Improving understandability of feature contributions in model-agnostic explainable AI tools

Improving understandability of feature contributions in model-agnostic explainable AI tools

Sophia Hadash
s.hadash@tue.nl
Jheronimus Academy of Data Science
's Hertogenbosch, Noord-Brabant, The Netherlands
Human-Technology Interaction Department, Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands, The Netherlands

Martijn C. Willemsen
m.c.willemsen@tue.nl
Jheronimus Academy of Data Science
's-Hertogenbosch, The Netherlands, The Netherlands
Human-Technology Interaction Department, Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands, The Netherlands

Chris Snijders
c.c.p.snijders@tue.nl
Human-Technology Interaction Department, Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands, The Netherlands

Wijnand A. IJsselsteijn
w.a.ijsselsteijn@tue.nl
Human-Technology Interaction Department, Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands, The Netherlands

ABSTRACT

Model-agnostic explainable AI tools explain their predictions by means of 'local' feature contributions. We empirically investigate two potential improvements over current approaches. The first one is to always present feature contributions in terms of the contribution to the outcome that is perceived as positive by the user ("positive framing"). The second one is to add "semantic labeling", that explains the directionality of each feature contribution ("this feature leads to +5% eligibility"), reducing additional cognitive processing steps. In a user study, participants evaluated the understandability of explanations for different framing and labeling conditions for loan applications and music recommendations. We found that positive framing improves understandability even when the prediction is negative. Additionally, adding semantic labels eliminates any framing effects on understandability, with positive labels outperforming negative labels. We implemented our suggestions in a package ArgueView[11].

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; **Interactive systems and tools**; Information visualization.

KEYWORDS

interpretable machine learning, explanations, argumentation, natural language

ACM Reference Format:
Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3491102.3517650

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517650

1 INTRODUCTION

While AI systems are increasingly applied to inform a broad array of real-world recommendation and decision-making processes, including financial decision making, medical recommendations, and personnel selection, many AI systems are notorious for their black box nature, leaving users to guess how or why certain decisions have been reached. The field of Explainable AI (XAI) has emerged to address challenges associated with interpreting black box models, for instance by coming up with ways in which to present model outcomes that are easier to interpret. The growing effort in the Explainable Artificial Intelligence (XAI) domain has resulted in several explanation tools such as Shap [18], LIME [25], and Anchor [26]. These interpretability tools take one of two approaches. Some approaches exploit a model's inner workings. This is feasible when models are fairly simple such as linear systems [5] and point systems [4, 32]. Other techniques are model agnostic. Such techniques often probe the actual model in the neighborhood of a prediction of interest to generate a simpler model that behaves similar to the black-box model (e.g. [18, 25, 26]).

As we will show, the state-of-the-art interpretability tools have several characteristics that make interpretation more difficult than it needs to be, especially for users that are less experienced with statistical or machine-learning models. We discuss two adaptations that can improve the users' understanding of algorithmic suggestions. The first one is to use positive framing and the second is to offer semantic feature contributions.

In some cases, decisions are inherently positive or negative for a user. For instance, when applying for a loan, getting the loan granted is the obvious positive outcome for the consumer. However, current interpretability tools do not consider whether the decision-class is positive ("consumer gets the loan") or negative ("consumer does not get a loan"). Instead, these tools label feature contributions as positive when they contribute to the suggested decision-class. When a model suggests an outcome that is negative for a user ("no loan granted"), this can lead to the confusing situation where

positively displayed contributions are in fact negative experiences for a user. To prevent this, we propose to frame explanations as contributing to the class that is conceived as the positive outcome ("positive framing").

The second improvement is also aimed at reducing the cognitive load of users. Currently, state-of-the-art tools use positive or negative values, or bars, to represent the contributions that features make to a decision. However, interpretation of those values requires that users make an additional cognitive step to understand what these values or bars imply. For example, in the case of loan applications, users must deduce whether a contribution of "+5%" implies an increase or a decrease in loan applicability. We therefore propose to explicitly label (using text, not numbers) what a feature contributes to the decision, making it easier for users to interpret the importance of a given feature correctly ("semantic labeling"). We give examples of positive framing and semantic labeling in the next section.

We tested these two improvements in a user study where we show explanations with and without these potential improvements to (lay) participants. We found that explanations that use positive framing, without semantic labels, lead to higher understandability compared to explanations with a negative framing (irrespective of the decision-class). When semantic labels are used, the effect of framing on understandability disappears, with positive labels always being more effective than negative labels, irrespective of the decision class.

In what follows, we first review the state-of-the-art explanation tools and the improvements we propose in more detail. We then report the design and results of our study. A discussion about the implications of our results for the design of explainability tools and the solutions we provide in the ArgueViewcompletes this paper.

2 STATE-OF-THE-ART EXPLANATION TOOLS

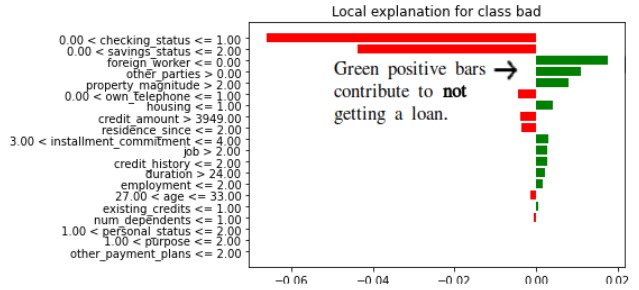
While the field of explainable AI (XAI) really took off around 2015 (consider, for instance, the strong increase in frequency of search terms such as "explainable AI" [22]), trying to explain models has a longer history. Linear regression models are an example of interpretable models which were used by Gauss, Legendre, and Quetelet [2] as early as the beginning of the 19th century. Especially linear regression (type) models allow for relatively easy interpretations, although even then it is easy to misinterpret model results [13]. In machine learning, there is a (much) less obvious connection between model input and output as they often relate in a non-linear way. This causes machine learning models to be less interpretable than regression type models and they can be considered "black-box" even to the most knowledgeable of model-makers. There are several publicly available tools to interpret model output from black-box models (see [1] for a review). Some of these tools are model-specific, for example to interpret deep neural networks [15, 17, 33] or tree ensembles [7, 12, 20, 28]. Other tools (e.g. [6, 18, 21, 23, 25–27]) are model-agnostic and can be used to explain decisions independent of the type of model used [24]. To explain the output for a certain set of inputs, the model-agnostic interpretability tools typically construct a simpler surrogate model by sampling the original (black-box) model near the inputs of interest, and then offering an explanation based on this simpler model. We focus on these model-agnostic interpretable AI tools that use surrogate models and their functionality towards explaining singular decisions.

A commonality in model-agnostic explanation tools is that, to construct explanations, most of the tools supply in their output a list of input features and their 'importance' for a given decision (see e.g. [18, 21, 25, 27]). These feature importance values are computed in different ways, depending on the tool, but these values always represent the extent to which a feature 'contributes' to a decision. Besides computing these feature importance values, some tools provide visualizations of these values using bar charts, waterfall plots, or "force plots" [18, 25]. While it is certainly useful to have interpretability tools for people who work with machine learning outputs, the usefulness of these tools can reach beyond those with technical skills. Explanations may also help people who use the model or people whose lives are affected by a model's output ([3, 30]). One could think of examples in the legal domain [16] for issues of accountability, in the medical domain for medical decision-making [8], or of all recipients of (algorithmic) decisions who would want to know how a certain decision that affects them (e.g. a mortgage decision, a job application) came about. In fact, since May 2018, consumers have the "right to explanation" by law in Europe [10]. The explanations that the tools provide should be as clear as possible, given that one cannot count on the fact that the user is tech-savvy enough to figure out what the technical output conveys.

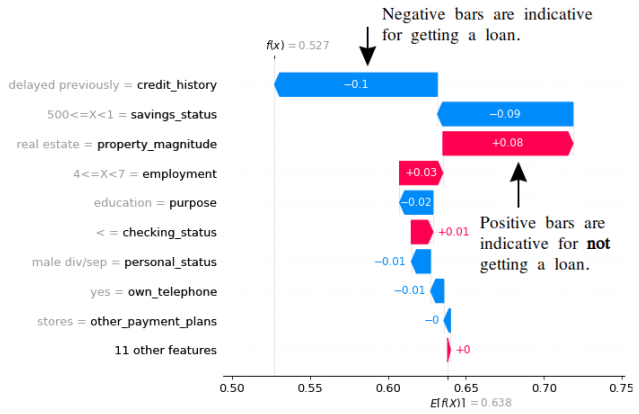
3 ISSUES AND IMPROVEMENTS

LIME and Shap are two of the most commonly used interpretability tools in XAI (e.g. see its implementations [1]). LIME is a model-agnostic explanation tool developed by Ribeiro, Singh, and Guestrin [25]. It generates its explanations using a surrogate linear model and proximity sampling. The generated explanation is in the format of a feature-importance list and a visualization. Similar to LIME, Shap [18] is another popular explanation framework that provides tools to interpret model output. Shap uses a different approach to calculate the surrogate model, but it also outputs a feature-importance list and visualizations, similar to LIME (Shap is also able to provide global explanations and visualizations, but these are outside the scope of this paper). We focus on Shap's waterfall plot, arguably the most straightforward and easiest to understand local explanation visualization.

Figure 1 shows the visualizations of LIME (top) and Shap (bottom) of an example case in the loan application domain. The underlying machine-learning model predicts whether a user is eligible or ineligible for a loan. We chose a case in which the algorithm decided that the user is ineligible for the loan. Because we picked a case where the decision class was negative to the user, the contributions that LIME and Shap show use positive values when features contribute towards not getting the loan. This may seem logical and straightforward for data scientists who work with prediction models regularly. But users unfamiliar with visualizations from interpretability tools and regression models may find it more intuitive when positive values imply a contribution to the outcome that is perceived as positive. We find evidence for this idea in cognitive psychology. A running theme in cognition is that people are better in handling positive information than negative information [19].

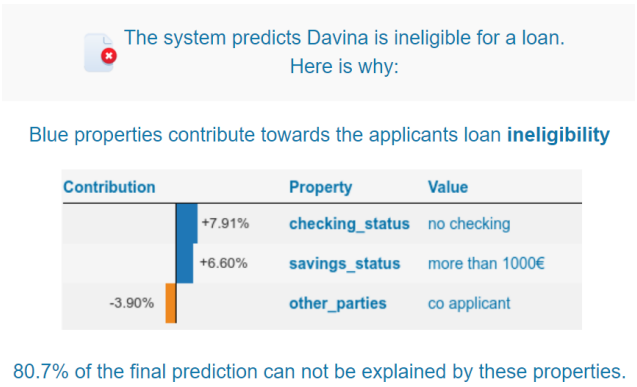


(a) LIME's map visualization

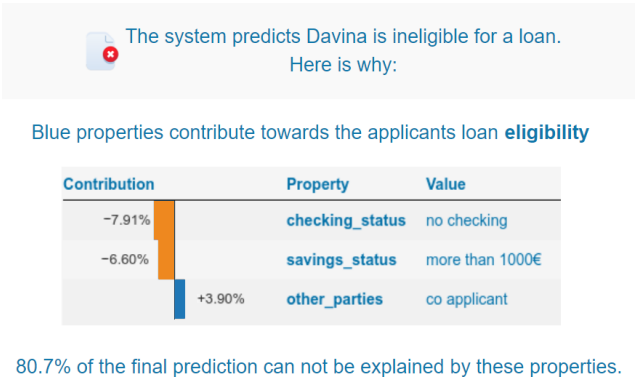


(b) Shap's waterfall visualization

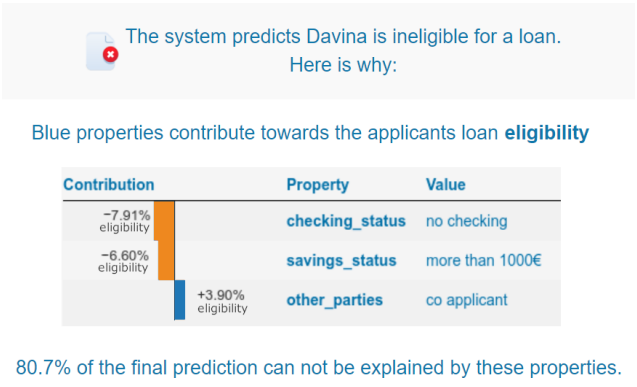
Figure 1: Example visualizations generated by LIME [25] (a) and Shap [18] (b). In this example, both visualizations use positive bars and numbers for contributions that are negative for the user, i.e. counter-intuitive framing. The example is from the OpenML credit-g dataset (case 496). The Shap visualization is generated using the waterfall visualization of Shap's perturbation explainer.



(a) Alignment as in LIME/Shap (based on decision-class)



(b) Contributions aligned with outcome



(c) Explicit contribution implications

Figure 2: Explanation changes according to the proposed improvements shown for an explanation of a negative decision-class. (a) Explanation similar to current LIME and Shap where contribution implications are based on the decision-class. (b) Proposal 1: framing contributions such that positive contributions imply positive outcomes. (c) Proposal 2: explicit contribution implications using labels.

For example, we understand sentences better if they use affirmative wording ("Mary is honest") than in negative wording ("Mary is not dishonest") [14]. Furthermore, Hearst found that we tend to perform better on various tasks if the information is emotionally pleasant rather than unpleasant, which is called the "Pollyana principle". It even seems to be the case that our cognitive system first makes people accept or believe information that comes in, before rejecting it when it appears to be false [9]. These observations suggest that we should frame explanations positively whenever possible and avoid (double) negatives in the cognitive interpretation process. In LIME and Shap, contributions are always framed as adding to the current decision-class, irrespective of whether the user perceives this as positive or negative. Thus, when a decision-class is perceived as negative for a user ("loan not granted"), positive numbers and bars are used for contributions with implications that are perceived as negative, and negative numbers result in implications that are perceived as positive. We propose to change this by framing feature contributions such that positive values imply positive outcomes (see the change from Figure 2a to Figure 2b). We hypothesize that **(H1)** framing explanations positively even when a decision has a negative implication improves the understandability of explanations.

There is, however, a more general underlying issue. Good explanations should be self-explanatory and depend as little as possible

on implied knowledge. We therefore propose to remove the ambiguity of how contributions are interpreted by making the implications of feature contributions explicit, by adding semantic labels to each feature contribution. For instance, this would imply changing "+5%" to "+5% eligibility" (see Figure 2c), making the direction of the contribution even more explicit. We hypothesize that **(H2a)** explanations with semantic labels are more understandable than explanations without semantic labels. When choosing the labels, we can opt for positively framed labels ("eligibility") or negatively framed labels ("ineligibility"). Since research suggests positive information can be cognitively handled more easily [19], we anticipate that **(H2b)** positively framed labels result in a better understandability compared to negatively framed labels. Furthermore, making the feature implications explicit might reduce the need for positive framing. Users no longer need to correctly deduce what contributions implicate. As such, we hypothesize that **(H3)** effects of contribution framing on understandability will be smaller when semantic labels are introduced. An implementation of both improvements has been implemented in a package[1] which also resolves some additional issues with LIME and Shap's visualizations, such as visual additivity, the inclusion of additional information about the features, and readability issues.

4 STUDY

A user study was designed to evaluate our proposed improvements. The study was reviewed and approved by the institutional review board. Participants received a number of loan application and music recommendation cases with explanations that varied in terms of the predictions, labels and framing. Participants interpreted the explanations and rated their understandability. The visualizations used in the study are shown in Figure 3.

4.1 Design

The study used a 3x2x2 mixed-effects design: labels x domain x framing. Semantic Labels (none / positive /negative) were varied between-subjects to avoid any confusion when labels change and to prevent carry-over effects due to this. Domain and framing were varied within-subject to increase the number of observations and statistical power as we did not expect these variations to cause confusion or carry-over effects. Each within-subjects condition (2 framing x 2 domains, total 4) consisted of 6 trials (i.e. explanations) of which 3 had a positive prediction and 3 had a negative prediction. The order of the within-subjects conditions was randomized.

4.1.1 Label condition. The label condition was chosen as a between-subjects condition (e.g. each participant only sees one type of label). The options for the labels are "none" for no labels (which is similar to LIME and Shap), "eligible/like" for positive semantic labels ("eligible" in the loan application condition, "like" in the music recommender condition), and "ineligible/dislike" for negative semantic labels. The example explanation in Figure 3 shows the positive label condition.

[1]package is available in our open-source PyPi and NPM package ArgueView. It can be used in conjunction with LIME and Shap to generated explanations with adjustable contribution framing and contribution labels.

4.1.2 Domain condition. In the loan application domain, participants were required to interpret explanations of a loan application prediction system. The system would predict for a number of cases whether someone is loan eligible or loan ineligible. In the music recommender domain participants could choose one of four profiles: "hiphop", "jazz", "pop", "rock". Based on their selected profile they received music recommendations with various predictions (e.g. "you like this" or "you dislike this"). An explanation showed why the prediction was made based on the audio features of the song and their similarity to the selected profile. Participants started with trials in a randomly chosen domain. After all the trials for that domain were completed (12, 2 conditions x 6 trials), they continued to the trials for the other domain.

4.1.3 Framing condition. Framing was a within-subjects condition that affects the explanation underneath the prediction (i.e. "Blue properties contribute towards the applicants loan (in)eligibility"), see Figure 3. Depending on the framing condition, the sentence mentioned the color that contributes towards the positive or negative decision-class. Furthermore, in the absence of labels the framing condition also affected the directionality of the feature contributions. In the positive framing condition, positive contributions indicated positive implications (i.e. 'increased eligibility') whereas in the negative framing condition positive contributions indicated negative implications (i.e. 'increased ineligibility'). When labels were presented the framing manipulation was less strong, because then the direction is determined by the labels and the framing solely manipulated the explanation underneath the prediction (Figure 3).

4.2 Measurements

4.2.1 Perceived understandability. The primary (dependent) variable of interest in this study was the perceived understandability of the explanations which was measured using two scales during the study: at the trial-level, and at the condition-level. The trial-level scale consisted of one item: "How well do you understand the explanation of how the properties contribute to the prediction (i.e. the contribution column)?", which was answered on a 4-pt Likert scale ("Not at all", "Somewhat", "Mostly, "Completely"). At the condition level (i.e., after every 6 trials), we used a four item questionnaire that measured the overall understandability of the condition. The items were combined into one scale using factor analysis, see Table 1.

4.2.2 Co-variates. Apart from several demographic variables, the main co-variate that was measured was user agreement. User agreement might affect understandability through confirmation bias [29]. People tend to conform their beliefs rather than disprove them, so when a decision is in line with their beliefs, they are more likely to accept it, which might influence their perceived understanding. User agreement was measured at the trial level using the item "Do you agree with this prediction? (No-Yes)".

4.3 Participants

The study sample consisted of 133 participants (male = 61). Participants were sampled from the university database (n = 91) and convenience sampling (n = 42). 79% of participants were younger than 34. Participants from the database were rewarded using a raffle

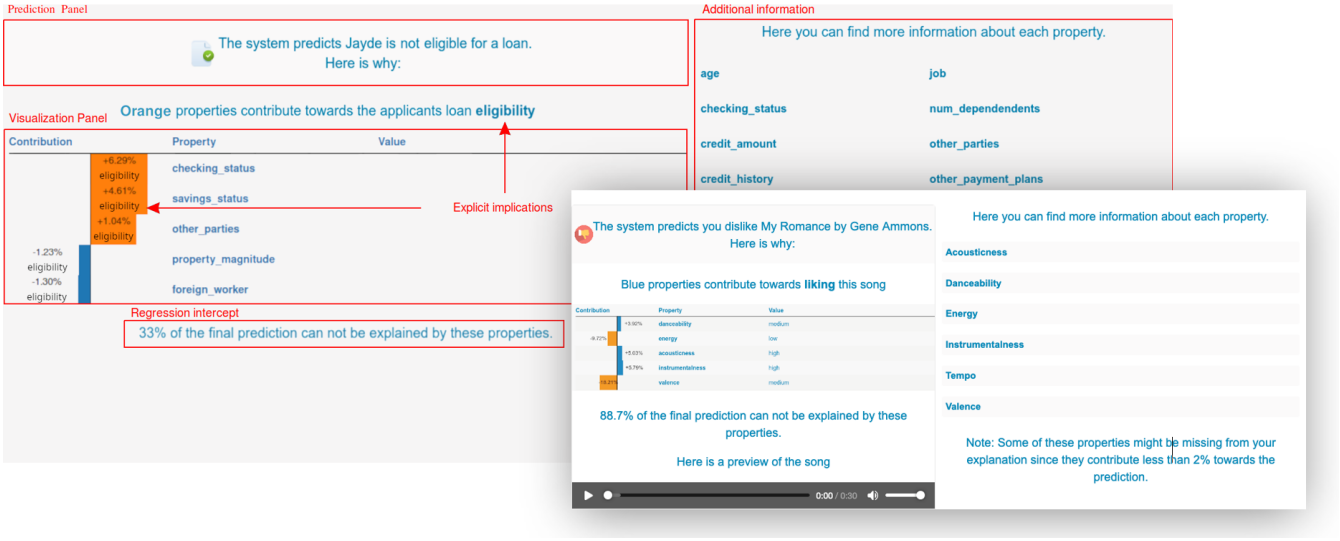


Figure 3: Examples of explanations as shown to participants. The loan application domain is shown in the background. On the left side from top to bottom: prediction panel, contribution explanation sentence, visualization panel, regression intercept. On the right side: additional information panel. Uses the OpenML credit-g dataset (case 496). The visualization for the music recommendation domain is shown in the foreground, where the only difference is the playback bar in the bottom and the different features and decision-classes.

Item	Factor Loading	Specific Variance	Communality
The explanations helped me get more insight into the given prediction.	0.44	0.60	0.19
The explanation felt clear to me.	0.77	0.48	0.60
I felt that the explanations took me a lot of time to comprehend.	1.17	0.65	1.37
The explanations were confusing to me.	1.13	0.53	1.28

Table 1: The perceived understandability scale (condition level). Cronbach's $\alpha = .73$. The items used a 5-pt Likert scale and were preceded by the question "To what extent do you agree with the following statements". Factor loadings are calculated using a principal component analysis without rotation. PCA was appropriate for this scale as shown by Kaiser-Meyer-Olkin's $MSA = 0.71$ and Bartlett's test of sphericity $\chi^2(6) = 531, p < .001$ [31].

with a 20% chance to win 25 euro. The other participants received a 5 euro gift card.

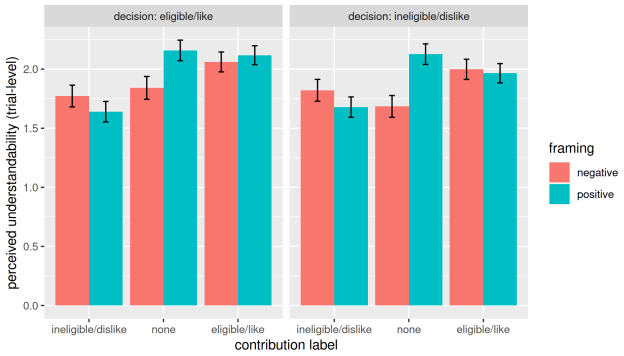
4.4 Procedure

After agreeing with the informed consent, participants were directed to a demographics form and explained the study procedure. Participants were to either start with the loan application domain or with the music recommender domain. In the loan application domain, participants were instructed to imagine a scenario where they are loan auditors tasked with evaluating the understandability of a new decision-support system for loan applications. In the music recommender domain, participants were instructed to select one out of four genres (rock, pop, hip-hop, or jazz). Then, they were instructed that an AI system would use their preferred genre to predict songs based on audio features.

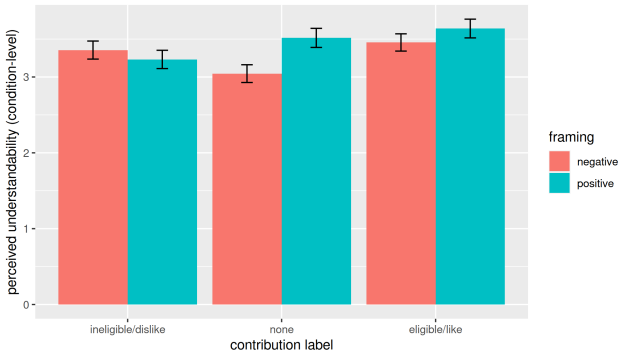
After the instructions the participants were directed to the trials. The trials consist of a decision, an explanation, an information panel, and a question panel. The explanation consists of three elements. The primary component is the list visualizer from the ArgueView-package (Figure 3). On top of the visualizer there was a sentence explaining the contribution of the features. This sentence and the color direction were manipulated by the *framing* condition. Below the visualizer was an indicator for the contribution that could not be explained by the features (i.e. unexplained variance). The question panel asked about perceived understandability and agreement (one question at a time). After every condition with six trials participants received the perceived understandability scale and an (objective) understandability check, before moving to the next condition. Finally, after the 4 conditions (24 trials) were presented, participants were thanked for their participation.

4.5 Results

The study ran online between 19 March and 18 April of 2021. Participants with incomplete submissions (n = 12) or with < 10 min completion time (n = 7) were excluded. Several trials were deleted because of coding error in the visualizations that were shown. These errors were mostly in the negative framing condition in the music



(a) trial-level



(b) condition-level

Figure 4: Perceived understandability for the various study conditions: decision-class x framing x latent continuous variable label. (a) Trial-level understandability metric. (b) Condition-level understandability metric. Error bars indicate one standard error of the mean.

domain in the hip-hop, jazz, and pop profiles. Fortunately, a large proportion of people (46%) chose the rock profile which was free of errors. The results did not change significantly with inclusion or exclusion of these trials, and we observed similar effects when only analyzing the trials of the loan domain that did not have this coding error.

The data analysis consists of two multi-level regression analyses, for the trial-level understandability (Model 1) and for the condition-level understandability (Model 2), see Table 2. The estimated means across the conditions of both regression analyses are visualized in Figure 4.

The variables in the regression are coded as follows. The decision class and the framing are mean centered (positive decision/framing is coded positively) while the three label conditions are dummy coded. Using this coding, we can easily compare the baseline model (the no-label condition) against the label conditions that are represented by interactions with the label dummies. In our results section we first look at the data for the conditions without labels, which will be used to answer H1. After this we look at the effects of adding feature labels and answer H2 and H3.

4.5.1 Co-variates. Each regression analysis includes the following co-variates: unexplained model contribution, domain, user agreement, and whether it is the first trial of a condition (as understandability might be lower in the first trial when all information is new). Additionally, models that use the condition-level understandability metric include the understandability check as a co-variate.

There are several significant main effects of the co-variates. The explanations in the loan application domain are more understandable compared to the music domain, $\beta = -0.106, p < 0.001$, but this effect is only seen in the trial-level understandability metric. When a user agreed with the explanation, they found the explanation more understandable, $\beta = 0.434, p < 0.001$ (trial-level), $\beta = 0.187, p < 0.001$ (condition-level). Whether it was the first trial or whether the understandability check was answered correctly did not have an effect on perceived understandability.

4.5.2 Does positive framing improve understandability? (H1). For this analysis we only consider the no-label condition to measure the pure effect of framing. We hypothesized in H1 that using positive framing improves understandability, even if the decision class is negative. Figure 4, suggests this is the case, as the center columns of both the trial and condition-level show higher perceived understandability for the positive frame than the negative frame. Indeed our regression (Table 2) shows a positive main effect of framing for model 1 (trial) and model 2 (condition level). At the trail level[2] the effect seems slightly stronger for the positive than negative decision class, but we do not observe a significant interaction (β = -0.031, p = 0.224), indicating that positive framing leads to higher understandability irrespective of decision-class. Both LIME and Shap are currently framing explanations based on the decision-class, which can lead to explanations where contributions with positive outcomes are actually framed negatively (i.e. with negative values/bars), and hence our finding shows that changing the framing in these situations to positive would improve understandability.

4.5.3 Do feature labels increase understandability? (H2). We hypothesized that adding feature contribution labels to the visualization would increase understandability. Figure 4 shows a mixed picture. Comparing across both framing conditions, positive labels (right bars in the graphs) seem to be a bit more understandable than the no-labels at the condition level, which is supported by a positive significant main effect of positive labels (β = 0.267, p = 0.024). However, for the trial-level the differences seem much smaller and no effect of positive label on understandability is observed. The negative label conditions (leftmost bars in Figure 4) seem to result in somewhat lower understandability than the no label condition (middle bars), but mostly for the trial-level results. Indeed only Model 1 shows a negative effect of the negative label on understandability at the trial level (β = -0.225, p = .039). Taken together, we do not find strong support for H2a that stated that any type of labels improves understandability compared to no-labels. A post-hoc contrast analysis combining the effects of both label conditions confirms that using any label compared to not using a label does not lead to higher understandability (Model 1, β = -0.142, $t_{ratio}(128)$ = -0.773, p = 0.441).

[2]The condition-level metric was measured across three negative and three positive decision trials, so no interactions with decision-class can be modeled

	Model 1 Trial-level			Model 2 Condition-level		
	β	SE	t	β	SE	t
H1: pos. framing improves understandability						
decision (c)	0.047	0.047	1.480			
framing (c)	0.190***	0.027	7.123	0.236***	0.061	3.855
decision (c) x framing (c)	-0.031	0.026	-1.217			
H2: feature labels improve understandability						
positive label	0.083	0.104	0.801	0.267*	0.117	2.288
negative label	-0.225*	0.108	-2.090	0.013	0.120	0.111
decision (c) x positive label	0.007	0.035	0.197			
decision (c) x negative label	-0.069	0.037	-1.854			
H3: labels reduce framing usefulness						
framing (c) x positive label	-0.184***	0.035	-5.235	-0.144	0.079	-1.820
framing (c) x negative label	-0.259***	0.037	-7.037	-0.297***	0.082	-3.610
decision (c) x framing (c) x positive label	0.054	0.034	1.563			
decision (c) x framing (c) x negative label	0.034	0.036	0.938			
Co-variates						
unexplained (c)	-0.017	0.021	-0.822	-0.004	0.046	-0.092
music domain	-0.106***	0.030	-3.472	0.086	0.083	1.039
agreement	0.434***	0.034	12.681	0.187***	0.038	4.922
first trial	0.026	0.037	0.708			
correct understandability check				0.029	0.074	0.393
Intercept	1.629***	0.081	20.144	3.280***	0.105	31.202
Statistics						
N	2490			532		
Log likelihood	-2769.2			-662.2		
$R^2_{GLMM(c)}$	0.381			0.311		
Random effects						
# of participants	128			133		
Participant *SD*	0.455			0.408		

Table 2: Multi-level regression analysis results. (c) = centered, unexplained = the amount of unexplained variance in the explanation shown to the user (normalized, centered), agreement = user agreement scale (yes/no), first trial = observation is the first trial in a session (yes/no), correct check = understandability check was answered correctly (yes/no), N = number of observations, $R^2_{GLMM(c)}$ = R^2 statistic of the full model (fixed+random effects), # = number, SD = standard deviation, cond. = condition, pos. = positive. 'agreement' and 'unexplained' are averaged over all the trials in the condition and normalized. ***p < .001; **p < .01; *p < .05.

We do find support for H2b, that states that positively framed labels work better than negatively framed ones. Figure 4 shows that understandability is higher for positive labels than for negative labels. A post-hoc contrast analysis supports this: positive labels are more understandable than negative labels (trial-level: β = 0.308, $t_{ratio}(127)$ = 2.959, p = 0.004; condition-level: β = 0.254, $t_{ratio}(132)$ = 2.176, p = 0.031).

4.5.4 Does positive framing still have additional benefits when labels are added? (H3). When labels are added, users do not need to deduce the implication of a value anymore, which we expected to reduce the advantage of positive over negative framing. Indeed, 4 shows that at both the trial level and the condition level, we do not observe large differences anymore between the red and blue bars for the label conditions (outer columns), as we do for the no-label condition (middle columns).

Our trial-level regression model corroborate this finding as the framing effect we found for the no labels condition (β = 0.190, p < 0.001) is completely counteracted by negative interaction of framing with positive labels (β = -0.184, p < 0.001) and negative labels (β = -0.259, p < 0.001). A post-hoc contrast analysis confirms this as we find no significant total effect of framing (main effect + interaction effect positive labels: β = 0.048, $t_{ratio}(2369)$ = 1.409, p = 0.159; main effect + interaction effect negative labels: β = -0.026, $t_{ratio}(2365)$ = -0.701, p = 0.483).

The condition-level model (Table 2, Model 2) shows similar results with a positive main effect of framing (β = 0.236, p < 0.001), counteracted by a negative interaction effect with positive labels

(β = -0.144, p = 0.070), and negative labels (β = -0.297, p < 0.001). Post-hoc contrast analysis yields that the effect of framing is reduced but not completely eliminated for positive labels: main effect + interaction effect positive labels (β = 0.18, $t_{ratio}(410)$ = 2.149, p = 0.0322), but it is completely eliminated for negative labels: main effect + interaction effect negative labels (β = 0.0269, $t_{ratio}(404)$ = 0.312, p = 0.756).

In summary, the effect of framing on understandability is largely eliminated when labels are introduced (with the exception of the condition-level metric with positive labels). We find support for H3: positive framing is no longer any better than negative framing when labels are added.

5 DISCUSSION AND CONCLUSION

In this paper we identified and addressed several issues in the most popular interpretability tools, LIME and Shap. Both tools use framing based on the decision-class which can cause counter-intuitive explanations in which positive contributions imply outcomes that are perceived as negative. We proposed to address this issue by always framing feature contributions with respect to the class that is perceived by the user as positive. The results showed that it is always better to frame the explanation using the positive decision class, even when the prediction class is negative. This result builds on existing evidence in cognitive psychology, where there are many studies that show people are better at handling positive information (see [19]). As obvious as this may seem, positive framing is not the default in many local interpretability tools (e.g. [18, 21, 23, 25]), in all likelihood because the software itself cannot determine what the positive outcome is. Allowing the model-maker or the user to set which of the outcomes is perceived to be the positive one can significantly improve the understandability of explanations. In addition, we proposed it may be even better to make the implications of feature contributions explicit using semantic labeling. Current interpretability tools display their results using (positive or negative) numeric feature contributions that can be hard to interpret. Hence, we proposed that semantically labeling the feature contributions improves understandability, while reducing the importance of positive framing. The results showed that labeling indeed improved understandability, but only when positive labels are used (e.g. 'eligibility' instead of 'ineligibility'). However, the more prominent finding is that when semantic labels were used, the positive framing no longer affected understandability. We anticipate that the reason for this is the removal of a cognitive step. Without labels participants had to deduce the meaning of feature contributions from the context, but with the labels they could directly see what a contribution implies and hence did not need to spend the cognitive effort to deduce the meaning. Although understandability was highest in the positive semantic label conditions, it was not substantially higher than the understandability in the (non-labeled) positive framing conditions. So both solutions seem equally effective in improving understandability: positive framing of the contributions or using positive semantic labels. Further research is needed to understand better whether and how cognitive effort is reduced when using labels.

Our study has several limitations. First, the decisions had a limited direct importance to the participants. The loan applications were for imaginary people with no relation to the participant, hence participants might have been indifferent to the actual prediction. Similarly, in the music domain, though the predictions were personalized based on their preferred genre, there was no consequence for the participant to whether the prediction was positive or negative. Another issue is that we presented participants with 24 trials which might have caused some learning effects or decreased attention during the trial. Then again, the entire study did not take longer than 20-25 minutes and our randomization of domains and conditions balances any effects of fatigue on our data. Also, including a variable 'first trial' in our model did not show systematic effects of fatigue or learning across the 6 trials in each condition. While acknowledging the above-mentioned limitations, our work addresses an important challenge of the CHI community: designing AI decision tools in such a way that their decisions or recommendations are easily interpretable, not just by experts but also by non-experts. This requires a more explicit connection between the XAI design community and the significant body of work in cognitive psychology and decision sciences. Our empirical contribution has highlighted the added value of positive framing and (positive) semantic labeling in aiding AI understandability. Positive framing should of course only be used in situations where there is a decision-class that users value positively as it might otherwise backfire. It is also complicated to have the software automatically deduce which decision-class is the positive one. Additionally, frequent users of interpretation tools may have come to expect positive contributions to imply positive contribution to the decision-class. Semantic labeling of feature contributions, on the other hand, can and should always be used to reduce user error and improve the understandability of feature contributions, so we propose to always include such labeling in future implementations.

ACKNOWLEDGMENTS

This work is part of the TEPAIV project number 612.001.752, which is financed by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] Namita Agarwal and Saikat Das. 2020. Interpretable Machine Learning Tools: A Survey. *2020 IEEE Symposium Series on Computational Intelligence, SSCI 2020* (2020), 1528–1534. https://doi.org/10.1109/SSCI47803.2020.9308260

[2] H Ahrens. 1988. Stigler, Stephen M.: The History of Statistics. The Measurement of Uncertainty before 1900. The Belknap Press of Harvard University, Cambridge, Mass., & London 1986; XVI, 410 S. *Biometrical Journal* 30, 5 (1988), 631–632. https://doi.org/10.1002/bimj.4710300527

[3] Vaishak Belle and Ioannis Papantonis. 2020. Principles and practice of explainable machine learning. *arXiv* (2020). arXiv:2009.11698

[4] Adrian Brasoveanu, Megan Moodie, and Rakshit Agrawal. 2020. Textual evidence for the perfunctoriness of independent medical reviews. *CEUR Workshop Proceedings* 2657 (2020), 1–9. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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*. ACM New York, NY, USA, Sydney, NSW, Australia, 1721–1730. https://doi.org/10.1145/2783258.2788613

[6] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. 2021. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence* 296 (2021), 103471. https://doi.org/10.1016/j.artint.2021.103471

[7] Jan Forberg, Annett Mitschick, Martin Voigt, and Raimund Dachselt. 2019. Interactive Exploration of Large Decision Tree Ensembles. (2019), 0–5. https://doi.org/10.1145/1122445.1122456

[8] John Fox. 2017. Cognitive systems at the point of care : The CREDO program. *Journal of Biomedical Informatics* 68 (2017), 83–95. https://doi.org/10.1016/j.jbi.2017.02.008

[9] Daniel T Gilbert, Douglas S Krull, and Patrick S Malone. 1990. Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of personality and social psychology* 59, 4 (1990), 601.

[10] B. Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741 arXiv:1606.08813

[11] Sophia Hadash. 2021. https://pypi.org/project/argueview/

[12] Satoshi Hara and Kohei Hayashi. 2016. Making Tree Ensembles Interpretable. arXiv:1606.05390 [stat.ML]

[13] Andrew F Hayes, Carroll J Glynn, and Michael E Huge. 2012. Cautions Regarding the Interpretation of Regression Coefficients and Hypothesis Tests in Linear Models with Interactions. *Communication Methods and Measures* 6, 1 (2012), 1–11. https://doi.org/10.1080/19312458.2012.651415

[14] Eliot Hearst. 1991. Psychology and Nothing. *American Scientist* 79, 5 (1991), 432–443. http://www.jstor.org/stable/29774477

[15] Peter K Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, and Steffan B Paul. 2021. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLOS Computational Biology* 17, 5 (2021), 1–21. https://doi.org/10.1371/journal.pcbi.1008925

[16] Joshua A Kroll, Joanna Huey, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2017. Accountable Algorithms. *Pennsylvania law review* 165, 3 (2017), 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

[17] Xuan Liu, Xiaoguang Wang, and Stan Matwin. 2018. Improving the Interpretability of Deep Neural Networks with Knowledge Distillation. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. 905–912. https://doi.org/10.1109/ICDMW.2018.00132

[18] Scott M Lundberg and Su-in Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. (Eds.). Curran Associates, Inc., Long Beach, CA, USA, 4765–4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[19] M.W. Matlin. 2008. *Cognition*. Wiley. https://books.google.nl/books?id=BUEdAQAAMAAJ

[20] Rory Mitchell, Eibe Frank, and Geoffrey Holmes. 2021. GPUTreeShap: Massively Parallel Exact Calculation of SHAP Scores for Tree Ensembles. arXiv:2010.13972 [cs.LG]

[21] Christoph Molnar. 2018. iml: An R package for Interpretable Machine Learning. *Journal of Open Source Software* 3, 26 (2018), 786. https://doi.org/10.21105/joss.00786

[22] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Communications in Computer and Information Science* 1323, 01 (2020), 417–431. https://doi.org/10.1007/978-3-030-65965-3_28 arXiv:2010.09337

[23] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. (2019), 1–8. arXiv:1909.09223 http://arxiv.org/abs/1909.09223

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, Been Kim, Dmitry M. Malioutov, and Kush R. Varshney (Eds.). ArXiv, New York, NY, USA, 91–95. arXiv:1606.05386 http://arxiv.org/abs/1606.05386

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (2016). https://doi.org/10.18653/v1/N16-3020 arXiv:1602.04938

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors : High-Precision Model-Agnostic Explanations. (2018).

[27] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 618–628. https://doi.org/10.1145/3351095.3375234

[28] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. 2020. Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference* (Virtual Event, USA) *(FODS '20)*. Association for Computing Machinery, New York, NY, USA, 23–34. https://doi.org/10.1145/3412815.3416893

[29] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[30] Adrian Weller. 2017. Challenges for Transparency. (2017). https://doi.org/10.1063/1.523063 arXiv:1708.01870

[31] Brett Williams, Andrys Onsman, and Ted Brown. 1996. Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care* 19, May (1996), 42–50. https://doi.org/10.1080/09585190701763982 arXiv:1512.00567

[32] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 180, 3 (2017), 689–722. https://doi.org/10.1111/rssa.12227 arXiv:1503.07810

[33] Xinyang Zhang, Ren Pang, Shouling Ji, Fenglong Ma, and Ting Wang. 2021. i-Algebra: Towards Interactive Interpretability of Deep Neural Networks. arXiv:2101.09301 [cs.LG]