From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence

David Alvarez-Melis,[1] Harmanpreet Kaur,[2]
Hal Daumé III,[1, 3] Hanna Wallach,[1] Jennifer Wortman Vaughan[1]
[1] Microsoft Research
[2] University of Michigan
[3] University of Maryland
{alvarez.melis,hal3,hanna,jenn}@microsoft.com, harmank@umich.edu

Abstract

We take inspiration from the study of human explanation to inform the design and evaluation of interpretability methods in machine learning. First, we survey the literature on human explanation in philosophy, cognitive science, and the social sciences, and propose a list of design principles for machine-generated explanations that are meaningful to humans. Using the concept of weight of evidence from information theory, we develop a method for generating explanations that adhere to these principles. We show that this method can be adapted to handle high-dimensional, multi-class settings, yielding a flexible framework for generating explanations. We demonstrate that these explanations can be estimated accurately from finite samples and are robust to small perturbations of the inputs. We also evaluate our method through a qualitative user study with machine learning practitioners, where we observe that the resulting explanations are usable despite some participants struggling with background concepts like prior class probabilities. Finally, we conclude by surfacing design implications for interpretability tools in general.

Introduction

Interpretability has long been a desirable property of machine learning (ML) models. With the success of complex models like neural networks, and their expanding reach into high-stakes and decision-critical applications, explaining ML models' predictions has become even more important. Interpretability can enable model debugging and lead to more robust ML systems, support knowledge discovery, and boost trust (Hong, Hullman, and Bertini 2020). It can also help to mitigate unfairness by surfacing undesirable model behavior (Tan et al. 2018; Dodge et al. 2019), lead to increased accountability by enabling auditing (Selbst and Barocas 2018), and enable ML practitioners to better communicate model behavior to stakeholders (Veale, Van Kleek, and Binns 2018; Hong, Hullman, and Bertini 2020).

There are two primary techniques for achieving interpretability of ML models. The first is to train transparent, or *glass-box*, models that are intended to be inherently interpretable, such as decision trees (Quinlan 1986) and sets (Lakkaraju, Bach, and Jure 2016), simple point systems (Zeng, Ustun, and Rudin 2017; Jung et al. 2020), and generalized additive models (Hastie and Tibshirani 1990; Caruana et al. 2015). Although some researchers have argued that glass-box models should always be used in high-stakes scenarios (Rudin 2019), complex *black-box* models, such as neural networks, random forests, and ensemble methods, are very widely used in practice. As a result, other ML researchers have gravitated towards interpretability methods that generate post-hoc *local* explanations for individual predictions produced by such models (e.g., Simonyan, Vedaldi, and Zisserman 2013; Selvaraju et al. 2017; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Alvarez-Melis and Jaakkola 2017).

Local explanations aim to answer the question of why a model $\mathcal{M}$ predicted a particular output y for some input x. There are many ways of operationalizing this abstract question, but most methods do so by addressing the proxy question of how much the value of each input feature x_i contributed to the prediction y. Thus, in practice, the explanations generated by many such methods consist of importance scores indicating the positive or negative relevance of each input feature x_i. Although the way these scores are computed varies from method to method, most start from an axiomatic or algorithmic derivation of some notion of feature importance, and only later investigate whether the resulting explanations are useful to humans. Some methods forgo this last step altogether, relying exclusively on intrinsic evaluation of mathematical properties of explanations, such as robustness or faithfulness to the underlying model (Alvarez-Melis and Jaakkola 2018a,b; Jacovi and Goldberg 2020).

Interpretability, however, is fundamentally a human-centered concept. In light of this, we put human needs at the center of both the design and evaluation of interpretability methods. Our work builds upon and weaves together two literatures that study the relationship between humans and explanations. First, researchers in philosophy, cognitive science, and the social sciences have long studied what it means to explain, and how humans do it (e.g., Pitt 1988; Miller 2019b, and references therein). Second, a recent line of work within the human–computer interaction community has focused on how humans understand and utilize interpretability tools (i.e., software implementations of interpretability methods) (Lim and Dey 2011; Bunt, Lount, and Lauzon 2012; Bussone, Stumpf, and O'Sullivan 2015; Hohman et al. 2019; Lage et al. 2019; Abdul et al. 2020; Poursabzi-Sangdeh et al. 2021), where a common finding is that practitioners misunderstand, over-trust, and misuse these tools (e.g., Kaur et al. 2020).

Inspired by Miller (2019b), we start by surveying the literature on the nature of explanation, revealing recurring characteristics of human explanation that are often missing from interpretability methods. We distill these characteristics into design principles that we argue human-centered machine-generated explanations should satisfy. We then realize our design principles using the concept of *weight of evidence* (WoE) from information theory (Good 1985), which has recently been advocated for by Spiegelhalter (2018), but, to the best of our knowledge, has yet to be investigated in the context of interpretability. We demonstrate that WoE can be adapted to handle high-dimensional, multi-class settings, yielding a suitable theoretical foundation for interpretability. We provide a general, customizable meta-algorithm to generate explanations for black-box models. We also show experimentally that WoE can be estimated from finite samples and is robust to small perturbations of the inputs.

Evaluation of interpretability methods is notoriously difficult (Doshi-Velez and Kim 2017; Kaur et al. 2020). Although recent work has focused on abstract, intrinsic metrics such as robustness or faithfulness to the underlying model (Alvarez-Melis and Jaakkola 2018a), considerably less attention has been given to understanding how the resulting explanations are used in practice. This discrepancy between the intended use of the explanations — by a human, for a specific goal such as auditing, debugging, or building trust in a model — and their experimental evaluation — typically performed using abstract, intrinsic metrics, in generic settings — hampers understanding of the benefits and failure points of different interpretability methods.

We build on a recent thread of work (e.g., Lage et al. 2019; Nourani et al. 2019; Li et al. 2020; Kaur et al. 2020; Vaughan and Wallach 2021; Poursabzi-Sangdeh et al. 2021) — including several recent papers from the human computation community — that argues that evaluations should be grounded in concrete use cases and should put humans at the center, taking into account not only how they use interpretability tools, but how well they understand the principles behind them. We carry out an artifact-based interview study with ten ML practitioners to investigate their use of a tool implementing our meta-algorithm in the context of a practical task. Qualitative themes from this study suggest that most participants successfully used the tool to answer questions, despite struggling with background concepts like prior class probabilities. Although the study was designed to identify preferences for different tool modalities, participants often used all of them and requested the option to switch between them interactively. Our results additionally highlight the importance of providing well-designed tutorials for interpretability tools — even for experienced ML practitioners — which are often overlooked in the literature on interpretability methods, and which we argue should be an integral part of any interpretability tool.

Human-Centered Design Principles

What it means to explain and how humans do it have long been studied in philosophy, cognitive science, and the social sciences. We draw on this literature to propose human-centered design principles for interpretability methods.

Hempel and Oppenheim (1948) and van Fraassen (1988) define an explanation as consisting of two main pieces: the *explanandum*, a description of the phenomenon to be explained, and the *explanans*, the facts or propositions that explain the phenomenon, which may rely on relevant aspects of context. As is often done colloquially, we will refer to the explanans as the *explanation*. Different ways of formalizing the explanation have given rise to various theories, ranging from logical deterministic propositions (Hempel and Oppenheim 1948) to probabilistic ones (Salmon 1971; van Fraassen 1988). An excellent historical overview can be found in the surveys by Pitt (1988) and Miller (2019b).

In the context of local explanations for predictions made by ML models, the phenomenon to be explained is why a model $\mathcal{M}$ predicted output y for input x. This why-question can be operationalized in different ways. The facts used to explain this phenomenon may include information about the input features, the model parameters, the data used to train the model, or the manner in which the model was trained.

Although the nature of explanation is far from settled, recurring themes emerge across disciplines. At the core of the theories by van Fraassen (1988) and Lipton (1990) is the hypothesis that humans tend to explain in contrastive terms (e.g., "a fever is more consistent with the flu than with a cold"), with explanations that are both factual and counterfactual (e.g., "had the patient had chest pressure too, the diagnosis would instead have been bronchitis"). Yet, the explanations produced by most current interpretability methods refer only to why the input x points to a single hypothesis (i.e., the prediction y) rather than ruling out all alternatives.[1] In light of this, we propose our first two design principles:

1. **Explanations should be contrastive**, i.e., explicate why the model predicted y instead of alternative y'.
2. **Explanations should be exhaustive**, i.e., provide a justification for why every alternative y' was not predicted.

Another theme, featured prominently by Hempel (1962), is that human explanations decompose into simple components. In other words, humans usually explain using multiple simple accumulative statements, each addressing a few aspects of the evidence (e.g., "a fever rules out a cold in favor of bronchitis or pneumonia; among these, chills suggest the latter"). Each component is intended to be understood without further decomposition. Again, this contrasts with current interpretability methods that explain in one shot, for example, by providing importance scores for all features simultaneously. Our next two design principles are therefore:

3. **Explanations should be modular and compositional**, breaking up predictions into simple components.

[1] Exceptions include recent work advocating for contrastive or counterfactual explanations (Wachter, Mittelstadt, and Russell 2017; Miller 2019a; van der Waa et al. 2018), partly inspired by contrast sets (Azevedo 2010; Bay and Pazzani 1999; Webb, Butler, and Newlands 2003; Novak, Lavrač, and Webb 2009).

4. **Explanations should rely on easily-understandable quantities**, so that each component is understandable.

Another recurring theme is minimality. In a survey of over 250 papers, Miller (2019b) argued that it is important, but underappreciated in ML, that only the most relevant facts be included in explanations. Thus, our final principle is:

5. **Explanations should be parsimonious**, i.e., only the most relevant facts should be provided as components.

These design principles are not exhaustive; each could be refined or generalized, and other principles could be derived from the same literature. However, we posit that these principles provide a reasonable starting point because they capture some of the most apparent discrepancies between human and machine-generated explanations. More generally, these principles point to a broader theme of human explanations as a *process* rather than (only) a *product* (Miller 2019b; Lombrozo 2012). Therefore, these principles work to shift interpretability methods from the latter towards the former.

Explaining with the Weight of Evidence

The set of design principles proposed in the previous section outlines a framework for human-centered interpretability in ML. In this section, we show how this framework can be operationalized by means of the *weight of evidence*, a simple but powerful concept from information theory. We operationalize the question of why model $\mathcal{M}$ predicted output y for input x in terms of how much *evidence* each input feature x_i (or feature group) provides in favor of y relative to alternatives. An explanation based on this question adheres to our design principles because it is based on a familiar concept (evidence) that is grounded in common language, it naturally evokes a contrastive statement (evidence *for* or *against* something), and, as we explain below, it can be formalized using simple quantities that admit modularity.

Weight of Evidence: Foundations

The weight of evidence (WoE) is a well-studied probabilistic approach for analyzing variable importance that traces its origins back to Peirce (1878), but was popularized by Good (1950, 1968, 1985), whose definition and notation we follow here. Given a hypothesis and some evidence, the WoE seeks to answer the following question: *"How much does the evidence speak in favor of or against the hypothesis?"*

The WoE is usually defined for some evidence e, a hypothesis h, and its logical complement $\bar{h}$. For example, in a simple binary classification setting, $e = (X_1, \dots, X_n)$, $h : Y = 1$, and $\bar{h} : Y = 0$. The WoE of e in favor of h is the log-odds ratio between h conditioned on e and h marginally:

$$\mathrm{woe}(h : e) \triangleq \log \frac{O(h \mid e)}{O(h)}, \qquad (1)$$

where $O(\cdot)$ denotes the odds of a hypothesis, i.e.,

$$O(h) \triangleq \frac{P(h)}{P(\bar{h})} \qquad \text{and} \qquad O(h \mid e) \triangleq \frac{P(h \mid e)}{P(\bar{h} \mid e)}. \qquad (2)$$

Using Bayes' rule, $\mathrm{woe}(h : e)$ can also be defined as

$$\mathrm{woe}(h : e) \triangleq \log \frac{P(e \mid h)}{P(e \mid \bar{h})}. \qquad (3)$$

These two equivalent definitions provide complementary views of the WoE: the *hypothesis-odds* and *evidence-likelihood* interpretations. Using Equation (1), $\mathrm{woe}(h : e) > 0$ indicates that the odds of h are higher under e than marginally. Equivalently, using Equation (3), it indicates that the likelihood of e is larger when conditioning on h than on its complement. In other words, the evidence *speaks in favor of* hypothesis h. Analogously, if $\mathrm{woe}(h : e) < 0$ we would say that the evidence *speaks against* h. The quantities in Equations (1) and (3) are contrastive (cf. Principle 1) — that is, defined in terms of ratios.

As a concrete example, suppose that a doctor wants to know whether a patient's symptoms indicate the presence of a certain disease, say, the flu. Denote $e =$ "*the patient has a fever*," $h =$ "*the patient has the flu*," and $\bar{h} =$ "*the patient doesn't have the flu*." The doctor might know that for a patient, the odds of having the flu roughly double once the patient's fever is taken into account (i.e., the hypothesis-odds interpretation), which corresponds to $\mathrm{woe}(h : e) \approx \log 2$. Alternatively, using the evidence-likelihood interpretation, the doctor could conclude that a patient is twice as likely to have a fever if they have the flu compared to when they do not. Note that neither interpretation tells us anything about the base rate of the flu.

The WoE generalizes beyond these simple scenarios. For example, it can be conditioned on additional information c:

$$\mathrm{woe}(h : e \mid c) \triangleq \log \frac{P(e \mid h, c)}{P(e \mid \bar{h}, c)}.$$

It can also contrast h to an arbitrary alternative hypothesis h' instead of $\bar{h}$ (e.g., evidence in favor of the flu and against a cold): $\mathrm{woe}(h/h' : e) \triangleq \mathrm{woe}(h : e \mid h \vee h')$. Thus, we can, in general, talk about the strength of evidence in favor of h and against h' provided by e (perhaps conditioned on c).

When the evidence is decomposable into multiple parts — that is, when $e = \bigcup_{i=1}^{n} e_i$ — the WoE is also decomposable:

$$\mathrm{woe}(h/h' : e) = \sum_{i=1}^{n} \log \frac{P(e_i \mid e_{i-1}, \dots, e_1, h)}{P(e_i \mid e_{i-1}, \dots, e_1, h')}. \qquad (4)$$

This is crucial to defining an extension of the WoE to high-dimensional inputs that adheres to Principle 3 (modularity).

A further appealing aspect of the WoE is its immediate connection to Bayes' rule through the following identity:

$$\underbrace{\log \frac{P(h \mid e)}{P(h' \mid e)}}_{\text{Posterior log odds}} = \underbrace{\log \frac{P(h)}{P(h')}}_{\text{Prior log odds}} + \underbrace{\log \frac{P(e \mid h)}{P(e \mid h')}}_{\text{Weight of evidence}}. \qquad (5)$$

In other words, the WoE can be understood as an adjustment to the prior log odds caused by observing the evidence. In a simple binary classification setting, this amounts to

$$\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} = \log \frac{P(Y=1)}{P(Y=0)} + \mathrm{woe}(Y = 1 : X),$$

which shows that a positive (respectively, negative) WoE implies that the posterior log odds of $Y = 1$ versus $Y = 0$ are higher (lower) than the prior log odds, indicating that the evidence makes $Y = 1$ more (less) likely than it was *a priori*.

Equation (5) shows that the WoE is modular (cf. Principle 3) in another important way: it disentangles prior class probabilities and input likelihoods. This is important because of the *base rate fallacy* studied in the behavioral science literature (Tversky and Kahneman 1974; Bar-Hillel 1980; Eddy 1982; Koehler 1996). This cognitive bias, prevalent even among domain experts, is characterized by a frequent misinterpretation of posterior probabilities, primarily caused by a neglect of base rates (i.e., prior probabilities). Despite this, many interpretability methods do not explicitly display prior probabilities, and even when they do, they focus on explaining posterior probabilities, which invariably entangle information about priors and the input being explained.

Additionally, the units in which the WoE is expressed (log-odds ratios) are arguably easily understandable (cf. Principle 4). There is evidence from the cognitive-neuroscience literature that log odds are a natural unit in human cognition. For example, degrees of confidence expressed by humans are proportional to log odds (Peirce and Jastrow 1885), people are less biased when responding in log odds that in linear scales (Phillips and Edwards 1966), and there exist plausible neurological hypotheses for encoding of log odds in the human brain (Gold and Shadlen 2001, 2002). We refer the reader to Zhang and Maloney (2012) for a meta-analysis of these various studies of log odds.

We provide additional properties of the WoE, along with an axiomatic derivation, in the appendix, which can be found in the longer version of this paper, available online.[2]

Composite Hypotheses and Evidence

Traditionally, the WoE has been mostly used in simple settings, such as a single binary output and only a few input features. Its use in the more complex settings typically considered in modern ML therefore poses new challenges.

The first such challenge is that in multi-class settings, there is flexibility in choosing the hypotheses h and h' to contrast. The obvious choice of letting h correspond to the predicted class y^* and h' its complement is unlikely to yield useful explanations when the number of classes is large (e.g., explaining the evidence in favor of one disease against one hundred thousand others). Following Principle 3 (modularity), and taking inspiration from Hempel's model (1962) and the view of explanation as a process (Lombrozo 2012; Miller 2019b), we address this by casting explanation as a sequential procedure, whereby a subset of the possible classes is ruled out at each step. For example, in medical diagnosis, we might first explain why bacterial diseases were ruled out in favor of viral ones, and then explain why a specific viral disease was predicted instead of the others. In general, for a classification problem over labels $\mathsf{Y} = \{1, \dots, k\}$, we will consider a (given or constructed) nested partitioning of Y into a sequence of T-many subsets U_i of classes such that $\{y^*\} \triangleq \mathsf{U}_T \subset \mathsf{U}_{T-1} \subset \cdots \subset \mathsf{U}_0 \triangleq \mathsf{Y}$. As we show in Figure 4 in the appendix, this partition implies a sequence of pairs of hypotheses $(h_t, h'_t) = (y \in \mathsf{U}_t, y \in \mathsf{U}_{t-1} \setminus \mathsf{U}_t)$.

Algorithm 1 WoE meta-algorithm for complex models

```
 1: Input: Instance $X \in \mathbb{R}^n$, prediction $y^* \in \{1, \dots, k\}$
 2: Parameters: Features $\mathcal{A} = \{\{1\}, \dots, \{n\}\}$ or feature
    groups $\mathcal{A} = \{S_1, \dots, S_m\}$
 3: Initialize $\mathsf{U}_0 \leftarrow \{1, \dots, k\}$
 4: $t \leftarrow 0$
 5: while $|\mathsf{U}_t| > 1$ do
 6:     $t \leftarrow t + 1$
 7:     $\mathsf{U}_t \leftarrow$ SelectHypothesis$(\mathsf{U}_{t-1}, y^*)$
 8:     $\overline{\mathsf{U}}_t \leftarrow \mathsf{U}_{t-1} \setminus \mathsf{U}_t$ {relative complement}
 9:     $\pi(\mathsf{U}_t) \leftarrow \log \frac{P(y \in \mathsf{U}_t)}{P(y \in \overline{\mathsf{U}}_t)}$ {prior log odds}
10:     for $i = 1, \dots, |\mathcal{A}|$ do
11:         $\omega_i^t \leftarrow \mathrm{woe}(y \in \mathsf{U}_t / y \in \overline{\mathsf{U}}_t : X_{\mathcal{A}_i} \mid X_{\mathcal{A}_{i-1}}, \dots, X_{\mathcal{A}_1})$
12:     end for
13:     $\Omega_t \leftarrow \sum_{i=1}^{|\mathcal{A}|} \omega_i^t$
14:     DisplayExplanation$(\mathsf{U}_t, \overline{\mathsf{U}}_t, \mathcal{A}, \pi(\mathsf{U}_t), \{\omega_i^t\}_i, \Omega_t)$
15: end while
```

A second challenge arises when the the number of input features is large. For very high-dimensional inputs (such as images or detailed health records), providing a WoE value for each feature will rarely be informative. Again, imagine our hypothetical doctor having to simultaneously analyze the relevance of thousands of symptoms. For such cases, we propose aggregating the input features into feature groups (e.g., super-pixels for images or groups of related symptoms for medical diagnosis). Formally, for an input X of dimension n, we partition the feature indices into m disjoint subsets, with $\mathsf{S}_1 \cup \cdots \cup \mathsf{S}_m = \{1, \dots, n\}$. Equation (4) (or, equivalently, the chain rule of probability) allows for arbitrary groupings, so for any such partition we can compute

$$\mathrm{woe}(h/h' : X) = \sum_{i=1}^{m} \underbrace{\log \frac{P(X_{\mathsf{S}_i} \mid X_{\mathsf{S}_{i-1}}, \dots, X_{\mathsf{S}_1}, h)}{P(X_{\mathsf{S}_i} \mid X_{\mathsf{S}_{i-1}}, \dots, X_{\mathsf{S}_1}, h')}}_{= \mathrm{woe}(h/h' : X_{\mathsf{S}_i} \mid X_{\mathsf{S}_{i-1}}, \dots, X_{\mathsf{S}_1})} \tag{6}$$

where $X_{\mathsf{S}_i} = \{X_j\}_{j \in \mathsf{S}_i}$ is the ith feature group, or "atom."

A Meta-Algorithm for WoE Explanations

Using these extensions of the WoE, we propose a meta-algorithm for generating explanations for complex classifiers (Algorithm 1). Given a model, an input, and a prediction, the algorithm generates an explanation for the prediction sequentially by producing WoE values for progressively smaller nested hypotheses. Specifically, at every step t, a subset of classes $\mathsf{U}_t \subset \mathsf{U}_{t-1}$ is selected and the remaining classes $\overline{\mathsf{U}}_t = \mathsf{U}_{t-1} \setminus \mathsf{U}_t$ are ruled out. The user is shown a comparison of hypotheses $h_t : y \in \mathsf{U}_t$ and $h'_t : y \in \overline{\mathsf{U}}_t$ consisting of both their prior log odds $\pi(\mathsf{U}_t)$ (line 9) and the WoE in favor of h_t and against h'_t. WoE values are computed sequentially with each atom $\mathcal{A}_i$ (either an individual input feature or a group of features) as the evidence (line 11) and these values are summed to obtain the total WoE using the additive property (line 13). These values are presented to the user, and the process continues until all classes except the prediction y^* have been ruled out (cf. Principles 2–3).

Left unspecified in this meta-algorithm are four key choices that are application-dependent and require fur-

[2] https://arxiv.org/abs/2104.13299

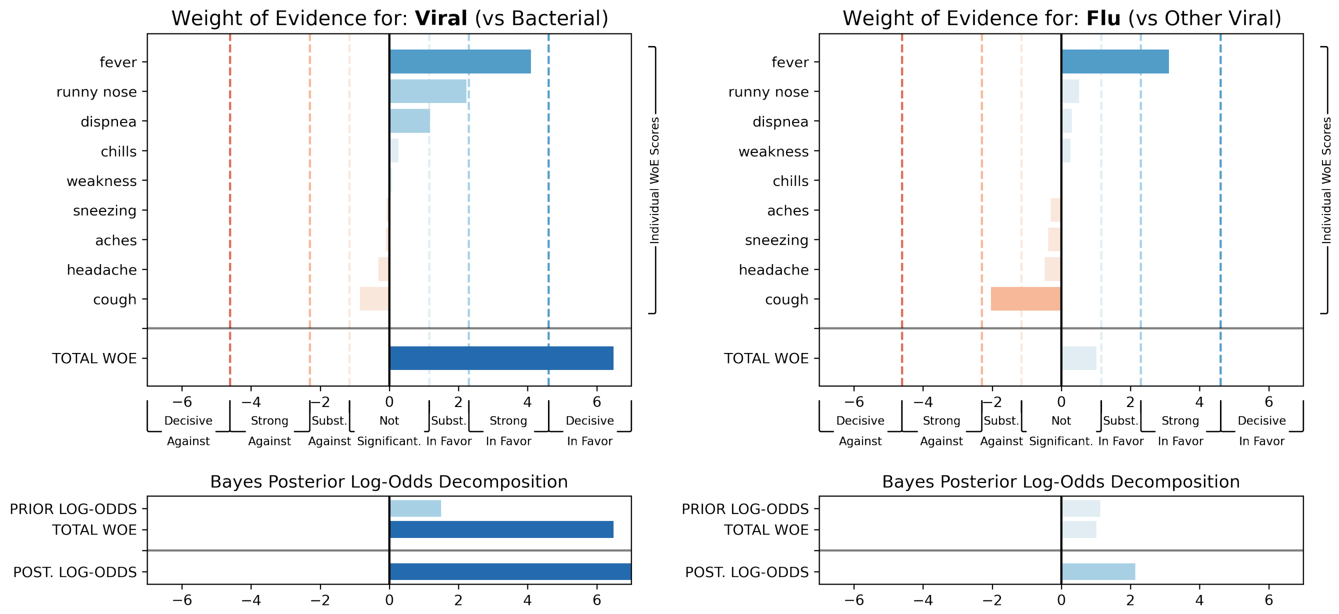

Figure 1: Example two-step explanation produced by our method for a model that predicted the flu for some input. The first step explains, using the weight of evidence, why the model favors viral diseases (instead of bacterial ones) for this input. Then, the second step explains why the model predicted the flu instead of the remaining possible classes (i.e., other viral diseases).

ther discussion. First is the question of how to define the SELECTHYPOTHESIS method to progressively partition the classes (line 7). If there is an inherent natural partitioning (e.g., "viral" versus "bacterial," as in the example discussed previously), then SELECTHYPOTHESIS simply amounts to retrieving the largest subset in the partition containing the prediction y^*. For the general case, we propose selecting the hypothesis that maximizes a WoE-based objective:

$$\mathsf{U}_t \leftarrow \operatorname*{argmax}_{\mathsf{U}\subset \mathsf{U}_{t-1};\, y^*\in \mathsf{U}} \mathrm{woe}(y \in \mathsf{U} \,/\, (y \in \mathsf{U}_{t-1}\setminus \mathsf{U}) : X) - R(\mathsf{U}),$$

where R is a cardinality-based regularizer. R should be chosen to penalize sets that are too small (which would yield granular explanations with many steps, in opposition to Principle 5) or too large (which would yield coarse explanations, to the detriment of Principle 3). Although the choice of R should ideally be informed by the application and the user, a sensible generic choice is $R(\mathsf{U}) \propto \big||\mathsf{U}| - \frac{1}{2}|\mathsf{U}_{t-1}|\big|^p$, normalized so that $R(\mathsf{U}) \in [0,1]$. Using this regularizer, Algorithm 1 approximately splits the remaining classes in half at every step, yielding roughly $O(\log k)$ steps in total.

Second, it should be noted that lines 10–12 in Algorithm 1 implicitly assume an ordering of the atoms, and that this ordering might affect the WoE values. In some applications, there might be a conditional independence structure known a priori that could inform the choice of atoms and their ordering (e.g., those simplifying the conditioning in line 11 the most). If not, the ordering can again be chosen randomly or based on the sorted per-atom conditional WoE values.

Third, computing the per-atom WoE (line 11) requires the conditional likelihoods $P(X_{\mathcal{A}_i} | X_{\mathcal{A}_{i-1}}, \dots, X_{\mathcal{A}_1}, Y)$. Ideally, the model would compute these likelihoods internally. If, instead, it computes only marginal feature likelihoods $P(X_{\mathcal{A}_i}|Y)$, we can use a naïve Bayes (NB) approximation — that is, use these in place of the conditional likelihoods in Equation (6). If the model is a black box or does not compute conditional likelihoods internally, then these must be estimated as we explain below.

Finally, there is the question of how to implement DISPLAYEXPLANATION. When the number of atoms is large, the WoE values for only the most salient atoms can be displayed (cf. Principle 5) — e.g., those with absolute WoE larger than a given threshold τ; Good (1985) suggests $\tau = 2$ as a rule of thumb. Otherwise, all per-atom WoE values can be displayed along with the total WoE and prior log odds.

An example two-step explanation produced by our method (on fabricated data from our user study tutorial) is shown in Figure 1. The sorting and color coding of the features by their WoE values makes it apparent which of these contribute the most evidence in favor or against the selected class (or set of classes), and the labeling along the x-axis provides guidelines for context. The visualization suggests the additive nature of these values (i.e., that stacking blue bars and subtracting red ones yields the total WoE). The bottom panel, a graphical representation of Equation (5), disentangles the model's estimated prior class odds (which *a priori* weakly favor the flu and other viral diseases), from its total WoE (very strong when contrasting viral and bacterial diseases, less so for the flu versus other viral diseases).

WoE Estimation for Black-box Models

As we noted above, computing the per-atom WoE (line 11 in Algorithm 1) requires evaluating the conditional likelihoods $P(X_{\mathcal{A}_i} | X_{\mathcal{A}_{i-1}}, \dots, X_{\mathcal{A}_1}, Y)$. In many practical settings, including those in which the model is a black box, these conditional likelihoods are not computed internally, so they must be estimated. In such settings, we propose fitting a conditional likelihood estimation model using the model's predictions $\hat{y} = \mathcal{M}(x)$ (not the true labels y) as a preliminary step. This conditional likelihood estimation model can then be called on demand when computing the per-atom WoE.

In some settings, it may be possible to fit a full (i.e., conditioned on *both* the class Y and all previous atoms $A_i - 1, \ldots, A_1$) conditional likelihood estimation model, for example, via kernel or spectral density estimation methods when working with low-dimensional data, or via autoregressive or recurrent neural networks when working with text or time-series data. For more complex types of data, such as images, methods based on normalizing flows and neural autoregressive models (e.g., Rezende and Mohamed 2015; Papamakarios, Pavlakou, and Murray 2017) are likely to be more appropriate. In settings where fitting a full conditional likelihood estimation model is infeasible, an NB approximation can be used to estimate class- (but not atom-) conditional likelihoods, for example, via a Gaussian NB classifier.

We emphasize that fitting a conditional likelihood estimation model — the main computational bottleneck of our method — must be done only once, potentially offline. This is in contrast to perturbation-based interpretability methods like LIME that fit a new model for every prediction.

We assess the quality of finite-sample WoE estimation experimentally in the "Quantitative Experiments" section.

Relation to LIME and SHAP

When viewed from a probabilistic perspective, most post-hoc interpretability methods revolve around a model's predictive posterior — that is, they seek explanations that deconstruct $P(Y = y^* \mid X)$ in various ways. For example, LIME (Ribeiro, Singh, and Guestrin 2016) seeks to approximate $f(x) = P(Y \mid X = x)$ in the vicinity of x_0 through a simpler, interpretable surrogate model $\tilde{f}(x)$. Similarly, SHAP (Lundberg and Lee 2017) quantifies variable importance by analyzing the effect on the posterior of "dropping" variables X_i from the input X. In contrast, the WoE focuses — directly, in the case of the evidence-likelihood interpretation and indirectly in the case of the hypothesis-odds interpretation — on the conditional likelihood $P(X = x \mid Y)$. In other words, for a given input x, a WoE explanation is based on the probability assigned by the model to x (or a subset of its features) given, e.g., $Y = y^*$.

When viewed in this way, the relationship between WoE explanations and other post-hoc interpretability methods like LIME and SHAP is akin to the relationship between generative and discriminative models. Indeed, as is the case for some pairs of generative and discriminative models (e.g., naïve Bayes and logistic regression), these different interpretability methods also turn out to be equivalent — two sides of the same coin — for some simple classifiers, as we show in the appendix for logistic regression. However, this is not generally the case. Moreover, even when the explanations generated by different interpretability methods qualitatively agree (i.e., the same features are highlighted as being important), the specific interpretations of the explanations will differ. Indeed, the WoE uses a different operationalization of the notion of feature importance, in turn entailing different units of explanation: log likelihoods and log-odds ratios in the case of the WoE and linear attribution scores for posterior probabilities in the case of LIME and SHAP.

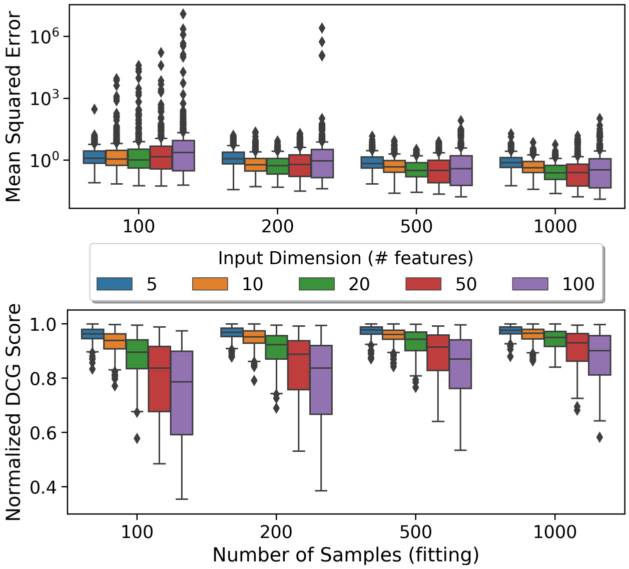

Figure 2: Quality of WoE estimation. Top: MSE. Bottom: NDCG, ranging from 0 (worst) to 1 (perfect) ranking quality.

Quantitative Experiments

Here, we assess the quality of finite-sample WoE estimation and the robustness of the WoE to perturbations of the inputs.

Quality of Finite-Sample WoE Estimation

Our first experiment evaluates the quality of WoE estimates from finite samples. As we explained above, such estimates are needed if the model is a black box or does not compute likelihoods internally. For evaluation purposes, we consider a model that, by construction, computes all the quantities required for exact WoE computation, but treat it as black box — that is, its internal WoE computation will be used only for evaluation, and is not available to our method. Instead, we separately fit a conditional likelihood estimation model by querying the model for a small number of inputs, and use this estimation model to compute WoE values at explanation time. We control for model misspecification by having both the model and our conditional likelihood estimation model use the NB assumption. Specifically, we use a smoothed Gaussian NB (GNB) classifier. This allows us to focus on the quality of WoE estimation from finite samples, but does not address model misspecification.

First, we generate a dataset of a given dimension. We train a model on a subset of this dataset of size $N_{\text{train}} = 1000$ and fit the conditional likelihood estimation model on a separate subset of size N_{fit}, which we vary. For every test input x_i ($N_{\text{test}} = 10$), we compute true WoE values for each input feature using the model's prior and posterior probabilities,

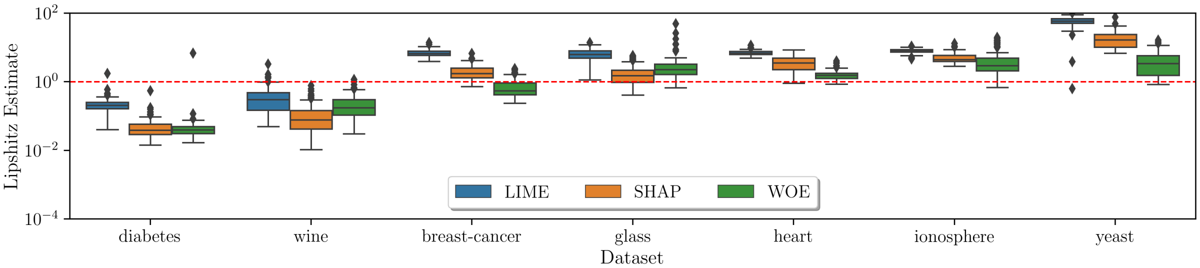

Figure 3: Explanation robustness across benchmark datasets. Extreme values away from $L = 1$ (dashed line) are undesirable.

and then compute estimated WoE values according to the estimation model in the appendix. We compare these using two metrics: mean squared error and normalized discounted cumulative gain (NDCG), a measure of ranking quality that might be relevant to practitioners, applied to the relative ranking of input features by their (true or estimated) WoE.[3]

Figure 2 shows these metrics as a function of input dimension and sample size N_{fit}. As expected, estimation quality improves with the number of samples used for fitting, and degrades gracefully as the input dimension increases. These results suggest that the WoE can be accurately estimated — even in relatively high dimensions — from finite samples, although we caution that these results may look different for other models, and we do not measure model misspecification error due to the NB approximation.

Robustness of WoE

Previous work has argued that interpretability methods should be robust in the sense that the explanations they provide should not vary dramatically when the input whose prediction is being explained changes by a small amount. To investigate the robustness of our method, we follow the setup of Alvarez-Melis and Jaakkola (2018a). Letting $\mathcal{E}(\cdot)$ be a function that maps feature vectors $x \in \mathbb{R}^n$ to explanation vectors (e.g., importance scores) $e \in \mathbb{R}^n$, we quantify its robustness around x_0 through its local Lipschitz constant:

$$L(x_0) = \max_{x_j \in \mathcal{B}_\varepsilon(x_0)} \frac{\|\mathcal{E}(x_j) - \mathcal{E}(x_0)\|}{\|x_j - x_0\|}, \qquad (7)$$

where $\mathcal{B}_\varepsilon(x_0) = \{x \mid \|x - x_0\| \leq \varepsilon\}$. Intuitively, $L(x_0)$ quantifies the largest relative change in importance scores in a small neighborhood around x_0. Extreme values are usually undesirable, as they indicate explanations that are either too sensitive (large L) or not responsive enough (very small L) to changes in the input features. In most settings, values below 1 but bounded away from 0 are preferable.

Concretely, we first train a GNB classifier. Then, for any input x we use the classifier to generate a prediction y, and input both of these to our method to generate $\mathcal{E}(x)$, a vector of WoE values for each feature x_i. Since computing the robustness metric (7) involves maximization, we estimate this quantity from finite samples using Bayesian optimization, making repeated calls to $\mathcal{E}(\cdot)$. We focus on standard benchmark classification datasets from the UCI repository (Dua and Graff 2017), and compare our method to LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017). Figure 3 shows the results for ten repetitions with different random seeds for each dataset and interpretability method pair. The red dashed line indicates the bound $L(x) = 1$. For all but one dataset, WoE explanations are, on average, as close or closer to the ideal Lipschitz robustness as explanations generated by LIME and SHAP.

User Study

Throughout this paper, we have argued that interpretability is fundamentally a human-centered concept and that the evaluation of interpretability methods should therefore focus on the needs of humans, exploring how they use interpretability tools, as well as their understanding of the concepts that underlie them. In this section, we present a user study that we carried out to assess the usefulness of our method in the types of scenarios in which it would be used in practice. Such studies are commonly used in the HCI community to distinguish between designers' intended use of a tool and users' mental models (Gibson 1977; Norman 2013).

Study Design

We conducted an artifact-based interview study with 10 ML practitioners to evaluate the use of our interpretability method, implemented in a simple tool (the artifact), in a controlled setting. In such qualitative studies, the goal is to ensure sufficient interaction time and nuanced data collection for each participant, which is typically only feasible for relatively small sample sizes (Hudson and Mankoff 2014; Olsen Jr 2007; Turner, Lewis, and Nielsen 2006). Our study followed a think-aloud protocol, in which participants were asked to verbalize their thought processes as they used the tool to perform specific tasks, in order to help us identify specific concepts and functionalities that might be confusing. We placed participants in a controlled setting (rather than observe them using the tool on their own models) because of challenges including data access, inconsistencies in the types of data analyzed by the participants, and the potential difficulty of establishing patterns across settings.

Our study consisted of two main parts: a tutorial intended to introduce the concepts and functionalities needed to use our tool, followed by the main study in which participants answered questions about a pre-trained ML model using the tool. We also conducted pre-study interviews to establish participants' backgrounds in ML and post-study interviews

[3] The NDCG is only defined for positive values, so we compute it separately for positive and negative values and average them.

in which participants reflected on their experiences with the tool and how they might use it in their ML pipelines. The study design was approved by our internal institutional review board. Excerpts (screenshots) from the Jupyter notebooks that we used in the tutorial and in the main study are in the appendix; the complete notebooks are available online.[4]

Tutorial Kaur et al. (2020) found that practitioners often use interpretability tools without fully understanding them, highlighting the importance of providing well-designed tutorials and other accompanying documentation. For our user study, we designed a tutorial to introduce the concepts and functionalities needed to use our tool, to evaluate participants' understanding of these concepts, and to check whether their responses in the main study were likely based on a sound understanding — without being too time-consuming or tedious. After several iterations and pilot studies, we converged on an approximately 40-minute-long tutorial based on a Jupyter notebook containing equations, text, and images. This tutorial covered log-odds ratios, weight of evidence, feature group for high-dimensional inputs, and sequential explanations for multi-class settings.

Main Study The goal of the main study was to assess participants' understanding of the WoE and to investigate their use of our interpretability tool in the context of a realistic ML task. Participants were given a Jupyter notebook that included a dataset, an ML model trained using the dataset, and our tool. They were then asked to answer several questions with the help of outputs (i.e., visualizations) from our tool.

The ML model was a random forest classifier trained using the Online News popularity dataset (Fernandes, Vinagre, and Cortez 2015), which consists of 39,797 news articles. Each article is represented using 59 features that capture metadata about the article, such as its length, any links, and its sentiment polarity. We trained the model to predict the category that the article was published under (e.g., "Lifestyle" or "Business"), creating a six-class classification task. We chose this dataset and this task because the domain is understandable without expert knowledge or prior experience, the number of classes is large enough to permit meaningful sequential explanations, and there are enough features to make explanations based on feature groups sufficiently different from explanations based on individual features.

The main study was itself divided into two parts, which were designed to let us observe the use of the two extensions of the WoE described in the "Composite Hypotheses and Evidence" section: feature groups and sequential explanations. In the first part, participants were given the option to view explanations based on feature groups or explanations based on individual features, and were asked questions that could be answered using either type of explanation (e.g., "What aspects of the news article contributed the most to this prediction?"). This part of the study was intended to surface participants' preferences. In the second part, participants were given the option to generate one-shot or sequential explanations, and were asked questions that could only be precisely answered using sequential explanations (e.g., "Why didn't the model predict [subset of classes]?"). This part of the study was intended to assess whether participants could successfully use sequential explanations.

Participants Potential participants were recruited via email. To be considered for the study, they were asked to complete a survey about their ML background and their experience with interpretability tools. Of 41 survey respondents, we randomly selected 10 to participate in the study. All participants were ML practitioners (e.g., data scientists) with 1–20 years of experience. On average, participants rated the role of ML in their jobs as 6.7 and their experience with interpretability tools as 3.2, both on a scale of 0 ("not at all") to 7 ("extremely"). Participants also rated their familiarity with concepts from probability relevant to the WoE on a scale of 0 to 7. Their average ratings were 2.7 for posterior class probabilities, 3.3 for log likelihoods, 3.2 for log-odds ratios, and 0.9 for the WoE. On average, participants took 1.7 hours to complete the study. Each participant was compensated with a $40 Amazon gift card.

Methods Participants' open-ended answers were scored by comparing them to an answer key prepared in advance by two of the authors. Answers that correctly identified key aspects (e.g., a feature with a large positive WoE value, pushing the model toward a particular prediction) were treated as correct even if the participants' specific language did not exactly match the language in the answer key. To examine patterns of tool use, the usability of our tool, participants' interpretability needs, and participants' general impressions, we analyzed automatically generated audio transcripts for high-level themes using inductive thematic analysis (Braun and Clarke 2012) and affinity diagramming.

Results

The results from our user study divide naturally into three categories: participants' understanding of relevant concepts, tool usability and participants' preferences, and general needs for interpretability tools. First, the pre-study interviews and answers to the checkpoint questions in the tutorial provided insight into participants' understanding of concepts relevant to the WoE. Second, participants' approaches to the questions in the main study and their patterns of tool use enabled us to examine the usability of our tool. Finally, via the main study and the post-study interviews, we were able to uncover participants' general interpretability needs and additional criteria (beyond our design principles) to consider when designing and evaluating interpretability tools.

Understanding of Relevant Concepts Our analysis showed that most participants (7/10) struggled to understand and use prior class probabilities in the tutorial. The section on this topic was time-consuming: on average, participants spent a third of their tutorial time on this section. Eventually, most participants either ignored the prior class probabilities or used them incorrectly, supporting the base rate fallacy. Nonetheless, participants were able to use the WoE to correctly answer questions in the main study for which prior class probabilities were relevant. This raises the possibility that although they struggled with the abstract concept,

[4] http://github.com/dmelis/interpretwoe

they were able to use the information indirectly (e.g., via displayed class probabilities). This finding is consistent with those of Kaur et al. (2020), who showed that data scientists struggle to explain the concepts underlying the explanations produced by generalized additive models (Hastie and Tibshirani 1990; Caruana et al. 2015) and SHAP (Lundberg and Lee 2017), even though they still find these tools useful.

Although participants generally understood the concept of WoE, some confused negative WoE values with negative values for input features, thus finding it challenging to make sense of the explanations. As a result, two participants provided incorrect answers for the questions in the main study.

Tool Usability and Participants' Preferences Participants had no overwhelming preference between explanations based on feature groups and explanations based on individual features. Indeed, they noted that the two levels of granularity provide complementary information, and switching between the two options was a clear pattern across all participants. Although feature groups provide a high-level overview, making it "easier to manage [reading the plot]...[and the] direction of analysis is a lot clearer" (P8), explanations based on individual features help participants in "looking into more details in general...to know exactly which feature it was [that was responsible for a prediction]" (P10). We observed some differences in behavior based on participants' roles and expertise, though of course these are inconclusive with our small sample size. Participants with more ML experience tended to rely on feature-level plots, while those with customer-facing jobs more often provided high-level answers based on feature groups, noting that feature groups "provide customer-friendly explanations" (P8).

Participants found sequential explanations to be a helpful breakdown of a larger explanation into parts. P3 noted these were like a "story of how the prediction was made." Sequential explanations prompted more detailed answers to our questions and most participants (7/10) accurately answered questions in the second part of the main study using sequential explanations. They explained that the type of questions — which required understanding how each of the classes were ruled out — could not be answered via one-shot explanations. P8 commented, "I find this to be quite helpful... I guess without this breaking it down to this point I wouldn't have thought twice really about this [input feature group] being a [differentiating] factor between the two [output classes]...I think that this would be a nice like final understanding [of the predicted class], this goes a lot deeper than I probably could have gone just looking at that without the tool. So I think it was very helpful in that case."

Although most participants were happy with the level of detail presented, some participants with more ML experience expressed a desire for deeper understanding of how the explanations were generated. They understood the underlying concepts, but were wary of anything that appeared automated, including the breakdown of class comparisons in sequential explanations (which was automated) and the feature groups (which were actually manually generated).

Even though participants said that the tool helped them understand the model's predictions, not all of them envisioned the tool being added to their ML pipelines. Participants with significant prior ML experience already had established ways of ensuring that model predictions are reasonable, but recognized other exciting use cases for the tool, such as communicating complex predictions to less experienced end users. Particularly for high-risk domains, visualizations from the tool could help users probe odd predictions. P7 noted, "With my focus on medical data, I do see the need in working with a customer...there this [tool] would be a must-have. My team, we are engaged with customers and we have to educate them fast... So for me model interpretability there comes very close side by side with fairness."

General Needs for Interpretability Tools Most interpretability tools, including ours, rely on tutorials and other accompanying documentation to provide an introduction to the tool's concepts and functionalities. All participants appreciated the information presented in our tutorial: "I can't imagine doing this [study] without the tutorial. I generally know a lot more about these concepts now" (P5). The tutorial seemed to impact participants' overall accuracy in answering the questions in the main study — those who spent longer on the tutorial tended to provide more accurate and more thoughtful answers. This manifested as longer time spent on exploring the tool in the tutorial and ensuring that their answers to the checkpoint questions were accurate and thorough. Even participants with less ML experience provided accurate answers when they devoted time to the tutorial. Participants appreciated the example in the tutorial and were able to generalize from this example to the questions the main study: "The tutorial...helps you start in the right place. I went back to the example in the tutorial to [determine how to] answer questions in the study" (P6).

Finally, participants expressed a desire to be able to more easily switch between different options (e.g., input features versus feature groups) rather than re-running code. Interactivity was consequently the most commonly requested functionality in the post-study interview. This is in line with prior work on human-centered design principles for ML (Amershi et al. 2019; Hohman et al. 2019; Weld and Bansal 2019).

Limitations

All interpretability methods, including ours, involve various design choices and assumptions (both implicit and explicit), many of which give rise to potential limitations. First, the concept of interpretability is notoriously ambiguous, and unlike supervised ML tasks, there is no ground truth to use for evaluation, even for proxy concepts like feature importance. As a result, different interpretability methods assume different notions of interpretability, propose different quantities to operationalize them, and (when needed) rely on different techniques to estimate them. In turn, these choices mean that no interpretability method will ever be universally ideal. Moreover, summarizing the behavior of complex models comes at a price (Rudin 2019) — that is, the explanations are partial, only hold in a small neighborhood (Ribeiro, Singh, and Guestrin 2016), or make strong assumptions about the data (Lundberg and Lee 2017). As a result, explanations generated by one interpretability method seldom strictly dominate explanations generated by another. Furthermore, different explanations might reveal information about different aspects of the underlying model's behavior.

Under this perspective, the design choices and assumptions involved in our interpretability method necessarily limit its scope and applicability. Starting from our decision to distill characteristics of human explanation into human-centered design principles, our method assumes that human characteristics are desirable for machine-generated explanations. And, although the specific characteristics that we focus on yield a coherent set of design principles, these principles are not exhaustive or universal. Some may not be necessary in all settings, and all are open to refinement.

By relying on the concept of weight of evidence (WoE) from information theory, our method inherits many of its strengths and limitations. Concretely, there are three main settings in which there is a clear case for using explanations based on the WoE: 1) when the underlying model is generative, 2) when the underlying model is log linear, and 3) when the underlying model is a multi-class classifier. We provide a detailed discussion of these three settings in the appendix. In terms of limitations, the WoE requires access to the conditional likelihoods $P(X_{\mathcal{A}_i}|X_{\mathcal{A}_{i-1}}, \dots, X_{\mathcal{A}_1}, Y)$, which limits its use to settings in which these are accessible or can be accurately estimated from finite samples. Estimating densities for more complex types of data, such as images, is an active area of research, and although it may be possible to integrate new advances into our method, its applicability to such types of data is currently limited. Other important design choices involved in our method include the technique for partitioning the classes in multi-class settings to yield sequential explanations and the technique for partitioning input features into feature groups. Although we chose generic solutions to these challenges, there are other techniques, which may be more appropriate in some settings, which will invariably lead to different explanations. We defer a thorough investigation of these choices for future work.

The main limitations of our user study are the number of participants, the type of participants, and the extent to which the study conditions mimic a realistic setting. We chose to conduct an artifact-based interview study to ensure sufficient interaction time and nuanced data collection for each participant, but this limited the number of participants that we could consider, thereby precluding statistical analyses. Our participants were also limited to ML practitioners. Following previous work (e.g., Kaur et al. 2020), we chose to focus on ML practitioners because they are frequent users of interpretability tools in the wild. Finally, although we tried to design the study so as to mimic a realistic setting, we cannot be sure that this experience was representative of participants' day-to-day experiences (e.g., working with their own datasets). Ideally, we would have run a longitudinal field study with multiple types of participants to enable us to observe participants' tool use over time as they gain expertise in using it. However, this would have required additional resources (e.g., to support multiple types of data) and was therefore infeasible. Instead, our user study serves as an initial evaluation of our interpetability method.

Discussion

In this paper, we take inspiration from the study of human explanation, drawing on the literature on human explanation in philosophy, cognitive science, and the social sciences to propose a list of design principles for machine-generated explanations that are meaningful to humans. We develop a method for generating explanations that adhere to these principles using the concept of weight of evidence from information theory. We show that this method can be adapted to meet the needs of modern ML — that is, high-dimensional, multi-class settings — and that the explanations can be estimated accurately from finite samples, are robust to small perturbations of the inputs, and are usable by ML practitioners.

This paper opens several avenues for future work. Adapting modern density estimation methods for complex types of data, such as images, might hold the key to wider applicability of our method. Regarding evaluation, an immediate next step would be to carry out a follow-up user study to investigate various design choices, such as the technique for partitioning the classes in multi-class settings. Ideally, future user studies should involve participants' own models and should rely on questions that attempt to uncover insights that are relevant to their day-to-day experiences.

The findings from our user study offer important lessons that we believe are generally applicable to other interpretability tools. Chief among these is the importance of user-friendly and engaging tutorials that provide users with the necessary understanding of the tool and its intended usage, and users' desire for flexibility in tools. These results underscore the importance of putting human needs at the center of the design and evaluation of interpretability methods. The human computation community is uniquely situated to drive this work as it requires interdisciplinary expertise in both ML and HCI, fields that are central to HCOMP.

In the spirit of developing AI responsibly, we believe that papers proposing new interpretability methods should also provide a discussion, as we have done in the previous section, of not only those settings for which the proposed method is suitable, but also those settings that fall outside its scope. In addition, we recommend that authors should explicitly describe the notion of interpretability (or explanation) that they aim to operationalize, allowing readers to situate the contributions in relation to other interpretability methods and to understand their scope and applicability.

References

Abdul, A.; von der Weth, C.; Kankanhalli, M.; and Lim, B. Y. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Alvarez-Melis, D.; and Jaakkola, T. S. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 412–421.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018a. On the Robustness of Interpretability Methods. In *ICML Workshop on Human Interpretability in Machine Learning*.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018b. Towards Robust Interpretability with Self-explaining Neural Networks. In *Neural Information Processing Systems*, 7775–7784. Curran Associates Inc.

Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P. N.; Inkpen, K.; et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. Association for Computing Machinery.

Azevedo, P. J. 2010. Rules for contrast sets. *Intelligent Data Analysis* doi:10.3233/IDA-2010-0444.

Bar-Hillel, M. 1980. The base-rate fallacy in probability judgments. *Acta Psychol.* 44(3): 211–233.

Bay, S. D.; and Pazzani, M. J. 1999. Detecting change in categorical data: Mining contrast sets. In *Conference on Knowledge discovery and data mining (KDD)*. Association for Computing Machinery. doi:10.1145/312129.312263.

Braun, V.; and Clarke, V. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.*, APA handbooks in psychology., 57–71. Washington, DC, US: American Psychological Association. doi:10.1037/13620-004.

Bunt, A.; Lount, M.; and Lauzon, C. 2012. Are Explanations Always Important? A Study of Deployed, Low-Cost Intelligent Interactive Systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces (IUI)*, 169–178. Association for Computing Machinery.

Bussone, A.; Stumpf, S.; and O'Sullivan, D. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI)*, 160–169. doi:10.1109/ICHI.2015.26.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. Association for Computing Machinery. doi: 10.1145/2783258.2788613.

Dodge, J.; Liao, Q.; Zhang, Y.; Bellamy, R. K. E.; and Dugan, C. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI)*, 275–285. Association for Computing Machinery.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* .

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Eddy, D. M. 1982. Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. In Kahneman, D.; Slovic, P.; and Tversky, A., eds., *Judgment Under Uncertainty: Heuristics and Biases*, 249–267. Cambridge University Press.

Fernandes, K.; Vinagre, P.; and Cortez, P. 2015. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In *Progress in Artificial Intelligence*, 535–546. Springer International Publishing. doi: 10.1007/978-3-319-23485-4_53.

Gibson, J. J. 1977. The theory of affordances. R. Shaw and J. Bransford (eds.), Perceiving, Acting and Knowing.

Gold, J. I.; and Shadlen, M. N. 2001. Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* 5(1): 10–16. ISSN 1364-6613, 1879-307X. doi: 10.1016/s1364-6613(00)01567-9.

Gold, J. I.; and Shadlen, M. N. 2002. Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* 36(2): 299–308. doi: 10.1016/s0896-6273(02)00971-6.

Good, I. J. 1950. *Probability and the Weighing of Evidence*. Charles Griffin & Company Limited: London.

Good, I. J. 1968. Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor. *Br. J. Philos. Sci.* 19(2): 123–143. doi:10.1093/bjps/19.2.123.

Good, I. J. 1985. Weight of evidence: A brief survey. *Bayesian statistics* 2: 249–270.

Hastie, T. J.; and Tibshirani, R. J. 1990. *Generalized Additive Models*. CRC Press.

Hempel, C. G. 1962. *Deductive-nomological vs. statistical explanation*, volume 3 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

Hempel, C. G.; and Oppenheim, P. 1948. Studies in the Logic of Explanation. *Philos. Sci.* 15(2): 135–175. doi: 10.1086/286983.

Hohman, F.; Head, A.; Caruana, R.; DeLine, R.; and Drucker, S. M. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Hong, S. R.; Hullman, J.; and Bertini, E. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*. Association for Computing Machinery.

Hudson, S. E.; and Mankoff, J. 2014. Concepts, values, and methods for technical human–computer interaction research. In *Ways of Knowing in HCI*, 69–93. Springer.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.386.

Jung, J.; Concannon, C.; Shroff, R.; Goel, S.; and Goldstein, D. G. 2020. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(3): 771–800. doi:https://doi.org/10.1111/rssa.12576.

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Vaughan, J. W. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Koehler, J. J. 1996. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behav. Brain Sci.* 19(1): 1–17.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2019. Human Evaluation of Models Built for Interpretability. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 59–67.

Lakkaraju, H.; Bach, S. H.; and Jure, L. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery. doi:10.1145/2939672.2939874.

Li, Q.; Chu, S.; Rao, N.; and Nourani, M. 2020. Understanding the Effects of Explanation Types and User Motivations on Recommender System Use. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* 8(1): 83–91.

Lim, B. Y.; and Dey, A. K. 2011. Investigating Intelligibility for Uncertain Context-Aware Applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp)*, 415–424.

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

Lombrozo, T. 2012. Explanation and Abductive Inference. In Keith J. Holyoak and Robert G. Morrison, ed., *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press. ISBN 9780199968718. doi:10.1093/oxfordhb/9780199734689.013.0014.

Lundberg, S.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4768–4777.

Miller, T. 2019a. Contrastive Explanation: A Structural-Model Approach. *arXiv preprint arXiv:1811. 03163* .

Miller, T. 2019b. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267: 1–38. ISSN 0004-3702.

Norman, D. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

Nourani, M.; Kabir, S.; Mohseni, S.; and Ragan, E. D. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP)* 7(1): 97–105.

Novak, P. K.; Lavrač, N.; and Webb, G. I. 2009. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *J. Mach. Learn. Res.* 10: 377–403. ISSN 1532-4435. doi:10.1145/1577069.1577083.

Olsen Jr, D. R. 2007. Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, 251–258.

Papamakarios, G.; Pavlakou, T.; and Murray, I. 2017. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 2335–2344. USA: Curran Associates Inc. ISBN 9781510860964.

Peirce, C.; and Jastrow, J. 1885. On small differences in sensation. *Memoirs of the National Academy of Science* 3: 73–83.

Peirce, C. S. 1878. Illustrations of the Logic of Science: IV The Probability of Induction. *Popular Science Monthly* 12: 705–718.

Phillips, L. D.; and Edwards, W. 1966. Conservatism in a simple probability inference task. *J. Exp. Psychol.* 72(3): 346–354. ISSN 0022-1015. doi:10.1037/h0023653.

Pitt, J. C. 1988. *Theories of explanation*. Oxford University Press. ISBN 9780195049701.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Quinlan, J. R. 1986. Induction of Decision Trees. *Mach. Learn.* ISSN 0885-6125, 1573-0565. doi:10.1023/A:1022643204877.

Rezende, D. J.; and Mohamed, S. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, 1530–1538.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 1135–1144. New York, NY, USA: ACM. ISBN 9781450342322. doi:10.1145/2939672.2939778.

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.

Salmon, W. C. 1971. *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press.

Selbst, A.; and Barocas, S. 2018. The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 87.

Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; and Batra, D. 2017. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. In *ICCV*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312. 6034* .

Spiegelhalter, D. 2018. Making Algorithms Trustworthy: What Can Statistical Science Contribute to Transparency, Explanation and Validation? Neural Information Processing Systems, Invited Talk (Breiman Lecture).

Tan, S.; Caruana, R.; Hooker, G.; and Lou, Y. 2018. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.

Turner, C. W.; Lewis, J. R.; and Nielsen, J. 2006. Determining usability test sample size. *International encyclopedia of ergonomics and human factors* 3(2): 3084–3088.

Tversky, A.; and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185(4157): 1124–1131. ISSN 0036-8075.

van der Waa, J.; Robeer, M.; van Diggelen, J.; Brinkhuis, M.; and Neerincx, M. 2018. Contrastive explanations with local foil trees. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*.

van Fraassen, B. 1988. The Pragmatic Theory of Explanation. In Pitt, J. C., ed., *Theories of Explanation*. Oxford University Press.

Vaughan, J. W.; and Wallach, H. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In Pelillo, M.; and Scantamburlo, T., eds., *Machines We Trust: Perspectives on Dependable AI*. MIT Press.

Veale, M.; Van Kleek, M.; and Binns, R. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31(2).

Webb, G. I.; Butler, S.; and Newlands, D. 2003. On detecting differences between groups. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 256–265. Association for Computing Machinery.

Weld, D. S.; and Bansal, G. 2019. The Challenge of Crafting Intelligible Intelligence. *Communications of the ACM* 62(6): 70–79.

Zeng, J.; Ustun, B.; and Rudin, C. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180(3): 689–722.

Zhang, H.; and Maloney, L. T. 2012. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neurosci.* 6. ISSN 1662-4548. doi:10.3389/fnins.2012.00001.