Minds and Machines (2022) 32:185–218
https://doi.org/10.1007/s11023-022-09598-7

ORIGINAL ARTICLE

Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice

David S. Watson[1] · **Limor Gultchin**[2,3] · **Ankur Taly**[4] · **Luciano Floridi**[5,6]

Received: 11 August 2021 / Accepted: 23 February 2022 / Published online: 16 March 2022

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. In this article, an expanded version of a paper originally presented at the 37th Conference on Uncertainty in Artificial Intelligence (Watson et al., 2021), we attempt to fill this gap. 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 propose a novel formulation of these concepts, and demonstrate its advantages over leading alternatives. We present a sound and complete algorithm for computing explanatory factors with respect to a given context and set of agentive preferences, allowing users to identify necessary and sufficient conditions for desired outcomes at minimal cost. Experiments on real and simulated data confirm our method's competitive performance against state of the art XAI tools on a diverse array of tasks.

Keywords Explainable artificial intelligence · Interpretable machine learning · Shapley values · Rule lists · Counterfactuals

David S. Watson and Limor Gultchin have contributed equally to this study.

✉ David S. Watson
david.watson@ucl.ac.uk

✉ Limor Gultchin
limor.gultchin@gmail.com

[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

[6] Department of Legal Studies, University of Bologna, Bologna, Italy

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. This raises fundamental issues of trust, fairness, and accountability that cannot be easily resolved. Post-hoc, model-agnostic local explanation tools—algorithms designed to shed light on the individual predictions of other algorithms—are at the forefront of a fast-growing research area dedicated to addressing these concerns. Prominent examples include feature attributions, rule lists, and counterfactuals, each of which will be critically examined below. The subdiscipline of computational statistics devoted to this problem is variously referred to as *interpretable machine learning* or *explainable artificial intelligence* (XAI). For recent reviews, see Murdoch et al. (2019), Rudin et al. (2021), and Linardatos et al. (2021).

Many authors have pointed out the inconsistencies between popular XAI tools, raising questions as to which method is more reliable in particular cases (Krishna et al., 2022; Mothilal et al., 2021; Ramon et al., 2020). Theoretical foundations have proven elusive in this area, perhaps due to the perceived subjectivity inherent to notions such as "simple" and "relevant" (Watson & 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). This method, formally defined in Sect. 4, quantifies the individual contribution of each feature toward a particular prediction. However, ambiguities regarding the underlying assumptions of existing software (Kumar et al., 2020) and the recent proliferation of mutually incompatible implementations (Merrick & Taly, 2020; Sundararajan & Najmi, 2019) have complicated this picture. Despite the abundance of alternative XAI tools (Molnar, 2019), 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 in the theory and practice of XAI. In this article, an expanded version of a paper originally presented at the 37th Conference on Uncertainty in Artificial Intelligence (Watson et al., 2021), we propose new formal and computational methods to operationalize this insight. Whereas our original publication focused largely on the properties and performance of our proposed algorithm, in this work we elaborate on the conceptual content of our approach, which relies on a subtle distinction between inverse and converse probabilities, as well as a pragmatic commitment to context-dependent, agent-oriented explanations.[1]

[1] Publishing journal versions of conference papers is relatively common in computer science; less so in philosophy. Our goal with this article is not only to expand upon the original work, but also to share it with a different readership that may be less likely to peruse the pages of the UAI proceedings. As a fundamentally interdisciplinary undertaking, this paper should, we hope, be of interest to researchers from both communities.

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. Our framework is flexible and pragmatic, enabling users to incorporate domain knowledge, search various subspaces, and select a utility-maximizing explanation. (2) We introduce novel measures of necessity and sufficiency that can be computed for any feature subset. Our definitions are uniquely expressive and accord better with intuition than leading alternatives on challenging examples. (3) We present a sound and complete algorithm for identifying explanatory factors, and illustrate its performance on a range of tasks.

The remainder of this paper is structured as follows. Following a review of related work (Sect. 2), we introduce a unified framework (Sect. 3) that reveals unexpected affinities between various XAI tools and fundamental quantities in the study of causation (Sect. 4). We proceed to implement a novel procedure for computing model explanations that improves upon the state of the art in quantitative and qualitative comparisons (Sect. 5). After a brief discussion (Sect. 6), we conclude with a summary and directions for future work (Sect. 7).

2 Necessity and Sufficiency

Necessity and sufficiency have a long philosophical tradition, 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 immediately suggest probabilistic relaxations, in which we measure the sufficiency of x for y by $P(y|x)$ and the necessity of x 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. This distinction between probabilistic conversion and inversion will be crucial to our proposal in Sect. 3, as well as our critique of alternative measures in Sect. 4. Counterintuitive implications of contrapositive relations abound, most famously in confirmation theory's raven paradox (Good, 1960; Hempel, 1945; Mackie, 1963), but also in the literature on natural language conditionals (Crupi & Iacona, 2020; Gomes, 2019; Stalnaker, 1981). Our formal framework aims to preserve intuition while extinguishing any potential ambiguity.

Logical and probabilistic definitions of necessity and sufficiency often fall short when we consider causal explanations (Tian & Pearl, 2000; Pearl, 2009). 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 (2009, Chap. 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. This problem is the primary focus of the copious literature on "actual causality", as famously laid out in a pair of influential articles by Halpern and Pearl (2005a, 2005b), and later given book-length treatment in a monograph by Halpern (2016). For a recent survey and refinement of the formal definitions, see Beckers (2021). The common thread in all these works, cashed out in various ways by philosophers including Mackie (1965) and Wright (2013), is that x causes y iff x is a necessary element of a sufficient set for y. These authors generally limit their analyses to Boolean systems with convenient structural properties. 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 pursue 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 (Grover et al., 2019; Lakkaraju et al., 2019). Others have combined classical work on prime implicants with recent advances in tractable Boolean circuits (Darwiche & Hirth, 2020). These methods typically derive 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 and scalable under certain restrictions on admissible feature subsets (see Sect. 6 for a discussion).

Mothilal et al. (2021) 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 (2009)'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 normative judgment that is inconsistent with many common uses of these terms.

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. As a running example, we will consider the case of a hypothetical loan applicant named Anne.[2]

3.1 The Basis Tuple

3.1.1 Target Function

Post-hoc explainability methods assume access to a target function $f : \mathcal{X} \mapsto \mathcal{Y}$, i.e. the machine learning 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, Anne 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. We generally regard f as deterministic, although stochastic variants can easily be accommodated.

3.1.2 Context

The context $\mathcal{D}$ is a probability distribution over which we quantify sufficiency and necessity.[3] Contexts may be constructed in various ways but always consist of at least some *input* (point or space) and *reference* (point or space). For

[2] In what follows, we use uppercase italics to represent variables, e.g. X; lowercase italics to represent their values, e.g. x; uppercase boldface to represent matrices, e.g. $\boldsymbol{X}$; lowercase boldface to represent vectors, e.g. $\boldsymbol{x}$; and calligraphic type to represent distributions or their support, e.g. $\mathcal{X}$. Occasional deviations, e.g. lowercase italic f to represent a function or uppercase C to represent a set, should be clear from the context.

[3] This use of "context" is not to be confused with the same term in the causal literature, where it typically refers to values for a set of unobserved exogenous features that serve as input to a structural causal model. See Pearl (2009) and Halpern (2016).

example, say Anne's loan application is denied. The specific values of all her recorded features constitute an input point. To figure out why she was unsuccessful, Anne may want to compare herself to some similar applicant who succeeded (i.e., a reference point), or perhaps the set of all successful applicants (i.e., a reference space). Alternatively, she may expand the input space to include all unsuccessful applicants of similar income and age range, and compare them to a reference class of successful applicants in this same income and age range. Anne may make this comparison by (optionally) exploring intermediate inputs that gradually make the input space more reference-like or vice versa. For instance, Anne may change the income of all applicants in the input space to some reference income. Contexts capture the range of all such intermediate inputs that Anne examines in comparing the input(s) and reference(s). This distribution provides a semantics for explanatory measures by bounding the scope of necessity and sufficiency claims.

Observe that the "locality" of Anne's explanation is determined by the extent to which input and reference spaces are restricted. An explanation that distinguishes *all* successful applicants from *all* unsuccessful applicants is by definition global. One that merely specifies why Anne failed, whereas someone very much like her succeeded, is local—perhaps even maximally so, if Anne's successful counterpart is as similar as possible to her without crossing the decision boundary. In between, we find a range of intermediate alternatives, characterized by spaces that overlap with Anne's feature values to varying degrees. Thus we can relax the hard boundary between types and tokens, so pervasive in the philosophical literature on explanation (Hausman, 2005), and admit instead a spectrum of generality that may in some cases be precisely quantified (e.g., with respect to some distance metric over the feature space).

In addition to predictors and outcomes, the context can optionally include information exogenous to f. A set of auxiliary variables $\boldsymbol{W}$ may span sensitive attributes like gender and race that are not recorded in $\boldsymbol{X}$, which Anne could use to audit for bias on the part of her bank. Other potential auxiliaries include engineered features, such as those learned via neural embeddings, or metadata about the conditioning events that characterize a given distribution. Crucially, such conditioning events need not be just observational. If, for example, Anne wants to compare her application to a treatment group of customers randomly assigned some promotional offer ($W = 1$), then her reference class is sampled from $P(\boldsymbol{X}|do(W = 1))$. Alternatively, W may index different distributions, serving the same function as so-called "regime indicators" in Dawid (2002, 2021)'s decision-theoretic approach to statistical causality. This augmentation allows us to evaluate the necessity and sufficiency of factors beyond those observed in $\boldsymbol{X}$. Going beyond observed data requires background assumptions (e.g., about structural dependencies) and/or statistical models (e.g., learned vector representations). Errors introduced by either may propagate to final explanations, so both should be handled with care. Contextual data take the form $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{W}) \sim \mathcal{D}$. We extend the target function to augmented inputs by defining $f(\boldsymbol{z}) := f(\boldsymbol{x})$.

3.1.3 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. Say Anne wants to know how her odds of receiving the bank loan might change following an intervention that sets her income to at least \$50,000. Then a relevant factor may be $c(\boldsymbol{z}) = \mathbb{1}[\boldsymbol{x}[\mathsf{gender} = \text{“female”}] \wedge \boldsymbol{w}[do(\mathsf{income} > \$50\text{k})]]$, which checks whether the random sample $\boldsymbol{z}$ corresponds to a female drawn from the relevant interventional distribution. 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. $\boldsymbol{x}$.[4] We assume a finite space of factors $\mathcal{C}$.

3.1.4 Partial Order

When multiple factors pass a given necessity or sufficiency threshold, users will tend to prefer some over others. Say Anne learns that either of two changes would be sufficient to secure her loan: increasing her savings or getting a college degree. She has just taken a new job and expects to save more each month as a result. At this rate, she could hit her savings target within a year. Quitting her job to go to college, by contrast, would be a major financial burden, one that would take years to pay off. Anne therefore judges that boosting her savings is preferable to getting a college degree—i.e., the former precedes the latter in her partial ordering of possible actions.

To the extent that XAI methods consider agentive preferences at all, they tend to focus on *minimality*. The idea is that, all else being equal, factors with fewer conditions and smaller changes are generally preferable to those with more conditions and greater changes. 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., Jeffrey, 1965; Savage, 1954; von Neumann & Morgenstern, 1944), thereby linking our formalization with a rich tradition of decision theory and associated methods for expected utility maximization.

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 possible factors, and $\preceq$ is a partial ordering on $\mathcal{C}$.

[4] For more on Pearl’s causal hierarchy and the distinction between observational, interventional, and counterfactual probabilities, see Pearl and Mackenzie (2018) and Bareinboim et al. (2021).

Table 1 Confusion matrix of labels (rows) and factors (columns), with accompanying definitions of the four fundamental explanatory probabilities

	$c(\boldsymbol{z})$	
$f(\boldsymbol{z})$	1	0
y	q_{11}	q_{10}
$1-y$	q_{01}	q_{00}

$$PS(c,y) = q_{11}/(q_{11}+q_{01})$$
$$PN(c,y) = q_{11}/(q_{11}+q_{10})$$
$$PS(1-c,1-y) = q_{00}/(q_{10}+q_{00})$$
$$PN(1-c,1-y) = q_{00}/(q_{01}+q_{00})$$

3.2 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:

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

Our definitions cast sufficiency and necessity as *converse* probabilities. We argue that this has major advantages over the more familiar inverse formulation, which has been dominant since Tian and Pearl (2000)'s influential paper, further developed and popularized in several subsequent publications (Halpern, 2016; Halpern & Pearl, 2005b; Pearl, 2009). To see why, observe that our notions of sufficiency and necessity 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, such as the true negative rate $P(c(\boldsymbol{z}) = 0 \mid f(\boldsymbol{z}) \neq y)$ and the negative predictive value $P(f(\boldsymbol{z}) \neq y \mid c(\boldsymbol{z}) = 0)$, which are contrapositive transformations of our proposed measures (see Table 1). We can recover these values exactly via $PN(1-c, 1-y)$ and $PS(1-c, 1-y)$, respectively. When necessity and sufficiency are defined as probabilistic inversions (rather than conversions), such

Table 2 Toy example of a contingency table for Anne's loan application

	BA	No BA	Total
Approved	5	10	15
Denied	45	40	85
Total	50	50	100

The gap between the number of successful applicants with a BA and the number of unsuccessful applicants without a BA pulls inverse and converse formulations of necessity apart

transformations are impossible. This is a major shortcoming given the explanatory significance of all four quantities, which correspond to probabilistic variants of the classical logical formulae for necessity and sufficiency. Definitions that can describe only two are fundamentally impoverished, bound to miss half the picture.

Pearl (2009) motivates the inverse formulation by interpreting his probabilities of causation as the tendency for an effect to respond to its cause in both ways—turning off in its absence, and turning on in its presence. As we show in the next section, these are better understood as two different sorts of sufficiency, i.e. the sufficiency of x for y and the sufficiency of $\neg x$ for $\neg y$ (see Proposition 4 for an exact statement of the correspondence). Our definition of necessity starts from a different intuition. We regard an explanatory factor as necessary to the extent that it covers all possible pathways to a given outcome. This immediately suggests our converse formulation, where we condition on the prediction itself—the "effect" in a causal scenario—and observe how often the factor in question is satisfied. Large values of $PN(c, y)$ suggest that there are few alternative routes to y except through c, which we argue is the essence of a necessary explanation.

In many cases, differences between inverse and converse notions of necessity will be negligible. Indeed, the two are strictly equivalent when classes are perfectly balanced (i.e., when $P(c(\boldsymbol{z}) = 1) = P(f(\boldsymbol{z}) = y) = 0.5$), or when the relationship between a factor and an outcome is deterministic (in which case we are back in the logical setting). More generally, the identity is obtained whenever $q_{11} = q_{00}$, to use the labels from Table 1. However, the greater the difference between these values, the more these two ratios diverge. Consider Anne's loan application. Say she wants to evaluate the necessity of college education for loan approval, so defines a factor that indicates whether applicants attained a bachelor's degree (BA). She samples some 100 individuals, with data summarized in Table 2. Observing that successful applicants are twice as likely to have no BA as they are to have one, we judge college education to be largely unnecessary for loan approval. Specifically, we have that $P(\text{"BA"}|\text{"Approved"}) = 1/3$. On an inverse notion of necessity, however, we get a very different result, with $P(\text{"Denied"}|\text{"No BA"}) = 4/5$. This counterintuitive conclusion overestimates the necessity of education by a factor of 2.4. A more persuasive interpretation of this quantity is that lacking a BA is often sufficient for loan denial—an informative discovery, perhaps, but not an answer to the original question, which asked to what extent college education was necessary for loan approval.

Pearl may plausibly object that this example is limited to observational data, and therefore uninformative with respect to causal mechanisms of interest. In fact,

our critique is far more general. For illustration, imagine that Table 2 represents the results of a randomized control trial (RCT) in which applicants are uniformly assigned to the "BA" and "No BA" groups.[5] Though counterfactual probabilities remain unidentifiable even with access to experimental data, Tian and Pearl (2000) demonstrate how to bound their probabilities of causation with increasing tightness as we make stronger structural assumptions. However, we are unconvinced that counterfactuals are even required here—and not just because of lingering metaphysical worries about the meaning of unobservable quantities such as $P(y_x, y_{x'})$ (Dawid, 2000; Quine, 1960). Instead, we argue that the relevant probabilities for causal sufficiency and necessity are simpler. Using the notation of regime indicators (Correa & Bareinboim, 2020; Dawid, 2021), let P_σ denote the probability distribution resulting from the stochastic regime imposed by our RCT, i.e. a trial in which college education is randomly assigned to all applicants with probability $1/2$. Then our arguments from above go through just the same, with the context $\mathcal{D}$ now given by P_σ.[6] We emphasize once again that we are perfectly capable of recovering Pearl's counterfactual definitions in our framework—see Proposition 4 below—but reiterate that probabilistic conversions are preferable to inversions even in causal contexts.

These toy examples illustrate a more general point. The converse formulation of necessity and sufficiency is not just more expressive than the inverse alternative, but also aligns more closely with our intuition when class imbalance pulls the two apart. In the following section, we present an optimal procedure for computing these quantities on real-world datasets, unifying a variety of XAI methods in the process.

[5] The plausibility of such a trial is beside the point. We could easily relabel the columns "Drug" and "Placebo", with rows "Response" and "Non-response".

[6] Observational and interventional probabilities only align under the assumption of conditional ignorability. However, nothing in our argument turns on this. We recycled Table 2 for ease of illustration. We only require some class imbalance to differentiate between converse and inverse formulations, regardless of whether this is observed in experimental or nonexperimental data.

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

3.3 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 Algorithm 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 $c' \prec c$.

Our next step is to span the τ-minimal factors and compute their cumulative *PN* (see probNec). Since no strictly preferable factor can match the sufficiency of a τ-minimal c, in reporting probability of necessity we expand C to its upward closure.

Theorems 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 1 for all proofs).

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

Population proportions may be obtained if the target function f is deterministic and data fully saturate the context $\mathcal{D}$, a plausible prospect with 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 $c \in \mathcal{C}$, Algorithm 1 is uniformly most powerful. That is, Algorithm 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).

Figure 1 provides a visual example of LENS outputs for a hypothetical loan application. We compute the minimal subvectors most likely to preserve or alter a given prediction, as well as cumulative necessity scores for all subsets. 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*. As noted above, the resulting trade-off is similar to that of a precision-recall curve quantifying and qualifying errors in classification tasks (see Fig. 2). Different degrees of necessity may be warranted for different tasks, depending on how important it is to (approximately)

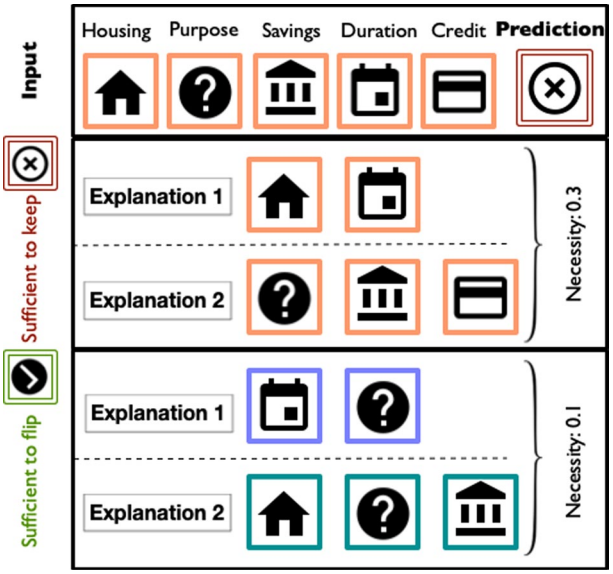

Fig. 1 A schematic overview of LENS outputs for an unsuccessful loan applicant. We describe minimal sufficient factors (here, sets of features) for a given input (top row), with the aim of preserving or flipping the original prediction. We report a sufficiency score for each set and a cumulative necessity 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)

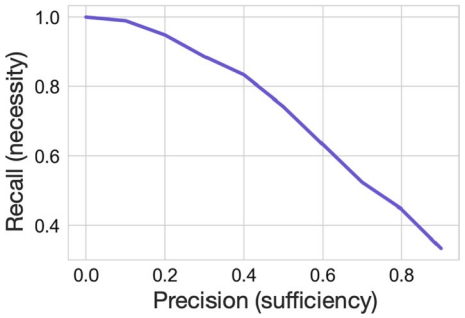

Fig. 2 An example curve visualizing the relationship between sufficiency and necessity from the `German` credit dataset (see Sect. 5). Setting τ amounts to thresholding the x-axis at a fixed point, with *PN* given by the corresponding y-coordinate of this curve

exhaust all paths towards an outcome. Users can therefore adjust τ to accommodate desired levels of *PN* over successive calls to LENS.

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.

4.1 Feature Attributions

Several popular feature attribution algorithms are based on Shapley values (Shapley, 1953), originally proposed as a solution to the attribution problem in cooperative game theory, which asks how best to distribute the surplus generated by a coalition of players. Substituting features for players and predictions for surplus, researchers have repurposed the method's combinatoric strategy for XAI to quantify the contribution of each input variable toward a given output. The goal is to decompose the predictions of any target function as a sum of weights over d features:

$$f(\boldsymbol{x}_i) = \sum_{j=0}^{d} \phi_j(i), \tag{1}$$

where $\phi_0(i)$ represents a baseline expectation and $\phi_j(i)$ the weight assigned to X_j at point $\boldsymbol{x}_i$.[7] Let $v : [n] \times 2^d \mapsto \mathbb{R}$ be a value function such that $v(i, S)$ is the payoff associated with feature subset $S \subseteq [d]$ for sample i and $v(i, \{\emptyset\}) = 0$ for all $i \in [n]$. 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:

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

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

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

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.

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 $\mathcal{D}(\cdot|S)$.[8]

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

[7] Shapley values can be computed for regression or classification tasks, although in the latter case class probabilities are required. While we treat f as binary for our purposes, most classifiers (including all those used in our experiments) also generate probabilities, which we use for benchmarking against Shapley values below (see Sect. 5.)

[8] 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 & 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.

Thus, the Shapley value $\phi_j(i)$ 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.

4.2 Rule Lists

Rule lists are sequences of if-then statements that describe hyperrectangles 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 (Friedman & Popescu, 2008; Letham et al., 2015), more recent efforts have prioritized local explanation tasks (Lakkaraju et al., 2019; Sokol & Flach, 2020).

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*, $\text{prec}(A)$, 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 Definition 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* $\text{prec}(A) = PS(c_A, y)$.

While Anchors output 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. Finally, by exhaustively searching over a space of candidate factors rather than engineering auxiliary variables on the fly, our method is certifiably sound and complete, whereas Anchors are at best probably approximately correct (i.e., satisfy a PAC bound).

4.3 Counterfactuals

Counterfactual explanations are rooted in the seminal work of Lewis (1973a, 1973b), who famously argued that a causal account of an event x should appeal to the nearest possible world in which $\neg x$. In XAI, this is accomplished by identifying 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).[9] The optimization problem is:

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

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 (Karimi et al., 2020a; Ustun et al., 2019; Wexler et al., 2020). 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. We interpret the *cost* function as encoding agentive preferences by representing the partial ordering on factors. This can be guaranteed under some constraints on $\preceq$; see Steele and Stefánsson (2020) for an overview of representation theorems in decision theory.

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* $c^*(\boldsymbol{z}) = 1$.

4.4 Probabilities of Causation

Our framework can describe Pearl (2009)'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* $X = 1, Y = 1$ *but intervene to set* $X = 0$*; and a reference space* $\mathcal{R}$*, in which we observe* $X = 0, Y = 0$ *but intervene to set* $X = 1$*. 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 input* $(W = 0)$ *or reference* $(W = 1)$ *distribution. Define* $c(\boldsymbol{z}) = w$*. Then we have* $\texttt{suf}(x, y) = PS(c, y)$ *and* $\texttt{nec}(x, y) = PS(1 - c, 1 - y)$.

[9] Confusingly, the term "counterfactual" in XAI refers to any point with an alternative outcome, whereas in the causal literature it denotes a space characterized by incompatible conditioning events (see Sect. 2). We will use the word in both senses, but strive to make our intended meaning explicit in each case.

In other words, we regard Pearl's notion of necessity as *sufficiency of the negated factor for the alternative outcome*. By contrast, Pearl (2009) has no analogue for our probability of necessity. This is true of any measure that defines necessity and sufficiency via inverse, rather than converse probabilities. While conditioning on the same variable(s) for both measures may have some intuitive appeal, especially in the causal setting, it comes at a substantial cost to expressive power. Whereas our framework can recover all four fundamental 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 1 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 (Blaauw, 2013; Kahneman & Miller, 1986; Lipton, 1990) and not especially restrictive, given that partitions of arbitrary complexity may be defined over $\boldsymbol{Z}$ and Y.

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 3. A comprehensive discussion of experimental design, including datasets and pre-processing pipelines, is left to Appendix 2. Code for reproducing all results is available at https://github.com/limorigu/LENS.

5.1 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 the sufficiency/necessity of the original model prediction, and I2R for examining the sufficiency/necessity of a contrastive model prediction. We sample from the empirical data in all experiments, except in Sect. 5.6.3, where we assume access to a structural causal model (SCM).

Table 3 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	$\preceq_{cost}$

5.2 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.

5.3 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 ($k \in [9]$) as identified by SHAP (Lundberg & Lee, 2017) and LENS. Fig. 3 shows results from the R2I setting for `German` credit (Dua & Graff, 2017) and `SpamAssassin` datasets (SpamAssassin, 2006). Our method attains higher PS for all cardinalities, indicating that our ranking procedure delivers more informative explanations than SHAP at any fixed degree of sparsity. Results from the I2R setting can be found in Appendix 2.

5.4 Rule Lists

5.4.1 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 4).

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

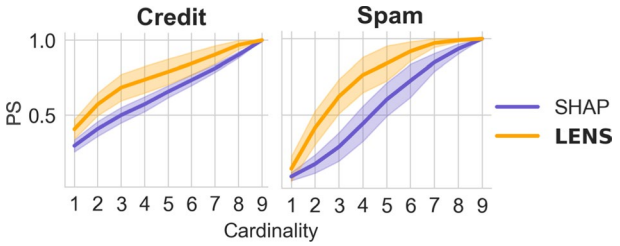

Fig. 3 Comparison of top k features ranked by SHAP against the best performing LENS subset of size k in terms of $PS(c, y)$. `German` results are over 50 inputs; `SpamAssassins` results are over 25 inputs. Shaded regions indicate 95% confidence intervals

Table 4 Example prediction given by an LSTM model trained on the `IMDB` dataset

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

We compare τ-minimal factors identified by LENS, based on $PS(c, y)$ and $PS(1 - c, 1 - y)$, and compare to output by Anchors

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 2), 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).

5.5 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. 4, and include additional comparisons in Appendix 2. We sample 100 inputs from the `German` dataset, and query both methods with $\tau = 0.9$ using the classifier from Sect. 5.3. Anchors satisfy a PAC bound controlled by parameter δ. At the default value $\delta = 0.1$, Anchors fail to meet the τ threshold on 14% of samples; LENS meets it on 100% of samples. This result accords with Theorem 1, and vividly demonstrates the benefits of our optimality guarantee. Note that we also go beyond

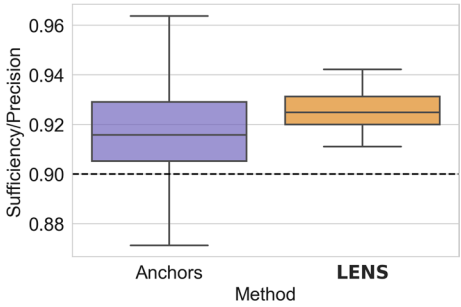

Fig. 4 We compare $PS(c, y)$ against precision scores attained by the output of LENS and Anchors for examples from `German`. We repeat the experiment for 100 inputs, and each time consider the single example generated by Anchors against the mean $PS(c, y)$ among LENS's candidates. Dotted line indicates $\tau = 0.9$

Anchors in providing multiple explanations instead of just a single output, as well as a cumulative probability measure with no analogue in their algorithm.

5.6 Counterfactuals

5.6.1 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 (Lakkaraju & Bastani, 2020; Ribeiro et al., 2018b), 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 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 5). These results can improve understanding of both a model's behavior and a dataset's properties.

5.6.2 Diverse Counterfactuals

Our explanatory measures can also be used to secure algorithmic recourse. For this experiment, we benchmark against DiCE (Mothilal et al., 2020), 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 & Becker, 1996), using an MLP and following the procedure from the original DiCE paper.

According to DiCE, a diverse set of counterfactuals is 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: $\mathsf{age} = 91, \mathsf{occupation} =$ "retired"; CF2: $\mathsf{age} = 44, \mathsf{occupation} =$ "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

Table 5 (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

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', $c \succeq_{cost} c'$ (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. 5, 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 emphasizes *minimality* and *diversity* of intervention targets, while still identifying low cost intervention values.

5.6.3 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. 6, 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 (Heskes et al., 2020; Karimi et al., 2020b).] We construct D by sampling from the SCM under a series of different possible interventions. Table 6 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. 6. However, encoding structural

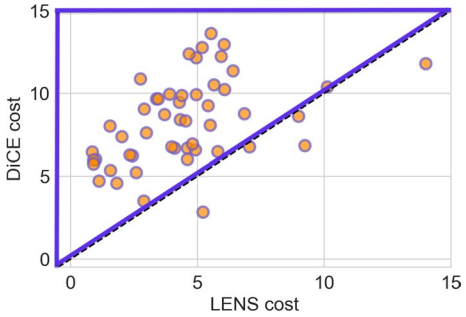

Fig. 5 A comparison of mean cost of outputs by LENS and DiCE for 50 inputs sampled from the `Adult` dataset

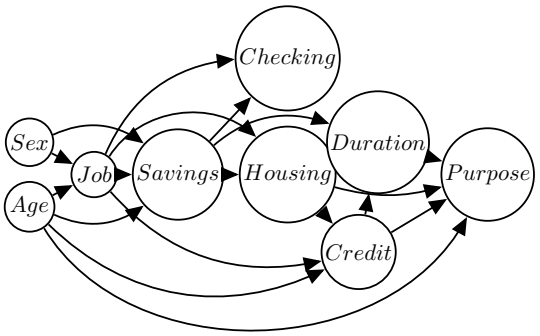

Fig. 6 Example DAG for `German` dataset

Table 6 Recourse example comparing causal and non-causal (i.e., feature independent) $\mathcal{D}$.

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

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

relationships in $\mathcal{D}$, 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 true context $\mathcal{D}$ and the complete enumeration of all relevant feature subsets. Neither may be feasible in practice. When elements of $\boldsymbol{Z}$ are based on assumptions about structural dependencies or estimated from data via some statistical model, 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, regarding the difficulty of the optimal subset selection procedure, is somewhat subtler. First, we observe that the problem is only NP-hard in the worst case. Partial orderings may vastly reduce the complexity of the task by, for instance, encoding a preference for greedy feature selection, or pruning the search space through branch and bound techniques, as our $\preceq_{subset}$ ordering does above. Thus agents with appropriate utility functions can always ensure efficient computation. Second, we emphasize that complex explanations citing 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 efficiently compute all τ-minimal factors for some 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 (Floridi, 2008), 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 (Hilgard et al., 2021; Kim et al., 2018; Koh et al., 2020) or data-driven clustering and dimensionality reduction techniques (Beckers et al., 2019; Chalupka et al., 2017; Kinney & Watson, 2020; Locatello et al., 2019). 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.

An anonymous reviewer raised concerns about the factor set $\mathcal{C}$, which is generally unconstrained in our formulation, and therefore may lead to explanations that are "not sensible". First, we note that unexplanatory factors should receive low probabilities of necessity and sufficiency, and therefore pose no serious problems in practice. Second, we observe that XAI practitioners generally query models with some hypotheses already in mind. For instance, Anne may want to know if her loan was denied due to her savings, her education, or her race. Perhaps none of these variables explains her unfavorable outcome, which would itself be informative. Her effort to understand the bank's credit risk model may well be circuitous, iterative, and occasionally less than fully sensible. Yet we strongly object to the notion that we could somehow automate the procedure of selecting the "right" factors in a subject-neutral, agent-independent

manner. We consider it a feature, not a bug, that LENS requires some user engagement to better understand model predictions. XAI is a tool, not a panacea.

7 Conclusion

We have presented a unified framework for XAI that foregrounds necessity and sufficiency, which we argue are the 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 expressive. It covers all four fundamental 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 argued that alternative formulations which rely on probabilistic inversion are better understood as alternative sufficiency measures. We illustrated illuminating connections between our framework and existing proposals in XAI, as well as Pearl (2009)'s probabilities of causation. We introduced a sound and complete algorithm for identifying minimally sufficient factors—LENS—and demonstrated its performance on a range of tasks and datasets. The approach is flexible and pragmatic, accommodating background knowledge and explanatory preferences as input. Though LENS prioritizes completeness over efficiency, the method may provide both for agents with certain utility functions. Future research will explore more scalable approximations and model-specific variants. User studies will guide the development of heuristic defaults and a graphical interface.

Appendix 1: Proofs

Theorems

Proof of Theorem 1

Theorem *With oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$, Algorithm 1 is sound and complete.*

Proof Soundness and completeness follow directly from the specification of (P1) $\mathcal{C}$ and (P2) $\preceq$ in the algorithm's input $\mathcal{B}$, along with (P3) access to oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$. Recall that the partial ordering must be complete and transitive, as noted in Sect. 3.

Assume that Algorithm 1 generates a false positive, i.e. outputs some c that is not τ-minimal. Then by Definition 4, either the algorithm failed to properly evaluate $PS(c, y)$, thereby violating (P3); or failed to identify some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. (i) contradicts (P3), and (ii) contradicts (P2). Thus there can be no false positives.

Assume that Algorithm 1 generates a false negative, i.e. fails to output some c that is in fact τ-minimal. By (P1), this c cannot exist outside the finite set $\mathcal{C}$.

Therefore there must be some $c \in \mathcal{C}$ for which either the algorithm failed to properly evaluate $PS(c, y)$, thereby violating (P3); or wrongly identified some c' such that (i) $PS(c', y) \geq \tau$ and (ii) $c' \prec c$. Once again, (i) contradicts (P3), and (ii) contradicts (P2). Thus there can be no false negatives. □

Proof of Theorem 2

Theorem *With sample estimates $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, Algorithm* 1 *is uniformly most powerful.*

Proof A testing procedure is uniformly most powerful (UMP) if it attains the lowest type II error β of all tests with fixed type I error α. Let Θ_0, Θ_1 denote a partition of the parameter space into null and alternative regions, respectively. The goal in frequentist inference is to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$ for some parameter θ. Let $\psi(X)$ be a testing procedure of the form $\mathbb{1}[T(X) \geq c_\alpha]$, where X is a finite sample, $T(X)$ is a test statistic, and c_α is the critical value. This latter parameter defines a rejection region such that test statistics integrate to α under H_0. We say that $\psi(X)$ is UMP iff, for any other test $\psi'(X)$ such that

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\psi'(X)] \leq \alpha,$$

we have

$$(\forall \theta \in \Theta_1)\ \mathbb{E}_\theta[\psi'(X)] \leq \mathbb{E}_\theta[\psi(X)],$$

where $\mathbb{E}_{\theta \in \Theta_1}[\psi(X)]$ denotes the power of the test to detect the true θ, $1 - \beta_\psi(\theta)$. The UMP-optimality of Algorithm 1 follows from the UMP-optimality of the binomial test (see Lehmann and Romano (2005), Chap. 3), which is used to decide between $H_0 : PS(c, y) < \tau$ and $H_1 : PS(c, y) \geq \tau$ on the basis of observed proportions $\hat{PS}(c, y)$, estimated from n samples for all $c \in \mathcal{C}$. The proof now takes the same structure as that of Theorem 1, with (P3) replaced by (P3′): access to UMP estimates of $PS(c, y)$. False positives are no longer impossible but bounded at level α; false negatives are no longer impossible but occur with frequency β. Because no procedure can find more τ-minimal factors for any fixed α, Algorithm 1 is UMP. □

Propositions

Proof of Proposition 1

Proposition *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 $v(S) = PS(c_S, y)$.*

As noted in the text, $\mathcal{D}(\boldsymbol{x}|S)$ may be defined in a variety of ways (e.g., via marginal, conditional, or interventional distributions). For any given choice, let $c_S(\boldsymbol{z}) = 1$ iff $\boldsymbol{x}$

is constructed by holding $\boldsymbol{x}_i^S$ fixed and sampling $\boldsymbol{X}^R$ according to $\mathcal{D}(\boldsymbol{x}|S)$. Since we assume binary Y (or binarized, as discussed in Sect. 3), we can rewrite Eq. 2 as a probability:

$$v(S) = P_{\mathcal{D}(\boldsymbol{x}|S)}(f(\boldsymbol{x}_i) = f(\boldsymbol{x})),$$

where $\boldsymbol{x}_i$ denotes the input point. Since conditional sampling is equivalent to conditioning after sampling, this value function is equivalent to $PS(c_S, y)$ by Definition 2.

Proof of Proposition 2

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

The proof for this proposition is essentially identical, except in this case our conditioning event is $A(\boldsymbol{x}) = 1$. Let $c_A = 1$ iff $A(\boldsymbol{x}) = 1$. Precision $\text{prec}(A)$, given by the lhs of Eq. 3, is defined over a conditional distribution $\mathcal{D}(\boldsymbol{x}|A)$. Since conditional sampling is equivalent to conditioning after sampling, this probability reduces to $PS(c_A, y)$.

Proof of Proposition 3

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

$$c^* = \underset{c \in \mathcal{C}}{\text{argmin}} \ \ cost(c) \text{ s.t. } PS(c, 1-y) \geq \tau, \tag{7}$$

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

There are two closely related ways of expressing the counterfactual objective: as a search for optimal *points*, or optimal *actions*. We use the latter interpretation, reframing actions as factors. We are only interested in solutions that flip the original outcome, and so we constrain the search to factors that meet an I2R sufficiency threshold, $PS(c, 1-y) \geq \tau$. Then the optimal action is attained by whatever factor (i) meets the sufficiency criterion and (ii) minimizes cost. Call this factor c^*. The optimal point is then any $\boldsymbol{z}$ such that $c^*(\boldsymbol{z}) = 1$.

Proof of Proposition 4

Proposition *Consider the bivariate Boolean setting, as in Sect. 2. We have two counterfactual distributions: an input space* $\mathcal{I}$, *in which we observe* $X = 1, Y = 1$ *but intervene to set* $X = 0$; *and a reference space* $\mathcal{R}$, *in which we observe* $X = 0, Y = 0$ *but intervene to set* $X = 1$. *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 input* $(W = 0)$ *or reference* $(W = 1)$ *distribution. Define* $c(\boldsymbol{z}) = w$. *Then we have* $\texttt{suf}(x, y) = PS(c, y)$ *and* $\texttt{nec}(x, y) = PS(1-c, 1-y)$.

Recall from Sect. 2 that (Pearl (2009), Ch. 9) defines $\texttt{suf}(x, y) := P(y_x|x', y')$ and $\texttt{nec}(x, y) := P(y'_{x'}|x, y)$. With the convention that $x' = 1 - x$, we may rewrite the former as $P_{\mathcal{R}}(Y = 1)$, where the reference space $\mathcal{R}$ denotes a counterfactual distribution conditioned on $X = 0, Y = 0, do(X = 1)$. Similarly, we may rewrite the latter as $P_{\mathcal{I}}(Y = 0)$, where the input space $\mathcal{I}$ denotes a counterfactual distribution conditioned on $X = 1, Y = 1, do(X = 0)$. Our context $\mathcal{D}$ is a uniform mixture over both spaces.

The key point here is that the auxiliary variable W indicates whether samples are drawn from $\mathcal{R}$ or $\mathcal{I}$. Thus conditioning on different values of W allows us to toggle between probabilities over the two spaces. Therefore, for $c(\boldsymbol{z}) = w$, we have $\texttt{suf}(x, y) = PS(c, y)$ and $\texttt{nec}(x, y) = PS(1 - c, 1 - y)$.

Appendix 2: Additional Discussions of Experimental Results

Data Pre-processing and Model Training

German Credit Risk

We first download the dataset from Kaggle,[10] which is a slight modification of the UCI version (Dua & Graff, 2017). We follow the pre-processing steps from a Kaggle tutorial.[11] In particular, we map the categorical string variables in the dataset (`Savings`, `Checking`, `Sex`, `Housing`, `Purpose` and the outcome `Risk`) to numeric encodings, and mean-impute values missing values for `Savings` and `Checking`. We then train an Extra-Tree classifier (Geurts et al., 2006) using scikit-learn, with random state 0 and max depth 15. All other hyperparameters are left to their default values. The model achieves a 71% accuracy.

German Credit Risk—Causal We assume a partial ordering over the features in the dataset, as described in Fig. 6. We use this DAG to fit a SCM based on the original data. In particular, we fit linear regressions for every continuous variable and a random forest classifier for every categorical variable. When sampling from $\mathcal{D}$, we let variables remain at their original values unless either (a) they are directly intervened on, or (b) one of their ancestors was intervened on. In the latter case, changes are propagated via the structural equations. We add stochasticity via Gaussian noise for continuous outcomes, with variance given by each model's residual mean squared error. For categorical variables, we perform multinomial sampling over predicted class probabilities. We use the same f model as for the non-causal `German` credit risk description above.

SpamAssassins The original spam assassins dataset comes in the form of raw, multi-sentence emails captured on the Apache SpamAssassins project, 2003-2015.[12] We segmented the emails to the following "features": `From` is the sender; `To` is the recipient; `Subject` is the email's subject line; `Urls` records any URLs found in

[10] See https://www.kaggle.com/kabure/german-credit-data-with-risk?select=german_credit_data.csv.
[11] See https://www.kaggle.com/vigneshj6/german-credit-data-analysis-python.
[12] See https://spamassassin.apache.org/old/credits.html.

the body; `Emails` denotes any email addresses found in the body; `First Sentence`, `Second Sentence`, `Penult Sentence`, and `Last Sentence` refer to the first, second, penultimate, and final sentences of the email, respectively. We use the original outcome label from the dataset (indicated by which folder the different emails were saved to). Once we obtain a dataset in the form above, we continue to pre-process by lower-casing all characters, only keeping words or digits, clearing most punctuation (except for '-' and '_'), and removing stopwords based on nltk's provided list (Bird et al., 2009). Finally, we convert all clean strings to their mean 50-dim GloVe vector representation (Pennington et al., 2014). We train a standard MLP classifier using scikit-learn, with random state 1, max iteration 300, and all other hyperparameters set to their default values.[13] This model attains an accuracy of 98.3%.

IMDB We follow the pre-processing and modeling steps taken in a standard tutorial on LSTM training for sentiment prediction with the IMDB dataset.[14] The CSV is included in the repository named above, and can be additionally downloaded from Kaggle or ai.standford.[15] In particular, these include removal of HTML-tags, non-alphabetical characters, and stopwords based on the the list provided in the ntlk package, as well as changing all alphabetical characters to lower-case. We then train a standard LSTM model, with 32 as the embedding dimension and 64 as the dimensionality of the output space of the LSTM layer, and an additional dense layer with output size 1. We use the sigmoid activation function, binary cross-entropy loss, and optimize with Kingma and Ba (2015). All other hyperparameters are set to their default values as specified by Keras.[16] The model achieves an accuracy of 87.03%.

Adult Income We obtain the adult income dataset via DiCE's implementation[17] and followed Haojun Zhu's pre-processing steps.[18] For our recourse comparison, we use a pretrained MLP model provided by the authors of DiCE, which is a single layer, non-linear model trained with TensorFlow and stored in their repository as 'adult.h5'.

Tasks

Comparison with Attributions

For completeness, we also include here comparison of cumulative attribution scores per cardinality with probabilities of sufficiency for the I2R view (see Fig. 7).

[13] See https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.

[14] See https://github.com/hansmichaels/sentiment-analysis-IMDB-Review-using-LSTM/blob/master/sentiment_analysis.py.ipynb.

[15] See https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews or http://ai.stanford.edu/~amaas/data/sentiment/.

[16] See https://keras.io.

[17] See https://github.com/interpretml/DiCE.

[18] See https://rpubs.com/H_Zhu/235617.

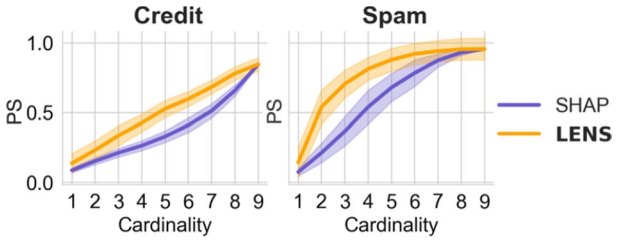

Fig. 7 Comparison of degrees of sufficiency in I2R setting, for top k features based on SHAP scores, against the best performing subset of cardinality k identified by our method. Results for `German` are averaged over 50 inputs; results for `SpamAssassins` are averaged over 25 inputs

Sentiment Sensitivity Analysis

We identify sentences in the original `IMDB` dataset that are up to 10 words long. Out of those, for the first example we only look at wrongly predicted sentences to identify a suitable example. For the other example, we simply consider a random example from the 10-word maximum length examples. We noted that Anchors uses stochastic word-level perturbations for this setting. This leads them to identify explanations of higher cardinality for some sentences, which include elements that are not strictly necessary. In other words, their outputs are not minimal, as required for descriptions of "actual causes" (Halpern, 2016; Halpern & Pearl, 2005a).

Comparison with Anchors

To complete the picture of our comparison with Anchors on the `German` Credit Risk dataset, we provide here additional results. In the main text, we included a comparison of Anchors's single output precision against the mean degree of sufficiency attained by our multiple suggestions per input. We sample 100 different inputs from the `German` Credit dataset and repeat this same comparison. Here we additionally consider the minimum and maximum $PS(c, y)$ attained by LENS against Anchors. Note that even when considering minimum PS suggestions by LENS, i.e. our worst output, the method shows more consistent performance. We qualify this discussion by noting that Anchors may generate results comparable to our own by setting the δ hyperparameter to a lower value. However, Ribeiro et al. (2018a) do not discuss this parameter in detail in either their original article or subsequent notebook guides. They use default settings in their own experiments, and we expect most practitioners will do the same.

Table 7 Recourse options for a single input given by DiCE and our method

input								DiCE output		**LENS output**	
Age	Wrkcls	Edu.	Marital	Occp.	Race	Sex	Hrs/week	Targets of intervention	Cost	Targets of intervention	Cost
42	Govt.	HS-grad	Single	Service	White	Male	40	Age, Edu., Marital, Hrs/week	8.13	Edu.	1
								Age, Edu., Marital, Occp., Sex, Hrs/week	5.866	Martial	1
								Age, Wrkcls, Educ., Marital, Hrs/week	5.36	Occp., Hrs/week	19.3
								Age, Edu., Occp., Hrs/week	3.2	Wrkcls, Occp., Hrs/week	12.6
								Edu., Hrs/week	11.6	Age, Wrkcls, Occp., Hrs/week	12.2

We report targets of interventions as suggested options, but they could correspond to different values of interventions. Our method tends to propose more minimal and diverse intervention targets. Note that all of DiCE's outputs are already subsets of LENS's two top suggestions, and due to τ-minimality LENS is forced to pick the next factors to be non-supersets of the two top rows. This explains the higher cost of LENS's bottom three rows

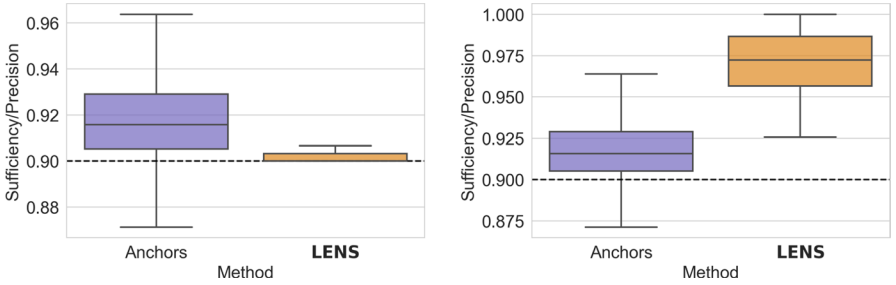

Fig. 8 We compare degree of sufficiency against precision scores attained by the output of LENS and Anchors for examples from `German`. We repeat the experiment for 100 sampled inputs, and each time consider the single output by Anchors against the min (left) and max (right) $PS(c, y)$ among LENS's multiple candidates. Dotted line indicates $\tau = 0.9$, the threshold we chose for this experiment

Recourse: DiCE Comparison

First, we provide a single illustrative example of the lack of diversity in intervention targets we identify in DiCE's output. Let us consider one example, shown in Table 7. While DiCE outputs are diverse in terms of values and target combinations, they tend to have great overlap in intervention targets. For instance, `Age` and `Education` appear in almost all of them. Our method would focus on minimal paths to recourse that would involve different combinations of features.

Next, we also provide additional results from our cost comparison with DiCE's output in Fig. 8. While in the main text we include a comparison of our mean cost output against DiCE's, here we additionally include a comparison of min and max cost of the methods' respective outputs. We see that even when considering minimum and maximum cost, our method tends to suggest lower cost recourse options. In particular, note that all of DiCE's outputs are already subsets of LENS's two top suggestions. The higher costs incurred by LENS for the next two lines are a reflection of this fact: due to τ-minimality, LENS is forced to find other interventions that are no longer supersets of options already listed above (Fig. 9).

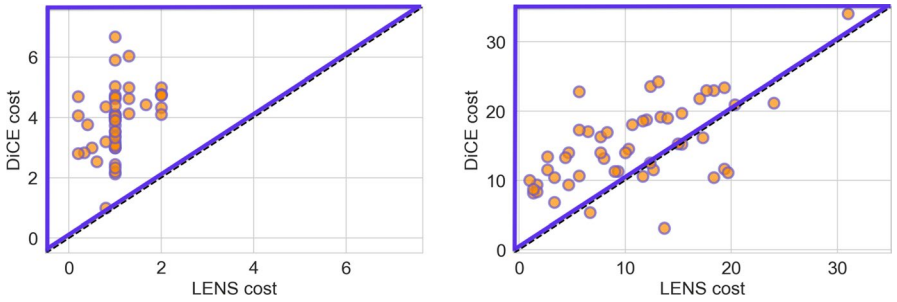

Fig. 9 We show results over 50 input points sampled from the original dataset, and all possible references of the opposite class, across two metrics: the min cost (left) of counterfactuals suggested by our method vs. DiCE, and the max cost (right) of counterfactuals

Acknowledgements DSW was supported by ONR Grant N62909-19-1-2096.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intellignece, 298*, 103502.

Bareinboim, E., Correa, J., Ibeling, D., & Icard, T. (2021). *On Pearl's hierarchy and the foundations of causal inference*. ACM.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 80–89).

Beckers, S. (2021). Causal sufficiency and actual causation. *Journal of Philosophical Logic 50*(6), 1341–1374.

Beckers, S., Eberhardt, F., & Halpern, J. Y. (2019). Approximate causal abstraction. In *Proceedings of the 35th conference on uncertainty in artificial intelligence* (pp. 210–219)

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 648–657).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly.

Blaauw, M. (Ed.). (2013). *Contrastivism in philosophy*. Routledge.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: An overview. *Behaviormetrika, 44*(1), 137–164.

Correa, J., & Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(6), 10093–10100.

Crupi, V., & Iacona, A. (2020). The evidential conditional. *Erkenntnis*. https://doi.org/10.1007/s10670-020-00332-2

Darwiche, A., & Hirth, A. (2020). On the reasons behind decisions. In *ECAI*.
Dawid, A. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95*(450), 407–424.
Dawid, A. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review 70*(2), 161–189.
Dawid, A. (2021). Decision-theoretic foundations for statistical causality. *Journal of Causal Inference, 9*(1), 39–77.
Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in neural information processing systems* (pp. 592–603).
Dua, D., & Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml
Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.
Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics, 2*(3), 916–954.
Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *SIGMOD*.
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.
Gomes, G. (2019). Meaning-preserving contraposition of conditionals. *Journal of Pragmatics, 1*(152), 46–60.
Good, I. (1960). The paradox of confirmation. *The British Journal for the Philosophy of Science, 11*(42), 145.
Grover, S., Pulice, C., Simari, G. I., & Subrahmanian, V. S. (2019). Beef: Balanced english explanations of forecasts. *IEEE Transactions on Computational Social Systems, 6*(2), 350–364.
Halpern, J. Y. (2016). *Actual causality*. MIT.
Halpern, J. Y., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887.
Halpern, J. Y., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science, 56*(4), 889–911.
Hausman, D. M. (2005). Causal relata: Tokens, types, or variables? *Erkenntnis, 63*(1), 33–54.
Hempel, C. G. (1945). Studies in the logic of confirmation (I). *Mind, 54*(213), 1–26.
Heskes, T., Sijben, E., Bucur, I. G., Claassen, T. (2020). Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in neural information processing systems*.
Hilgard, S., Rosenfeld, N., Banaji, M. R., Cao, J., & Parkes, D. (2021). Learning representations by humans, for humans. In *Proceedings of the 38th international conference on machine learning* (pp. 4227–4238).
Ignatiev, A., Narodytska, N., & Marques-Silva, J. (2019). Abduction-based explanations for machine learning models. In *AAAI* (pp. 1511–1519).
Jeffrey, R. C. (1965). *The logic of decision*. McGraw Hill.
Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153.
Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint. https://arxiv.org/abs/2010.04050
Karimi, A. H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. In *Advances in neural information processing systems*.
Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th international conference on machine learning* (pp. 2673–2682).
Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *The 3rd International conference for learning representations*.
Kinney, D., & Watson, D. (2020). Causal feature learning for utility-maximizing agents. In *Proceedings of the 10th international conference on probabilistic graphical models* (pp. 257–268). Skørping.
Kochavi, R., & Becker, B. (1996). Adult income dataset. https://archive.ics.uci.edu/ml/datasets/adult
Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020) Concept bottleneck models. In *Proceedings of the 37th international conference on machine learning* (pp. 5338–5348).

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, Z. S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint. https://arxiv.org/abs/2202.01602

Kumar, I., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th international conference on machine learning* (pp. 5491–5500).

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the 2020 AAAI/ACM conference on AI, ethics, and society* (pp. 79–85).

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 131–138).

Lehmann, E., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics, 9*(3), 1350–1371.

Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*, 556–567.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021) Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements, 27*, 247–266.

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

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th international conference on machine learning* (pp. 4114–4124).

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

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

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly, 2*(4), 245–264.

Mackie, J. L. (1963). The paradox of confirmation. *The British Journal for the Philosophy of Science, 13*(52), 265–277.

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

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

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

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models interpretable*. https://christophm.github.io/interpretable-ml-book/

Mothilal, R. K., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 652–663).

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America, 116*(44), 22071–22080.

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

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP* (pp. 1532–1543).

Quine. (1960). *Word and object*. MIT.

Ramon, Y., Martens, D., Provost, F., & Evgeniou, T. (2020). *A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC*. Advances in Data Analysis and Classification: LIME-C and SHAP-C.

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

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

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

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys, 16*, 1–85.

Savage, L. (1954). *The Foundations of Statistics*. New York: Dover Publications.

Shapley, L. (1953). A value for n-person games. In *Contributions to the theory of games* (Chap. 17, pp. 307–317). Princeton University Press.

Sokol, K., & Flach, P. (2020). LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. arXiv preprint. 2005.01427

SpamAssassin. (2006). Retrieved 2021, from https://spamassassin.apache.org/old/publiccorpus/

Stalnaker, R. C. (1981). *A theory of conditionals* (pp. 41–55). Springer.

Steele, K., & Stefánsson, H. O. (2020). Decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy, Winter* (2020th ed.). Metaphysics Research Laboratory, Stanford University.

Storey, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology, 69*(3), 347–368.

Sundararajan, M., & Najmi, A. (2019). *The many Shapley values for model explanation*. ACM.

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

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the 2019 conference on fairness, accountability, and transparency* (pp. 10–19).

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

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

Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. *Synthese, 198*, 9211–9242.

Watson, D. S., Gultchin, L., Taly, A., & Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*. *PMLR 161*, 1382–1392.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 56–65.

Wright, R. W. (2013). *The NESS account of natural causation: A response to criticisms* (pp. 13–66). De Gruyter.

Zhang, X., Solar-Lezama, A., & Singh R. (2018). Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in neural information processing systems* (pp. 4879–4890).