [2020] observe that, "since there is no standard procedure for converting Shapley values into a statement about a model's behavior, developers rely on their own mental model of what the values represent" (p. 8). By contrast, necessary and sufficient factors are more transparent and informative, offering a direct path to what Shapley values indirectly summarize.

Rule lists. Rule lists are sequences of if-then statements that describe a hyperrectangle in feature space, creating partitions that can be visualized as decision or regression trees. Rule lists have long been popular in XAI. While early work in this area tended to focus on global methods, more recent efforts have prioritized local explanation tasks.

We focus in particular on the Anchors algorithm [Ribeiro et al., 2018a], which learns a set of Boolean conditions A (the eponymous "anchors") such that $A(\boldsymbol{x}_i) = 1$ and

$$P_{\mathcal{D}(\boldsymbol{x}|A)}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})) \geq \tau. \tag{4}$$

The lhs of Eq. 4 is termed the *precision*, prec(A), and probability is taken over a synthetic distribution in which the conditions in A hold while other features are perturbed. Once τ is fixed, the goal is to maximize *coverage*, formally defined as $\mathbb{E}[A(\boldsymbol{x}) = 1]$, i.e. the proportion of datapoints to which the anchor applies.

The formal similarities between Eq. 4 and Def. 2 are immediately apparent, and the authors themselves acknowledge that Anchors are intended to provide "sufficient conditions" for model predictions.

Proposition 2. Let $c_A(\boldsymbol{z}) = 1$ iff $A(\boldsymbol{x}) = 1$. Then prec$(A) = PS(c_A, y)$.

While Anchors outputs just a single explanation, our method generates a ranked list of candidates, thereby offering a more comprehensive view of model behavior. Moreover, our necessity measure adds a mode of explanatory information entirely lacking in Anchors.

Counterfactuals. Counterfactual explanations identify one or several nearest neighbors with different outcomes, e.g. all datapoints $\boldsymbol{x}$ within an ϵ-ball of $\boldsymbol{x}_i$ such that labels $f(\boldsymbol{x})$ and $f(\boldsymbol{x}_i)$ differ (for classification) or $f(\boldsymbol{x}) > f(\boldsymbol{x}_i) + \delta$ (for regression).[2] The optimization problem is:

[1] The diversity of Shapley value algorithms is largely due to variation in how this distribution is defined. Popular choices include the marginal $P(\boldsymbol{X}^R)$ [Lundberg and Lee, 2017]; conditional $P(\boldsymbol{X}^R|\boldsymbol{x}^S)$ [Aas et al., 2021]; and interventional $P(\boldsymbol{X}^R|do(\boldsymbol{x}^S))$ [Heskes et al., 2020] distributions.

[2] Confusingly, the term "counterfactual" in XAI refers to any point with an alternative outcome, which is distinct from the causal sense of the term (see Sect. 2). We use the word in both senses here, but strive to make our intended meaning explicit in each case.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \text{CF}(\boldsymbol{x}_i)}{\text{argmin}} \; cost(\boldsymbol{x}_i, \boldsymbol{x}), \tag{5}$$

where $\text{CF}(\boldsymbol{x}_i)$ denotes a counterfactual space such that $f(\boldsymbol{x}_i) \neq f(\boldsymbol{x})$ and $cost$ is a user-supplied cost function, typically equated with some distance measure. [Wachter et al., 2018] recommend using generative adversarial networks to solve Eq. 5, while others have proposed alternatives designed to ensure that counterfactuals are coherent and actionable [Ustun et al., 2019; Karimi et al., 2020a]. As with Shapley values, the variation in these proposals is reducible to the choice of context $\mathcal{D}$.

For counterfactuals, we rewrite the objective as a search for minimal perturbations sufficient to flip an outcome.

Proposition 3. Let $cost$ be a function representing $\preceq$, and let c be some factor spanning reference values. Then the counterfactual recourse objective is:

$$c^* = \underset{c \in \mathcal{C}}{\text{argmin}} \; cost(c) \;\; \text{s.t. } PS(c, 1-y) \geq \tau, \tag{6}$$

where τ denotes a decision threshold. Counterfactual outputs will then be any $\boldsymbol{z} \sim \mathcal{D}$ such that $c^*(\boldsymbol{z}) = 1$.

Probabilities of causation. Our framework can describe Pearl [2000]'s aforementioned probabilities of causation, however in this case $\mathcal{D}$ must be constructed with care.

Proposition 4. Consider the bivariate Boolean setting, as in Sect. 2. We have two counterfactual distributions: an input space $\mathcal{I}$, in which we observe x, y but intervene to set $X = x'$; and a reference space $\mathcal{R}$, in which we observe x', y' but intervene to set $X = x$. Let $\mathcal{D}$ denote a uniform mixture over both spaces, and let auxiliary variable W tag each sample with a label indicating whether it comes from the original ($W = 1$) or contrastive ($W = 0$) counterfactual space. Define $c(\boldsymbol{z}) = w$. Then we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, y')$.

In other words, we regard Pearl's notion of necessity as *sufficiency of the negated factor for the alternative outcome*. By contrast, Pearl [2000] has no analogue for our probability of necessity. This is true of any measure that defines sufficiency and necessity via inverse, rather than converse probabilities. While conditioning on the same variable(s) for both measures may have some intuitive appeal, it comes at a cost to expressive power. Whereas our framework can recover all four explanatory measures, corresponding to the classical definitions and their contrapositive forms, definitions that merely negate instead of transpose the antecedent and consequent are limited to just two.

Remark 3. We have assumed that factors and outcomes are Boolean throughout. Our results can be extended to continuous versions of either or both variables, so long as $c(\boldsymbol{Z}) \perp\!\!\!\perp Y \mid \boldsymbol{Z}$. This conditional independence holds whenever $\boldsymbol{W} \perp\!\!\!\perp Y \mid \boldsymbol{X}$, which is true by construction since $f(\boldsymbol{z}) := f(\boldsymbol{x})$. However, we defend the Boolean assumption on the grounds that it is well motivated by contrastivist epistemologies [Kahneman and Miller, 1986; Blaauw, 2013] and not especially restrictive, given that partitions of arbitrary complexity may be defined over $\boldsymbol{Z}$ and Y.

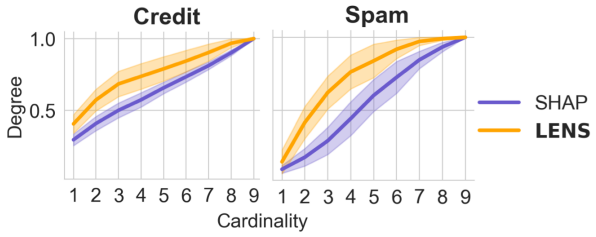

Figure 2: Comparison of top k features ranked by SHAP against the best performing LENS subset of size k in terms of $PS(c, y)$. `German` results are over 50 inputs; `SpamAssassins` results are over 25 inputs.

5 EXPERIMENTS

In this section, we demonstrate the use of LENS on a variety of tasks and compare results with popular XAI tools, using the basis configurations detailed in Table 1. A comprehensive discussion of experimental design, including datasets and pre-processing pipelines, is left to Appendix C. Code for reproducing all results is available at `https://github.com/limorigu/LENS`.

Contexts. We consider a range of contexts $\mathcal{D}$ in our experiments. For the input-to-reference (I2R) setting, we replace input values with reference values for feature subsets S; for the reference-to-input (R2I) setting, we replace reference values with input values. We use R2I for examining sufficiency/necessity of the original model prediction, and I2R for examining sufficiency/necessity of a contrastive model prediction. We sample from the empirical data in all experiments, except in Sect. 5.3, where we assume access to a structural causal model (SCM).

Partial Orderings. We consider two types of partial orderings in our experiments. The first, $\preceq_{subset}$, evaluates subset relationships. For instance, if $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\text{gender} = \text{“female”}]]$ and $c'(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\text{gender} = \text{“female”} \wedge \text{age} \geq 40]]$, then we say that $c \preceq_{subset} c'$. The second, $c \preceq_{cost} c' := c \preceq_{subset} c' \wedge cost(c) \leq cost(c')$, adds the additional constraint that c has cost no greater than c'. The cost function could be arbitrary. Here, we consider distance measures over either the entire state space or just the intervention targets corresponding to c.

Table 1: Overview of experimental settings by basis configuration.

Experiment	Datasets	f	$\mathcal{D}$	$\mathcal{C}$	$\preceq$
Attribution comparison	`German`, `SpamAssassins`	`Extra-Trees`	R2I, I2R	Intervention targets	-
Anchors comparison: Brittle predictions	`IMDB`	`LSTM`	R2I, I2R	Intervention targets	$\preceq_{subset}$
Anchors comparison: PS and Prec	`German`	`Extra-Trees`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Adverserial	`SpamAssassins`	`MLP`	R2I	Intervention targets	$\preceq_{subset}$
Counterfactuals: Recourse, DiCE comparison	`Adult`	`MLP`	I2R	Full interventions	$\preceq_{cost}$
Counterfactuals: Recourse, causal vs. non-causal	`German`	`Extra-Trees`	I2R$_{causal}$	Full interventions	$\preceq_{cost}$

5.1 FEATURE ATTRIBUTIONS

Feature attributions are often used to identify the top-k most important features for a given model outcome [Barocas et al., 2020]. However, we argue that these feature sets may not be explanatory with respect to a given prediction. To show this, we compute R2I and I2R sufficiency – i.e., $PS(c, y)$ and $PS(1-c, 1-y)$, respectively – for the top-k most influential features ($k \in [1, 9]$) as identified by SHAP [Lundberg and Lee, 2017] and LENS. Fig. 2 shows results from the R2I setting for `German` credit [Dua and Graff, 2017] and `SpamAssassin` datasets [SpamAssassin, 2006]. Our method attains higher PS for all cardinalities. We repeat the experiment over 50 inputs, plotting means and 95% confidence intervals for all k. Results indicate that our ranking procedure delivers more informative explanations than SHAP at any fixed degree of sparsity. Results from the I2R setting are in Appendix C.

5.2 RULE LISTS

Sentiment sensitivity analysis. Next, we use LENS to study model weaknesses by considering minimal factors with high R2I and I2R sufficiency in text models. Our goal is to answer questions of the form, "What are words with/without which our model would output the original/opposite prediction for an input sentence?" For this experiment, we train an LSTM network on the `IMDB` dataset for sentiment analysis [Maas et al., 2011]. If the model mislabels a sample, we investigate further; if it does not, we inspect the most explanatory factors to learn more about model behavior. For the purpose of this example, we only inspect sentences of length 10 or shorter. We provide two examples below and compare with Anchors (see Table 2).

Consider our first example: READ BOOK FORGET MOVIE is a sentence we would expect to receive a negative prediction, but our model classifies it as positive. Since we are investigating a positive prediction, our reference space is conditioned on a negative label. For this model, the classic UNK token receives a positive prediction. Thus we opt for an alternative, PLATE. Performing interventions on all possible combinations of words with our token, we find the conjunction of READ, FORGET, and MOVIE is a sufficient factor for a positive prediction (R2I). We also find that changing any of READ, FORGET, or MOVIE to PLATE would result in a negative prediction (I2R). Anchors, on the other hand, perturbs the data stochastically (see Appendix C), suggesting the conjunction READ AND BOOK. Next, we investigate the sentence: YOU BETTER CHOOSE PAUL VERHOEVEN EVEN WATCHED. Since the label here is negative, we use the UNK token. We find that this prediction is brittle – a change of almost any word would be sufficient to flip the outcome. Anchors, on the other hand, reports a conjunction including most words in the sentence. Taking the R2I view, we still find a more concise explanation: CHOOSE or EVEN would be enough to attain a negative prediction. These brief examples illustrate how LENS may be used to find brittle predictions across samples, search for similarities between errors, or test for model reliance on sensitive attributes (e.g., gender pronouns).

Anchors comparison. Anchors also includes a tabular variant, against which we compare LENS's performance in terms of R2I sufficiency. We present the results of this comparison in Fig. 3, and include additional comparisons in Appendix C. We sample 100 inputs from the `German` dataset, and query both methods with $\tau = 0.9$ using the classifier from Sect. 5.1. Anchors satisfies a PAC bound controlled by parameter δ. At the default value $\delta = 0.1$, Anchors fails to meet the τ threshold on 14% of samples; LENS meets it on 100% of samples. This result accords with Thm. 1, and vividly demonstrates the benefits of our optimality guarantee. Note that we also go beyond Anchors in providing multiple explanations instead of just a single output, as well as a cumulative probability measure with no analogue in their algorithm.

5.3 COUNTERFACTUALS

Adversarial examples: spam emails. R2I sufficiency answers questions of the form, "What would be sufficient for the model to predict y?". This is particularly valuable in cases with unfavorable outcomes y'. Inspired by adversarial interpretability approaches [Ribeiro et al., 2018b; Lakkaraju and Bastani, 2020], we train an MLP classifier on the `SpamAssassins` dataset and search for minimal factors sufficient to relabel a sample of spam emails as non-spam. Our examples follow some patterns common to spam emails: received from unusual email addresses, includes suspicious keywords such as ENLARGEMENT or

Table 2: Example prediction given by an LSTM model trained on the IMDB dataset. We compare τ-minimal factors identified by LENS (as individual words), based on $PS(c, y)$ and $PS(1-c, 1-y)$, and compare to output by Anchors.

Inputs		Anchors		LENS	
Text	Original model prediction	Suggested anchors	Precision	Sufficient R2I factors	Sufficient I2R factors
'read book forget movie'	wrongly predicted positive	[read, movie]	0.94	[read, forget, movie]	read, forget, movie
'you better choose paul verhoeven even watched'	correctly predicted negative	[choose, better, even, you, paul, verhoeven]	0.95	choose, even	better, choose, paul, even

Table 3: (Top) A selection of emails from SpamAssassins, correctly identified as spam by an MLP. The goal is to find minimal perturbations that result in non-spam predictions. (Bottom) Minimal subsets of feature-value assignments that achieve non-spam predictions with respect to the emails above.

From	To	Subject	First Sentence	Last Sentence
resumevalet info resumevalet com jacqui devito goodroughy ananzi co za rose xu email com	yyyy cv spamassassin taint org picone linux midrange com yyyyac idt net	adv put resume back work enlargement breakthrough zibdrzpay adv harvest lots target email address quickly	dear candidate recent survey conducted want	professionals online network inc increase size enter detailsto come open advertisement persons 18yrs old

Gaming options	Feature subsets for value changes	
1	From	To
	crispin cown crispin wirex com	example com mailing... list secprog securityfocus... moderator
2	From	First Sentence
	crispin cowan crispin wirex com	scott mackenzie wrote
3	From	First Sentence
	tim one comcast net tim peters	tim

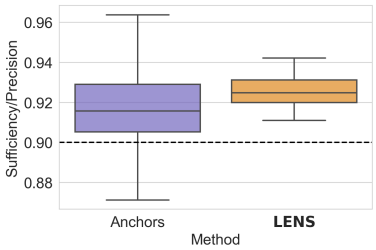

Figure 3: We compare $PS(c, y)$ against precision scores attained by the output of LENS and Anchors for examples from German. We repeat the experiment for 100 inputs, and each time consider the single example generated by Anchors against the mean $PS(c, y)$ among LENS's candidates. Dotted line indicates $\tau = 0.9$.

ADVERTISEMENT in the subject line, etc. We identify minimal changes that will flip labels to non-spam with high probability. Options include altering the incoming email address to more common domains, and changing the subject or first sentences (see Table 3). These results can improve understanding of both a model's behavior and a dataset's properties.

Diverse counterfactuals. Our explanatory measures can also be used to secure algorithmic recourse. For this experiment, we benchmark against DiCE [Mothilal et al., 2020b], which aims to provide diverse recourse options for any underlying prediction model. We illustrate the differences between our respective approaches on the Adult dataset [Kochavi and Becker, 1996], using an MLP and following the procedure from the original DiCE paper.

According to DiCE, a diverse set of counterfactuals is

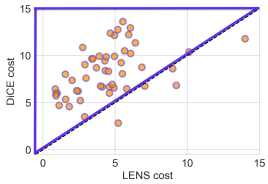

Figure 4: A comparison of mean cost of outputs by LENS and DiCE for 50 inputs sampled from the Adult dataset.

one that differs in *values* assigned to features, and can thus produce a counterfactual set that includes different interventions on the same variables (e.g., CF1: age $=$ 91, occupation = "retired"; CF2: age $=$ 44, occupation = "teacher"). Instead, we look at diversity of counterfactuals in terms of intervention *targets*, i.e. features changed (in this case, from input to reference values) and their effects. We present minimal cost interventions that would lead to recourse for each feature set but we summarize the set of paths to recourse via subsets of features changed. Thus, DiCE provides answers of the form "Because you are not 91 and retired" or "Because you are not 44 and a teacher"; we answer "Because of your age and occupation", and present the lowest cost intervention on these features sufficient to flip the prediction.

With this intuition in mind, we compare outputs given by DiCE and LENS for various inputs. For simplicity, we let all features vary independently. We consider two metrics for comparison: (a) the mean cost of proposed factors, and (b) the number of minimally valid candidates proposed, where a factor c from a method M is *minimally valid* iff for all c' proposed by M', $\neg(c' \prec_{cost} c)$ (i.e., M' does not report a factor preferable to c). We report results based on 50 randomly sampled inputs from the Adult dataset, where references are fixed by conditioning on the opposite prediction. The cost comparison results are shown in Fig. 4, where we find that LENS identifies lower cost factors for the vast majority of inputs. Furthermore, DiCE finds no minimally valid candidates that LENS did not already account for. Thus LENS

Table 4: Recourse example comparing causal and non-causal (i.e., feature independent) $\mathcal{D}$. We sample a single input example with a negative prediction, and 100 references with the opposite outcome. For I2R$_{causal}$ we propagate the effects of interventions through a user-provided SCM.

input									I2R		I2R$_{causal}$	
Age	Sex	Job	Housing	Savings	Checking	Credit	Duration	Purpose	τ-minimal factors ($\tau = 0$)	Cost	τ-minimal factors ($\tau = 0$)	Cost
23	Male	Skilled	Free	Little	Little	1845	45	Radio/TV	Job: Highly skilled	1	Age: 24	0.07
									Checking: NA	1	Sex: Female	1
									Duration: 30	1.25	Job: Highly skilled	1
									Age: 65, Housing: Own	4.23	Housing: Rent	1
									Age: 34, Savings: N/A	1.84	Savings: N/A	1

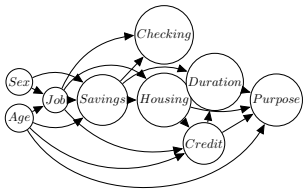

Figure 5: Example DAG for `German` dataset.

emphasizes *minimality* and *diversity* of intervention targets, while still identifying low cost intervention values.

Causal vs. non-causal recourse. When a user relies on XAI methods to plan interventions on real-world systems, causal relationships between predictors cannot be ignored. In the following example, we consider the DAG in Fig. 5, intended to represent dependencies in the `German` credit dataset. For illustrative purposes, we assume access to the structural equations of this data generating process. (There are various ways to extend our approach using only partial causal knowledge as input [Karimi et al., 2020b; Heskes et al., 2020].) We construct D by sampling from the SCM under a series of different possible interventions. Table 4 describes an example of how using our framework with augmented causal knowledge can lead to different recourse options. Computing explanations under the assumption of feature independence results in factors that span a large part of the DAG depicted in Fig. 5. However, encoding structural relationships in D, we find that LENS assigns high explanatory value to nodes that appear early in the topological ordering. This is because intervening on a single root factor may result in various downstream changes once effects are fully propagated.

6 DISCUSSION

Our results, both theoretical and empirical, rely on access to the relevant context $\mathcal{D}$ and the complete enumeration of all feature subsets. Neither may be feasible in practice. When elements of $\boldsymbol{Z}$ are estimated, as is the case with the generative methods sometimes used in XAI, modeling errors could lead to suboptimal explanations. For high-dimensional settings such as image classification, LENS cannot be naïvely applied without substantial data pre-processing. The first issue is extremely general. No method is immune to model misspecification, and attempts to recreate a data generating process must always be handled with care. Empirical sampling, which we rely on above, is a reasonable choice when data are fairly abundant and representative. However, generative models may be necessary to correct for known biases or sample from low-density regions of the feature space. This comes with a host of challenges that no XAI algorithm alone can easily resolve. The second issue – that a complete enumeration of all variable subsets is often impractical – we consider to be a feature, not a bug. Complex explanations that cite many contributing factors pose *cognitive* as well as computational challenges. In an influential review of XAI, Miller [2019] finds near unanimous consensus among philosophers and social scientists that, "all things being equal, simpler explanations – those that cite fewer causes... are better explanations" (p. 25). Even if we could list all τ-minimal factors for some very large value of d, it is not clear that such explanations would be helpful to humans, who famously struggle to hold more than seven objects in short-term memory at any given time [Miller, 1955]. That is why many popular XAI tools include some sparsity constraint to encourage simpler outputs. Rather than throw out some or most of our low-level features, we prefer to consider a higher level of abstraction, where explanations are more meaningful to end users. For instance, in our `SpamAssassins` experiments, we started with a pure text example, which can be represented via high-dimensional vectors (e.g., word embeddings). However, we represent the data with just a few intelligible components: `From` and `To` email addresses, `Subject`, etc. In other words, we create a more abstract object and consider each segment as a potential intervention target, i.e. a candidate factor. This effectively compresses a high-dimensional dataset into a 10-dimensional abstraction. Similar strategies could be used in many cases, either through domain knowledge or data-driven clustering and dimensionality reduction techniques. In general, if data cannot be represented by a reasonably low-dimensional, intelligible abstraction, then post-hoc XAI methods are unlikely to be of much help.

7 CONCLUSION

We have presented a unified framework for XAI that foregrounds necessity and sufficiency, which we argue are the fundamental building blocks of all successful explanations. We defined simple measures of both, and showed how they undergird various XAI methods. Our formulation, which relies on converse rather than inverse probabilities, is uniquely flexible and expressive. It covers all four basic explanatory measures – i.e., the classical definitions and their contrapositive transformations – and unambiguously accommodates logical, probabilistic, and/or causal interpretations, depending on how one constructs the basis tuple $\mathcal{B}$. We illustrated illuminating connections between our measures and existing proposals in XAI, as well as Pearl [2000]'s probabilities of causation. We introduced a sound and complete algorithm for identifying minimally sufficient factors, and demonstrated our method on a range of tasks and datasets. Our approach prioritizes completeness over efficiency, suitable for settings of moderate dimensionality. Future research will explore more scalable approximations, model-specific variants optimized for, e.g., convolutional neural networks, and developing a graphical user interface.

Author Contributions

David S. Watson and Limor Gultchin contributed equally to this paper.

Acknowledgements

DSW was supported by ONR grant N62909-19-1-2096.

References

K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artif. Intell.*, 298: 103502, 2021.

S. Barocas, A. D. Selbst, and M. Raghavan. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *FAT**, pages 80–89, 2020.

U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley. Explainable machine learning in deployment. In *FAT**, pages 648–657, 2020.

M. Blaauw, editor. *Contrastivism in Philosophy*. Routledge, New York, 2013.

A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, pages 592–603, 2018.

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

C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. *arXiv* preprint, 2001.07417, 2020.

S. Galhotra, R. Pradhan, and B. Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *SIGMOD*, 2021.

S. Grover, C. Pulice, G. I. Simari, and V. S. Subrahmanian. Beef: Balanced english explanations of forecasts. *IEEE Trans. Comput. Soc. Syst.*, 6(2):350–364, 2019.

J. Y. Halpern. *Actual Causality*. The MIT Press, Cambridge, MA, 2016.

J. Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach. Part II: Explanations. *Br. J. Philos. Sci.*, 56(4):889–911, 2005.

T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *NeurIPS*, 2020.

A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI*, pages 1511–1519, 2019.

D. Kahneman and D. T. Miller. Norm theory: Comparing reality to its alternatives. *Psychol. Rev.*, 93(2):136–153, 1986.

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

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

R. Kochavi and B. Becker. Adult income dataset, 1996. URL `https://archive.ics.uci.edu/ml/datasets/adult`.

I. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *ICML*, pages 5491–5500, 2020.

H. Lakkaraju and O. Bastani. "How do I fool you?": Manipulating user trust via misleading black box explanations. In *AIES*, pages 79–85, 2020.

H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Faithful and customizable explanations of black box models. In *AIES*, pages 131–138, 2019.

D. Lewis. Causation. *J. Philos.*, 70:556–567, 1973.

P. Lipton. Contrastive explanation. *Royal Inst. Philos. Suppl.*, 27:247–266, 1990.

Z. Lipton. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43, 2018.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774. 2017.

A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, 2011.

J. Mackie. Causes and conditions. *Am. Philos. Q.*, 2(4): 245–264, 1965.

L. Merrick and A. Taly. The explanation game: Explaining machine learning models using shapley values. In *CD-MAKE*, pages 17–38. Springer, 2020.

G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 101(2):343–352, 1955.

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

C. Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Münich, 2021. URL `https://christophm.github.io/interpretable-ml-book/`.

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

R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pages 607–617, 2020b.

N. Narodytska, A. Shrotri, K. S. Meel, A. Ignatiev, and J. Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *SAT*, pages 267–278, 2019.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Y. Ramon, D. Martens, F. Provost, and T. Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Adv. Data Anal. Classif.*, 2020.

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

M. T. Ribeiro, S. Singh, and C. Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *ACL*, pages 856–865, 2018b.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

L. Shapley. A value for *n*-person games. In *Contributions to the Theory of Games*, chapter 17, pages 307–317. Princeton University Press, Princeton, 1953.

A. SpamAssassin, 2006. URL `https://spamassassin.apache.org/old/publiccorpus/`. Accessed 2021.

J. D. Storey. The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Royal Stat. Soc. Ser. B Methodol.*, 69(3):347–368, 2007.

M. Sundararajan and A. Najmi. The many Shapley values for model explanation. In *ACM*, New York, 2019.

J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.*, 28(1-4):287–313, 2000.

B. Ustun, A. Spangher, and Y. Liu. Actionable recourse in linear classification. In *FAT**, pages 10–19, 2019.

T. J. VanderWeele and T. S. Richardson. General theory for interactions in sufficient cause models with dichotomous exposures. *Ann. Stat.*, 40(4):2128–2161, 2012.

T. J. VanderWeele and J. M. Robins. Empirical and counterfactual conditions for sufficient cause interactions. *Biometrika*, 95(1):49–61, 2008.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.*, 31 (2):841–887, 2018.

D. S. Watson and L. Floridi. The explanation game: a formal framework for interpretable machine learning. *Synthese*, 2020.

X. Zhang, A. Solar-Lezama, and R. Singh. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *NeurIPS*, page 4879–4890, 2018.