Generating Simple, Conservative and Unifying Explanations for Logistic Regression Models

Sameen Maruf[*]
Oracle
Melbourne, Australia
sameen.maruf@gmail.com

Ingrid Zukerman
Dept. of Data Science and AI
Faculty of Information Technology
Monash University, Australia
ingrid.zukerman@monash.edu

Xuelin Situ[*]
Oracle
Melbourne, Australia
situsnow@gmail.com

Cecile Paris
CSIRO Data61, Australia
Cecile.Paris@data61.csiro.au

Gholamreza Haffari
Dept. of Data Science and AI
Faculty of Information Technology
Monash University, Australia
gholamreza.haffari@monash.edu

Abstract

In this paper, we generate and compare three types of explanations of Machine Learning (ML) predictions: *simple*, *conservative* and *unifying*. Simple explanations are concise, conservative explanations address the surprisingness of a prediction, and unifying explanations convey the extent to which an ML model's predictions are applicable.

The results of our user study show that (1) conservative and unifying explanations are liked equally and considered largely equivalent in terms of completeness, helpfulness for understanding the AI, and enticement to act, and both are deemed better than simple explanations; and (2) users' views about explanations are influenced by the (dis)agreement between the ML model's predictions and users' estimations of these predictions, and by the inclusion/omission of features users expect to see in explanations.

1 Introduction

The increased accuracy of Machine Learning (ML) models has led to their widespread adoption by decision makers in vital domains, such as healthcare and finance. This highlights the need for explanations of the outcomes of these models to support decision making by practitioners and end users.

To generate explanations, we adopt the human-centered view in (Biran and McKeown, 2017), whereby an explanation is "not about the model, but about the evidence that led to the prediction" (according to the model). Our explanations are aimed at non-expert users, whose goals are to obtain a basic understanding of the reasons for a prediction, and to decide on a course of action. Specifically, we generate three types of explanations, *simple*, *conservative* and *unifying*,[1] and examine their influence on the achievement of these goals.

Table 1: Features and their values for an instance in the Car Evaluation dataset (top part), and explanations for the prediction made by the AI: features and values are *italicised*, predicted outcomes appear in ***boldface italics***, and unifying information is shaded.

Feature:	**Value**	**Feature:**	**Value**
Buying price:	*high*	*Maintenance cost:*	*high*
Number of doors:	*four*	*Seating capacity:*	*four*
Luggage boot size:	*big*	*Safety rating:*	*medium*
Simple explanation			
The AI system deems this car ***acceptable*** mainly because it has a *seating capacity of four* and a *medium safety rating*.			
Conservative explanation			
Even though this car has a *high buying price*, the AI system deems this car ***acceptable*** mainly because it has a *seating capacity of four* and a *medium safety rating*. However, if this car had a *seating capacity of two*, then the AI system would deem it ***unacceptable***.			
Unifying explanation			
The AI system deems this car ***acceptable*** mainly because it has a *seating capacity of four* and a *medium safety rating*. In fact, 85 out of 100 cars with a *seating capacity of four* and a *medium safety rating* are deemed ***acceptable*** by the AI system.			

Table 1 illustrates these explanations for our ML model's prediction for an instance in the *Car Evaluation* dataset (Dua and Graff, 2017), which contains features and feature values of cars, and their acceptance status (acceptable or unacceptable).

A *simple* explanation implements Ockham's Razor. It presents the most influential feature values

[*]Work done while the author was at Monash University.

[1]These terms and their meaning are sourced from the literature on *Explanatory Virtues* (Kuhn, 1977; van Cleave, 2016).

Proceedings of the 17th International Natural Language Generation Conference, pages 103–120
September 23–27, 2024.

that lead to a predicted outcome. These explanations are the baseline in our evaluation (Section 4).

A *conservative* explanation decreases the degree to which we find an outcome surprising (increases its expectedness). It comprises a simple explanation plus a concessive-contrastive and a counterfactual component — the former acknowledges feature values that would normally yield an outcome that differs from the predicted one, and the latter mentions the fewest changes required to get the *not-predicted* outcome. These components have strong support in the *eXplainable Artificial Intelligence* (*XAI*) literature (Biran and McKeown, 2017; Guidotti et al., 2019; Maruf et al., 2023; Miller, 2019; Sokol and Flach, 2020; Stepin et al., 2020; van der Waa et al., 2018).

Finally, a *unifying* explanation conveys the extent of the coverage of a prediction to other entities — in our case, these are instances that have the same influential feature values as those of the instance at hand (but may differ with respect to other values). It comprises the simple explanation plus a component that communicates the proportion of instances with the same influential feature values and the same predicted outcome as the current instance. This type of explanation has been considered only in (Buçinca et al., 2020).

In this paper, we offer new algorithms for generating simple, conservative and unifying explanations of the outcomes of logistic regression models. These models, which are widely used in healthcare and the social sciences, are considered *transparent*, i.e., they are "interpretable by a Machine Learning expert or a statistician" (Biran and McKeown, 2017). It is important to explain the predictions of transparent models because (1) these models are commonly used as *local surrogate explainer models* that approximate neural networks for an instance of interest (Section 2); (2) transparent models are employed when the data are insufficient for neural models; and (3) even if transparent models are understandable by ML experts, they may be unclear to lay practitioners and end users.

We conducted a user study to evaluate our explanations. Our main findings are that conservative and unifying explanations are deemed largely equivalent, are liked more than simple explanations, and are deemed more complete, more helpful for understanding the AI's reasoning and more enticing to act than simple explanations. Also, users' views about explanations are influenced by the (dis)agreement between the AI's predictions and users' estimates of these predictions, and by the inclusion/omission of features users expect to see in explanations.

This paper is organised as follows. Section 2 discusses related work, Section 3 describes our explanation-generation algorithms. Our user study appears in Section 4 and its results in Section 5. Section 6 discusses key findings and future work.

2 Related research

The sub-field of XAI focuses on explaining the predictions made by ML models. In particular, neural networks have received a lot of attention, owing to their superior performance on one hand, and their opaqueness on the other hand.

Transparent models as local surrogate explainers. Linear regression, decision rules and decision trees have been used to this effect. Under linear regression, an explanation is cast as a linear combination of the input features of a model, where the coefficients are learned by perturbing the features in the local neighbourhood of an instance of interest (Ribeiro et al., 2016), or by approximating a feature's Shapley value (Kokalj et al., 2021; Lundberg and Lee, 2017). The explanations generated by these systems comprise feature attributions that represent the contribution of important features to a model's prediction. Looking at decision rules, Ribeiro et al. (2018) search for the smallest set of "anchor rules" that describes the largest part of the input space and respects a precision threshold. The works that approximate the local neighbourhood of an instance via decision trees specify this neighbourhood in different ways; they also consider contrastive and counterfactual explanations (Guidotti et al., 2019; van der Waa et al., 2018).

Transparent models in their own right. There has also been research on directly explaining the predictions of two main types of transparent models, viz decision trees and linear classifiers, such as logistic regression and linear SVMs. Decision trees differ from linear models in that in decision trees, the contributions of feature values to a prediction are contextualised in light of the contributions of other feature values, and only the features that are relevant to a prediction appear in the path from the root of the tree to that prediction. In contrast, in linear models, the contributions of feature values are independent of each other, and all the feature

values contribute to the outcome, generally to different extents.

The predictions made by decision trees are generally explained by tracing the path from the root to a predicted outcome (Guidotti et al., 2019; Stepin et al., 2020). In addition, contrastive and/or counterfactual explanations have been generated to enhance the explanations of decision tree predictions (Maruf et al., 2023; Sokol and Flach, 2020; Stepin et al., 2020). Looking at linear classifiers, Biran and McKeown (2017) incorporated unexpected effects of individual features in their explanations of the predictions of a logistic regressor, but they did not consider unexpected predictions, as done in our concessive-contrastive explanations. Ustun et al. (2019) solved a discrete optimisation problem to generate a list of actionable changes in feature values that would cause a linear classification model to yield a desired outcome. Their approach aims to provide recourse to people who have been disadvantaged by such a model, rather than conveying the fewest changes that yield a different outcome.

3 Generating Explanations

Our explanation-generation algorithms receive three main inputs: an instance $\boldsymbol{x}$, a logistic regression model denoted $f_{\boldsymbol{\beta}}$, and an outcome y predicted by the model for the instance in question; the instance $\boldsymbol{x}$ comprises features $\{x_1, \ldots, x_N\}$, each associated with a value. In this section, we specify the logistic regression classifier employed in our research, and describe algorithms that generate simple, conservative and unifying explanations for the outcomes produced by this classifier.

3.1 Logistic regression model

Since our dataset comprises only categorical features, we used a one-hot vector representation, such that the logistic regression model learns a weight for each feature value, $\{x_{1,1}, \ldots, x_{1,m_1}, \ldots, x_{N,1}, \ldots, x_{N,m_N}\}$, where m_i denotes the number of values associated with a particular feature x_i, for $i = 1, \ldots, N$.

For a multinomial classification problem (one versus the rest), this yields a model $f_{\boldsymbol{\beta}}$ parameterised by an intercept $\beta_{c,0}$ for each class c (the intercepts are collectively denoted as $\boldsymbol{\beta}_0$), and coefficients for each feature value for each class c, $\boldsymbol{\beta}_c = \{\beta_{c,1,1}, \ldots, \beta_{c,1,m_1}, \ldots, \beta_{c,N,1}, \ldots, \beta_{c,N,m_N}\}$.

For a binary classification problem, $f_{\boldsymbol{\beta}}$ contains parameters (intercept and the coefficients for each feature value) only for the positive outcome; the parameters of the negative outcome are obtained by negating the parameters for the positive outcome. The intercept represents the log odds of the positive outcome for the reference feature values — for our one-hot vector representation, this corresponds to 0 for each feature value. For instance, the intercept -1.67 in Table 2 means that a car where all feature values are absent or unknown has a probability of $\frac{e^{-1.67}}{1+e^{-1.67}} = 0.158$ of being acceptable.

Table 2: ***Classes***, *features* and *feature values* (in descending order of desirability), logistic regression coefficients and intercept for the Car Evaluation dataset; feature values of the sample car from Table 1 are shaded.

Classes	***Acceptable***, ***Unacceptable***			
Feature	**Feature values and coefficients**			
Buying price	*low*	*medium*	*high*	*very high*
	0.94	0.62	−0.45	−1.11
Maintenance cost	*low*	*medium*	*high*	*very high*
	0.68	0.58	−0.29	−0.97
Number of doors	*five*	*four*	*three*	*two*
	0.25	0.19	0.10	−0.54
Seating capacity	*four*	*> four*		*two*
	1.48	1.28		−2.76
Luggage boot size	*big*	*medium*		*small*
	0.43	0.19		−0.63
Safety rating	*high*	*medium*		*low*
	1.64	0.94		−2.58
Intercept				−1.67

3.2 Generating simple explanations

Intuitively, the feature values of interest for explaining a prediction are those having positive coefficients for that prediction. To obtain this set of feature values, we first separate the feature values with positive and negative coefficients, and then sort the feature values with positive coefficients in descending order, starting with the most positive. The simplest explanation comprises $\hat{\boldsymbol{x}}_{\text{simp}}$ – the smallest set of feature values with positive coefficients that can overcome the net effect of the feature values with negative coefficients and a negative intercept in order to yield the predicted outcome. This reasoning is formalised in Algorithm 1 (Appendix A).

As an example, consider the feature values of the Car Evaluation dataset and their coefficients in a binary logistic regression model (Table 2), and the feature values of the sample car from Table 1 (shaded in Table 2). Those with positive coefficients are: *number of doors (four), seating capacity (four), luggage boot size (big)* and *safety rating (medium)*. *Buying price (high)* and *maintenance cost (high)* have negative coefficients. After sorting the feature values with positive coefficients, we get:

seating capacity > *safety rating* > *luggage boot size* > *number of doors*. The minimal set of feature values that can overcome the intercept and the feature values with negative coefficients is $\hat{\boldsymbol{x}}_{\text{simp}} = \{$*seating capacity (four), safety rating (medium)*$\}$.

After the feature values $\hat{\boldsymbol{x}}_{\text{simp}}$ have been selected, an explanation is produced by the following programmable template: "The AI system deems this car $Phrase_{outcome}(y)$ mainly because it has $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{simp}})$", where $Phrase_{outcome}(y)$ is a function that articulates an outcome (e.g., "acceptable"), and $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{simp}})$ is a function that articulates a list of feature values (e.g., [*maintenance cost: low* ⇒ "low maintenance cost"]) in decreasing order of importance for a prediction.[2] The resultant text appears in Table 1.

3.3 Generating conservative explanations

Conservative explanations account for outcomes that appear surprising in light of background knowledge (Schupbach and Sprenger, 2011; van Cleave, 2016). For instance, this happens in the car domain when a car with a *high buying price* and *high maintenance cost* is deemed acceptable (Table 1). Our conservative explanations address such surprises by including two components: concessive-contrastive and counterfactual. The concessive-contrastive component acknowledges feature values that would normally lead to an outcome that differs from the predicted one. These feature values are overcome by the feature values in the simple explanation, which explain the surprising (predicted) outcome. The counterfactual component conveys minimal changes in feature values that would yield the outcome that was not predicted.

Algorithm 2 (Appendix A) presents our procedure for generating a conservative explanation for a prediction made by a logistic regression classifier. First, we obtain the feature values that lead to the predicted outcome, i.e., those in the simple explanation ($\hat{\boldsymbol{x}}_{\text{simp}}$); next, we derive the feature values for the concessive-contrastive component ($\hat{\boldsymbol{x}}_{\text{cc}}$); and then we determine the feature values for the counterfactual component ($\hat{\boldsymbol{x}}_{\text{cf}}$).

Concessive-contrastive component (Algorithm 4, Appendix A). We first find the feature values whose coefficients disagree with the prediction, i.e., those with negative coefficients for the classifier of class y. We then select the most influential of these feature values as follows: (i) sort the feature values with negative coefficients in ascending order, starting with the most negative; and (ii) choose the feature value with the most negative coefficient, and all feature values with coefficients within $100 \times \tau\%$ of the most negative coefficient, where τ is a tunable parameter. For our experiments, we set τ to 0.75, which means that we include feature values whose coefficients are 75% or more of the most negative coefficient. This value of τ, which was empirically obtained, enables us to balance the influence of feature values and the number of feature values included in the concessive-contrastive component of an explanation.

To illustrate, let's revisit the sample car in Table 1. As seen in Table 2, the feature values that have negative coefficients are *high buying price* (-0.45) and *high maintenance cost* (-0.29). Since $0.29 < \tau \times 0.45$, $\hat{\boldsymbol{x}}_{\text{cc}} = \{$*buying price (high)*$\}$.

Counterfactual component (Algorithm 5, Appendix A). We find the minimal number of changes in feature values that yield an unsurprising (not predicted) outcome y'[3] — this approach is appropriate for logistic regression models, which assume that features are independent.

To determine the impact of all possible changes in the value of a feature on achieving the unsurprising outcome y', for each feature, we compute the difference between the coefficient for each value not in $\boldsymbol{x}$ and the coefficient of the value in $\boldsymbol{x}$ based on the classifier for y'; this yields a list of differences denoted $\boldsymbol{\delta}_{y'}$. A positive δ means that we are moving towards the unsurprising outcome y', while a negative δ means that we are moving away from y'; hence, we consider only positive δs. To propose the minimal number of changes, we first sort the features in descending order of their maximum potential impact (largest δ), and within each feature, we sort the change in value in ascending order of δ. That is, we start with the smallest change in the maximum-impact feature.

To illustrate, consider the changes depicted in Table 3, which decrease the acceptability of our sample car. After sorting the features in descending order of their highest δ, we get: *seating capacity (4.24)* > *safety rating (3.52)* > *luggage boot size (1.06)* > *number of doors (0.73)* > *maintenance cost (0.68)* > *buying price (0.66)*. We select *seating capacity*, and start by replacing the value *four* with

[2] We eschew varying the generated text, e.g., by using Large Language Models, as this may vitiate the experiment.

[3] We minimise the number of changes, rather than the magnitude of change, because the relative importance of different features (e.g., seating capacity versus maintenance cost) and feature values depends on users' priorities.

Table 3: Changes in feature values that would make the sample car less acceptable, and "gain" towards unacceptability (δ).

Feature	Value change(s)		(δ)	(δ)
buying price	*high*	$\Rightarrow$ *very high*	(0.66)	
maintenance cost	*high*	$\Rightarrow$ *very high*	(0.68)	
number of doors	*four*	$\Rightarrow$ *three*	(0.09); *two*	(0.73)
seating capacity	*four*	$\Rightarrow >$ *four*	(0.20); *two*	(4.24)
luggage boot size	*big*	$\Rightarrow$ *medium*	(0.24); *small*	(1.06)
safety rating	*medium*	$\Rightarrow$ *low*	(3.52)	

'$>$*four*'. Since this does not change the prediction, we replace it with *two*, which makes the car unacceptable. Hence, $\hat{\boldsymbol{x}}_{\text{cf}} = \{$*seating capacity (two)*$\}$. If the car had still been acceptable, we would have proceeded to *safety rating*, and so on.

Composing the explanation. After selecting the feature values $\hat{\boldsymbol{x}}_{\text{simp}}$, $\hat{\boldsymbol{x}}_{\text{cc}}$ and $\hat{\boldsymbol{x}}_{\text{cf}}$, an explanation is produced by the following template: "Even though this car has $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{cc}})$, the AI system deems this car $Phrase_{outcome}(y)$ mainly because it has $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{simp}})$. However, if this car had $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{cf}})$, then the AI system would deem it $Phrase_{outcome}(y')$." Table 1 shows the resultant text.

3.4 Generating unifying explanations

Unifying explanations embody an inductive reasoning style. They indicate the extent of the applicability of an ML model's predictions to other entities which are similar to the instance at hand.

Algorithm 3 (Appendix A) presents our procedure for generating these explanations. First, we obtain the feature values that lead to the predicted outcome, i.e., those in the simple explanation ($\hat{\boldsymbol{x}}_{\text{simp}}$). Next, we find the $\eta_{\hat{\boldsymbol{x}}_{\text{simp}}}$ training instances that have the feature values mentioned in the simple explanation of the current instance, and determine how many of these training instances have the same predicted outcome as the current instance, $\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y}$. A unifying explanation is produced by a programmable template that presents the simple explanation followed by the proportion of $\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y}$ out of the reference training instances $\eta_{\hat{\boldsymbol{x}}_{\text{simp}}}$: "The AI system deems this car $Phrase_{outcome}(y)$ mainly because it has $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{simp}})$. In fact, $Phrase_{prop}(\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y}, \eta_{\hat{\boldsymbol{x}}_{\text{simp}}})$ cars that have $Phrase_{feature}(\hat{\boldsymbol{x}}_{\text{simp}})$ are deemed $Phrase_{outcome}(y)$ by the AI system", where $Phrase_{prop}(\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y}, \eta_{\hat{\boldsymbol{x}}_{\text{simp}}})$ is articulated as "$100 \times \frac{\eta_{\hat{\boldsymbol{x}}_{simp},y}}{\eta_{\hat{\boldsymbol{x}}_{simp}}}$ out of 100" if the ratio is less than 1, and as "all 100" otherwise. We use proportion out of a referent, rather than percentage, in line with the recommendations in (Gigerenzer, 2003); the referent is set to 100 to avoid presenting referents of different magnitudes for different cars, which may introduce a *ratio bias* (Spiegelhalter, 2017). The resultant text appears in Table 1.

4 Experimental Setup

We consider two research questions:

RQ1: How do the three types of explanations compare to each other in terms of completeness (no missing information), presence of misleading/contradictory/irrelevant information, users' understanding of the AI's reasoning for a predicted outcome, and enticement to act on the prediction (Hoffman et al., 2018), and the extent to which an explanation is liked?

RQ2: Which independent variables influence users' views about the three types of explanations?

We first describe our dataset and classifier, followed by the user study and our results.[4]

4.1 Dataset and logistic regression model

We chose the Car Evaluation dataset from the UCI Machine Learning Repository (Dua and Graff, 2017), owing to the general accessibility of its domain and concepts — this dataset has relatively few features, and users are familiar with their semantics. The difficulty faced by users when predicting the acceptability of a car pertains to understanding the combined impact of several feature values, which may have opposite effects on an outcome.

The Car Evaluation dataset was pre-processed as described in Appendix B, yielding a balanced binary dataset comprising 518 acceptable cars and 518 unacceptable cars. The dataset was split into 80% training and 20% test sets using proportional sampling.

We trained a binary logistic regression model with the features shown in Table 2, using the API provided by *scikit-learn* (Pedregosa et al., 2011); the coefficients of this model appear in Table 2. This model achieved an accuracy of 96.26% and 95.67% on the training and test set respectively. We did not cross-validate, as average classifier accuracy is tangential to this research.

4.2 User study

After signing a consent form, participants filled a demographic questionnaire and proceeded to the body of the survey.

[4]We have addressed the recommendations for human evaluation in (Howcroft et al., 2020). The experiment and data are available here.

4.2.1 Survey design

The design of the survey was similar to that in (Maruf et al., 2023). The survey began with a narrative immersion, where participants were told that they have a car dealership, and are trialing an AI system to help them predict whether a car was acceptable or unacceptable for sale at their dealership. Participants were then shown the features and values that are input to the AI, and asked which features were important to them in order to determine the acceptability of a car; this was followed by a brief account of how an AI system makes predictions (Figure 1, Appendix C). To set up a baseline for users' pre-existing beliefs, next, participants were shown a test car, and for each feature value of this car, they were asked whether it should make the car more (un)acceptable for the AI; they were then asked to estimate the AI-predicted outcome for the test car, and to enter their confidence level in this estimate.

In the main part of the survey, participants were shown four car scenarios in random order. To detect unreliable responses, we inserted an attention question after each scenario, where users had to indicate whether a neutral statement about background information in the scenario or an explanation was true or false. A short version of the Matching Familiar Figures Test (Cairns and Cammock, 1978) was given between scenarios as a filler.

Scenarios. We chose four car scenarios with diverse feature values, where a car was predicted as acceptable in two scenarios and as unacceptable in the other two. Each scenario began by showing the features of a car with their values (Table 1). For each feature value of the car, users were asked whether it should make the car more (un)acceptable for the AI; they were then asked to estimate the outcome predicted by the AI, and to indicate their confidence in this estimate (Figure 2, Appendix C). On the next page, users were shown the prediction made by the logistic regressor, and given three side-by-side explanations for this prediction: simple, conservative and unifying (Figure 3, Appendix C). The side-by-side configuration of these explanations was randomised between scenarios, but all the participants saw the same configuration for a given scenario.

Participants' views about explanations. A 7-point Likert scale was used throughout our experiment, in line with recent best practice recommendations in (van der Lee et al., 2021). Participants were asked to enter their level of agreement ('Strongly disagree': 1 to 'Strongly agree': 7) with statements about four attributes of an explanation, sourced from Hoffman et al.'s (2018) *Explanation Satisfaction Scale*: (1) it is complete, (2) it contains misleading/contradictory/irrelevant information, (3) it helps understand the AI's reasoning, and (4) it entices to act on the prediction (Figure 3, Appendix C). Participants were then asked to rate how much they liked each explanation ('Dislike a great deal': 1 to 'Like a great deal': 7), and to indicate which features that had been omitted from the explanations they expected to see, followed by an attention question (Figure 4, Appendix C).

Table 4: Descriptive statistics – two options with the most participants; domain familarity was self-rated.

Question	Option	#Part. (40)
Gender	Male / Female	23 / 15
Age	25-34 / 35-44	17 / 12
Ethnicity	Caucasian / East Asian	30 / 6
English proficiency	High	40
Education	Bachelor / Some college	16 / 14
ML expertise	Low / Medium	23 / 17
Domain familiarity	Average / Good	15 / 13

4.3 Participants

Our survey was implemented in the Qualtrics platform, and conducted on CloudResearch (Litman and Robinson, 2020) and Connect (a CloudResearch platform). Participants spent about 25 minutes on the experiment on average, and they were paid $10 USD. Their responses were validated based on their answers to the attention questions and the time they spent on the experiment, yielding 40 valid responses out of 42.[5] Table 4 shows descriptive statistics for the 40 retained participants.

5 Results

We addressed the research questions as follows. (**RQ1**) We compared the ratings given by users to the simple, conservative and unifying explanations for the four explanatory attributes and the extent to which an explanation was liked (Section 5.1). (**RQ2**) We analysed the influence of three independent variables on users' ratings of our explanation types: *acceptance status of a car (acceptable or unacceptable)*, *(dis)agreement between the outcome predicted by the AI and users' estimates of these predictions*, and *whether features expected by users were omitted from explanations* (Section 5.2). According to Lombrozo (2016), explanation length

[5]The two rejected participants scored 50% on the attention questions, while most participants scored 100%.

Table 5: Comparison between ratings of explanation types: mean (standard deviation); a lower score is better for Misleading/Contradictory/Irrelevant, and a higher score is better for the other attributes.

	Mean (standard deviation)		
Attribute	**Simple**	**Conservative**	**Unifying**
Complete	3.71 (1.72)	5.02 (1.85)	4.78 (1.79)
Misleading/...	2.12 (1.37)	2.30 (1.52)	2.14 (1.39)
Understand AI	4.43 (1.72)	5.64 (1.37)	5.58 (1.36)
Entice to act	5.13 (1.56)	5.55 (1.54)	5.59 (1.48)
Liked by users	3.40 (1.63)	5.21 (1.81)	5.18 (1.52)

affects users' views. However, in our case, length is highly correlated with explanation type, hence length was excluded from our analysis.

Statistical significance was calculated using Wilcoxon rank-sum tests for unpaired variables, and Wilcoxon signed-rank tests for paired ratings of different types of explanations. Significance was adjusted using Holm-Bonferroni correction for multiple comparisons (Holm, 1979).

5.1 Comparison between explanation types

Table 5 shows the means and standard deviations of the users' ratings of the three explanation types for the four explanatory attributes and the extent to which an explanation was liked. We performed pairwise comparisons between the ratings of the explanation types (Wilcoxon signed-rank test; statistical significances appear in Table 9, Appendix D). Our results indicate that (i) there was no difference between the explanation types in terms of misleading/contradictory/irrelevant information; (ii) conservative and unifying explanations were deemed better than simple explanations for the other three explanatory attributes and the extent to which an explanation was liked (*p-value* < 0.001); and (iii) conservative and unifying explanations were deemed equivalent for all the explanatory attributes and the extent to which an explanation was liked, but there is a trend whereby conservative explanations were deemed more complete than unifying explanations ($0.05 <$ *p-value* < 0.1).

Finding 1 *Conservative and unifying explanations are deemed better than simple explanations, and unifying explanations are deemed largely equivalent to conservative explanations.*

Our finding about conservative versus simple explanations is consistent with the results in (Maruf et al., 2023) about contrastive versus simple explanations. However, our finding about unifying versus simple explanations is somewhat at odds with Buçinca et al.'s (2020), where simple explanations were preferred for decision-making tasks.

5.2 Effect of independent variables

Acceptance status of a car. Even though the acceptance status of a car is domain specific, we consider this variable, as the notions of acceptance and rejection are general. We split the participant responses according to the predicted outcome (acceptable or unacceptable), and for each outcome, we compared users' ratings of each pair of explanation types. Our results indicate that the statistical significances obtained from the initial pairwise comparisons between explanation types (Section 5.1) largely held (Table 10, Appendix D), except for enticement to act on the AI's prediction of an unacceptable outcome, where conservative and unifying explanations were deemed equivalent to simple explanations. Also, the trend whereby conservative explanations are deemed more complete than unifying explanations is exhibited only for unacceptable cars.

Finding 2 *The predicted outcome had little effect on the results reported in Finding 1.*

(Dis)agreement between the AI's predictions and users' estimations of these predictions. Maruf et al. (2023) found that contrastive explanations which address users' potential expectations are particularly valuable when an AI's predictions (made by a decision tree) disagree with users' estimates of these predictions. Here, we determine whether this finding holds for conservative explanations of the predictions of a logistic regressor, which have a contrastive aspect, and whether it extends to unifying explanations. To this effect, we compare users' ratings of each pair of explanation types for *AI Predict = User Predict* and *AI Predict* $\neq$ *User Predict* (84% and 16% of the responses respectively).

Our results indicate that the statistical significances obtained from the initial pairwise comparisons between explanation types (Section 5.1) held when the AI's predictions agreed with users' estimates of these predictions (Table 6). However, when they disagreed, conservative and unifying explanations were statistically significantly better than simple explanations only for liking an explanation (last row of Table 6). This result, which is not in line with the findings in (Maruf et al., 2023) for contrastive explanations, could be partially attributed to the small sample size of *AI Predict* $\neq$ *User Predict* (35 samples).

Finding 3 *Conservative and unifying explanations are deemed better than simple explanations when*

Table 6: Effect of (dis)agreement between ML model predictions and users' estimates of these predictions on ratings of explanations: mean (standard deviation) and statistical significance (Wilcoxon signed-rank test); a lower score is better for Misleading/Contradictory/Irrelevant, and a higher score is better for the other attributes; statistically significant results are **boldfaced**.

		Mean (standard deviation)			Statistical Significance		
Attribute	**AI Predict vs User Predict**	**Simple**	**Conservative**	**Unifying**	Simple vs Conservative	Simple vs Unifying	Unifying vs Conservative
Complete	AI=User	3.68 (1.70)	5.06 (1.81)	4.78 (1.76)	**6.88E-10**	**6.42E-10**	0.187
	AI≠User	3.84 (1.86)	4.80 (2.08)	4.76 (1.98)	0.819	0.826	1
Misleading/Contradictory/Irrelevant	AI=User	2.05 (1.29)	2.21 (1.42)	2.06 (1.34)	1	1	1
	AI≠User	2.52 (1.68)	2.76 (1.90)	2.60 (1.63)	1	1	1
Understand AI's reasoning	AI=User	4.41 (1.68)	5.69 (1.34)	5.64 (1.30)	**6.89E-12**	**3.31E-14**	1
	AI≠User	4.52 (1.98)	5.40 (1.52)	5.24 (1.61)	0.777	1	1
Entice to act	AI=User	5.28 (1.44)	5.71 (1.40)	5.73 (1.33)	**2.50E-03**	**4.87E-05**	1
	AI≠User	4.32 (1.90)	4.68 (1.97)	4.84 (2.01)	1	1	1
Liked by users	AI=User	3.46 (1.62)	5.25 (1.79)	5.20 (1.53)	**1.56E-10**	**6.60E-15**	1
	AI≠User	3.04 (1.64)	4.96 (1.94)	5.00 (1.50)	**0.024**	**4.99E-03**	1

the AI's predictions agree with users' estimates of these predictions, and are deemed at least as good as simple explanations when the predictions disagree.

Features omitted from an explanation. Dale and Reiter (1995) showed that descriptions with superfluous attributes were preferred to minimal descriptions. This prompted us to investigate whether omitting features that are not influential, but are expected by users, affects users' views about explanations. To this effect, we asked participants to point out features they expected to see, but were omitted from the explanations for each scenario. At least 75% of the participants selected *buying price* when it was omitted, and each omitted feature was chosen by at least six participants (Table 11, Appendix D).

We then compared the ratings of explanations that omitted expected features with the ratings of explanations that had no omissions. Since conservative explanations contain the largest number of features, and simple and unifying explanations contain only features with values that have a positive impact on a predicted outcome, we considered only conservative explanations in our analysis. We found that explanations that omit features expected by users were statistically significantly less liked and deemed less complete than explanations that include all expected features (Wilcoxon rank-sum test, $p\text{-value} < 0.05$; Table 7); and there is a trend whereby explanations that omit expected features were deemed to be more misleading/contradictory/irrelevant than explanations that have no omissions ($0.05 < p\text{-value} < 0.1$). These results indicate that users may perceive some domain-specific features to be essential, regardless of their influence on the outcome, and omitting these features from explanations adversely affects users' views.

Table 7: Effect of omitted feature values on ratings of conservative explanations: mean (std. dev.) and statistical significance (Wilcoxon rank-sum test); a lower score is better for Misleading/Contradictory/Irrelevant, and a higher score is better for the other attributes; statistically significant results are **boldfaced**, and trends ($0.05 < p\text{-value} < 0.1$) are *italicised*.

	Mean (std. dev.)		Stat. Sig.
Attribute	**Omitted**	**Not omitted**	Omit vs Not omit
Complete	4.84 (1.88)	5.76 (1.52)	**0.027**
Misleading/...	2.42 (1.56)	1.76 (1.14)	*0.064*
Understand AI	5.58 (1.34)	5.90 (1.49)	0.121
Entice to act	5.48 (1.54)	5.83 (1.53)	0.121
Liked by users	5.05 (1.84)	5.86 (1.56)	**0.022**

Finding 4 *Explanations that omit expected features are liked less and are deemed less complete than explanations that have no such omissions.*

6 Conclusion

We have offered algorithms that generate simple, conservative and unifying explanations for predictions made by a logistic regressor; and we reported the results of a user study where we evaluated these explanations in terms of the extent to which they were liked and four explanatory attributes, viz completeness, presence of misleading/contradictory/irrelevant information, helpfulness to understand the AI's reasoning, and enticement to act on the AI's prediction. We also considered the influence of three independent variables on users' views about our explanations, viz *predicted outcome*, *(dis)agreement between the AI's prediction and users' estimates of these predictions*, and *presence/absence of features users expect to see in explanations*.

Comparison between explanation types. Our results show that conservative and unifying explanations are better liked than their simple counterparts, and are deemed more complete, more helpful to understand the AI's reasoning, and more enticing to act on the AI's prediction; and that unifying explanations are deemed largely equivalent to conservative explanations. In the future, it would be interesting to compare an explanation that combines conservative and unifying explanations with each of these explanation types.

Effect of independent variables. Firstly, the outcome predicted by the AI has little effect on users' views about our explanations.

Second, conservative and unifying explanations are deemed better than simple explanations when the AI's predictions agree with users' estimates of these predictions. However, when they disagree, conservative and unifying explanations are only liked better than simple explanations, and are deemed equivalent for the other attributes. This result may be partially attributed to the small number of data points for disagreement. In addition, these findings with respect to conservative explanations, which have a contrastive component, are at odds with those in (Maruf et al., 2023), where contrastive explanations of decision-tree predictions were particularly favoured when the AI's predictions and users' estimates of these predictions disagreed. This suggests that the factors that affect users' views about explanations may be more nuanced than simply having a contrastive aspect, e.g., whether a contrastive component explicitly mentions the expectations it is addressing, as done in (Maruf et al., 2023).

Finally, users have domain-specific expectations about features that should appear in explanations, regardless of their effect on the outcome, and not meeting these expectations adversely affects users' views about explanations.

Limitations and future work

User study. We could not recruit real users who were personally engaged with our car-dealership setting. This is a well-known problem in evaluating NLG systems, which we tried to mitigate by using a generally accessible domain, and a narrative immersion at the start of our experiment.

Dataset and algorithms. Our dataset has only categorical features, which are handled by our one-hot encoding. In the future, we will adapt our algorithms to numerical and ordinal features.

Our dataset comprises six variables, each with 3-4 values. This relatively small number is consistent with the state-of-the-art for generating textual explanations of the outcomes of transparent ML models (Maruf et al., 2023; Stepin et al., 2020). However, in the future, our explanation-generation algorithms should be adapted to handle datasets with a large number of features — even though our algorithms select feature values with the highest impact, it is possible that when the feature set is large, the generated explanations could become quite lengthy.

Our algorithms for generating simple, concessive and counterfactual explanations are linear in the number of feature values, except for the sorting steps of positive or negative coefficients. Our algorithm for generating unifying explanations examines the training instances in the dataset to determine the model's predictions for instances with the same feature values as the instance at hand. However, sampling can be used, instead of examining the entire training set.

Our algorithm for generating unifying explanations is model agnostic, while the other algorithms were developed for logistic regressors. However, these algorithms are directly applicable to other feature-attribution models, and are generalisable to linear classifiers that use linear discriminant functions, such as perceptrons and linear SVMs, and log-linear models, such as Naïve Bayes.

Communicative goals and uncertainty. We considered two user goals: understanding the AI's reasoning and acting on its prediction. However, ML models are not 100% accurate, so another important goal is to enable users to determine the trustworthiness of a prediction (Buçinca et al., 2020; Cau et al., 2023). This goal is related to another limitation of our work, viz our explanations omit information about the accuracy of an ML model — an issue that is investigated in (Zukerman and Maruf, 2024).

Acknowledgments

This research was supported in part by grant DP190100006 from the Australian Research Council. Ethics approval for the user study was obtained from Monash University Human Research Ethics Committee (ID-24208). The authors are grateful to David Evans and Enes Makalic for their assistance in verifying our statistical analysis.

References

O. Biran and K. McKeown. 2017. Human-centric justification of Machine Learning predictions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI 2017, pages 1461–1467, Melbourne, Australia.

Z. Buçinca, P. Lin, K.Z. Gajos, and E. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, pages 454–464, Cagliari, Italy.

E. Cairns and T. Cammock. 1978. Development of a more reliable version of the matching familiar figures test. *Developmental Psychology*, 14(5):555.

F.M. Cau, H. Hauptmann, L.D. Spano, and N. Tintarev. 2023. Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, page 251–263, Sydney, Australia.

R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18(2):233–263.

D. Dua and C. Graff. 2017. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

G. Gigerenzer. 2003. *Reckoning with risk: Learning to live with uncertainty*. Penguin Books Ltd.

R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23.

R.R. Hoffman, S.T. Mueller, G. Klein, and J. Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

S. Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

D.M. Howcroft, A. Belz, M.A. Clinciu, D. Gkatzia, S.A. Hasan, S. Mahamood, S. Mille, E. Van Miltenburg, S. Santhanam, and V. Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, INLG 2020, pages 169–182, Dublin, Ireland.

E. Kokalj, B. Škrlj, N. Lavrač, S. Pollak, and M. Robnik-Šikonja. 2021. BERT meets Shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21, Online.

T. Kuhn. 1977. Objectivity, value judgment, and theory choice. In *The Essential Tension*. Chicago University Press.

L. Litman and J. Robinson. 2020. *Conducting online research on Amazon Mechanical Turk and beyond*. Sage Publications.

T. Lombrozo. 2016. Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10):748–759.

S.M. Lundberg and S-I. Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777, Long Beach, California.

S. Maruf, I. Zukerman, E. Reiter, and G. Haffari. 2023. Influence of context on users' views about explanations for decision-tree predictions. *Computer Speech & Language*, 81:101483.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

M.T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the ACM/SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD'16, pages 1135–1144, San Francisco, California.

M.T. Ribeiro, S. Singh, and C. Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI-18, pages 1527–1535, New Orleans, Louisiana.

J.N. Schupbach and J. Sprenger. 2011. The logic of explanatory power. *Philosophy of Science*, 78(1):105–127.

K. Sokol and P. Flach. 2020. One explanation does not fit all: The promise of interactive explanations for Machine Learning transparency. *Künstliche Intelligenz*, 34:235–250.

D. Spiegelhalter. 2017. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4(1):31–60.

I. Stepin, J.M. Alonso, A. Catala, and M. Pereira. 2020. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *Proceedings of the IEEE World Congress on Computational Intelligence*, WCCI, pages 1–8, Glasgow, Scotland.

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

M. van Cleave. 2016. *Introduction to Logic and Critical Thinking*. Lansing Community College.

C. van der Lee, A. Gatt, E. van Miltenburg, and E.J. Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:1–24.

J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, and M. Neerincx. 2018. Contrastive explanations with local foil trees. In *Proceedings of the ICML-18 Workshop on Human Interpretability in Machine Learning*, WHI'18, pages 41–46, Stockholm, Sweden.

L. Zhang, T. Geisler, H. Ray, and Y. Xie. 2022. Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, 49(13):3257–3277.

I. Zukerman and S. Maruf. 2024. Communicating uncertainty in explanations of the outcomes of machine learning models. In *Proceedings of the 17th International Conference on Natural Language Generation*, INLG 2024, Tokyo, Japan.

A Algorithms

Algorithm 1 Generate Simple Explanation

Input: $\boldsymbol{x}$: the feature values of an instance in the test set; $f_{\boldsymbol{\beta}}$: the logistic regression model; y: the model's prediction for instance $\boldsymbol{x}$; N: the number of features in the dataset

procedure GENERATESIMPLE($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y, N)
- ▷ get coefficients corresponding to the feature values in $\boldsymbol{x}$ from the classifier of class y
- $\boldsymbol{\beta}_y^{\boldsymbol{x}} \leftarrow$ getcoeff($\boldsymbol{x}$, $\boldsymbol{\beta}_y$)
- ▷ separate the indices of feature values with positive and negative coefficients
- $\boldsymbol{Index}^{pos} \leftarrow \emptyset$, $\boldsymbol{Index}^{neg} \leftarrow \emptyset$
- $\boldsymbol{\beta}_y^{\boldsymbol{x}_{pos}} \leftarrow \emptyset$ ▷ positive coefficients in $\boldsymbol{\beta}_y^{\boldsymbol{x}}$
- $\boldsymbol{x}_{neg} \leftarrow \emptyset$ ▷ feature values with negative coefficients
- **for** $i \leftarrow 1$ to N **do**
  - **if** $\beta_{y,i}^{\boldsymbol{x}} < 0$ **then**
    - ▷ collect indices of feature values with negative coefficients
    - $\boldsymbol{Index}^{neg} \leftarrow$ append($\boldsymbol{Index}^{neg}$, i)
    - ▷ collect feature values with negative coefficients
    - $\boldsymbol{x}_{neg} \leftarrow$ append($\boldsymbol{x}_{neg}$, x_i)
  - **else**
    - ▷ collect indices of feature values with positive coefficients
    - $\boldsymbol{Index}^{pos} \leftarrow$ append($\boldsymbol{Index}^{pos}$, i)
    - ▷ collect positive coefficients
    - $\boldsymbol{\beta}_y^{\boldsymbol{x}_{pos}} \leftarrow$ append($\boldsymbol{\beta}_y^{\boldsymbol{x}_{pos}}$, $\beta_{y,i}^{\boldsymbol{x}}$)
  - **end if**
- **end for**
- ▷ sort $\boldsymbol{Index}^{pos}$ in descending order of the positive coefficients
- $\boldsymbol{Index}^{\textit{pos-sorted}} \leftarrow$ sort($\boldsymbol{Index}^{pos}$, $\boldsymbol{\beta}_y^{\boldsymbol{x}_{pos}}$, descend)
- $i \leftarrow 1$
- ▷ get the feature value with the most positive coefficient
- $\hat{\boldsymbol{x}}_{\text{simp}} \leftarrow$ {get-feature-value($Index_i^{\textit{pos-sorted}}$, $\boldsymbol{x}$)}
- ▷ iteratively add feature values with positive coefficients until prediction y is obtained
- **while** $f_{\boldsymbol{\beta}}(\boldsymbol{x}_{neg} \cup \hat{\boldsymbol{x}}_{\text{simp}}) \neq y$ **do**
  - $i \leftarrow i + 1$
  - $\hat{\boldsymbol{x}}_{\text{simp}} \leftarrow$ append($\hat{\boldsymbol{x}}_{\text{simp}}$, get-feature-value($Index_i^{\textit{pos-sorted}}$, $\boldsymbol{x}$))
- **end while**
- **return** $\hat{\boldsymbol{x}}_{\text{simp}}$

end procedure

Algorithm 2 Generate Conservative Explanation

Input: $\boldsymbol{x}$: the feature values of an instance in the test set; $f_{\boldsymbol{\beta}}$: the logistic regression model; y: the model's prediction for instance $\boldsymbol{x}$; y': an alternative class ($\neq y$) for the counterfactual; N: the number of features in the dataset; τ: a threshold for selecting the concessive feature values; *feature-values*: the list of feature values in the dataset $\{x_{1,1}, \ldots, x_{1,m_1}, \ldots, x_{N,1}, \ldots, x_{N,m_N}\}$

procedure GENERATECONSERVATIVE($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y, y', N, τ, *feature-values*)
- ▷ get coefficients corresponding to the feature values in $\boldsymbol{x}$ from the classifier of class y
- $\boldsymbol{\beta}_y^{\boldsymbol{x}} \leftarrow$ getcoeff($\boldsymbol{x}$, $\boldsymbol{\beta}_y$)
- $\hat{\boldsymbol{x}}_{\text{simp}} \leftarrow$ GENERATESIMPLE($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y, N) ▷ Algorithm 1
- $\hat{\boldsymbol{x}}_{\text{cc}} \leftarrow$ GENERATECONCESSIVE($\boldsymbol{x}$, N, τ, $\boldsymbol{\beta}_y^{\boldsymbol{x}}$) ▷ Algorithm 4
- $\hat{\boldsymbol{x}}_{\text{cf}} \leftarrow$ GENERATECOUNTERFACTUAL($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y', *feature-values*) ▷ Algorithm 5
- **return** $\hat{\boldsymbol{x}}_{\text{cc}}$, $\hat{\boldsymbol{x}}_{\text{simp}}$, $\hat{\boldsymbol{x}}_{\text{cf}}$

end procedure

Algorithm 3 Generate Unifying Explanation

Input: $\boldsymbol{x}$: the feature values of an instance in the test set; $f_{\boldsymbol{\beta}}$: the logistic regression model; y: the model's prediction for instance $\boldsymbol{x}$; N: the number of features in the dataset; D: a set of training instances

procedure GENERATEUNIFYING($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y, N, D)
- $\hat{\boldsymbol{x}}_{\text{simp}} \leftarrow$ GENERATESIMPLE($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y, N) ▷ Algorithm 1
- ▷ find the instances in D with the same feature values as $\hat{\boldsymbol{x}}_{\text{simp}}$ and the same prediction
- $\eta_{\hat{\boldsymbol{x}}_{\text{simp}}} = 0$ ▷ same feature values
- $\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y} = 0$ ▷ same feature values and prediction
- **for** each $\hat{\boldsymbol{x}} \in D$ **do**
  - **if** $\hat{\boldsymbol{x}}_{\text{simp}} \subseteq \hat{\boldsymbol{x}}$ **then**
    - $\eta_{\hat{\boldsymbol{x}}_{\text{simp}}} = \eta_{\hat{\boldsymbol{x}}_{\text{simp}}} + 1$
    - **if** $f_{\boldsymbol{\beta}}(\hat{\boldsymbol{x}}) = y$ **then**
      - $\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y} = \eta_{\hat{\boldsymbol{x}}_{\text{simp}},y} + 1$
    - **end if**
  - **end if**
- **end for**
- **return** $\hat{\boldsymbol{x}}_{\text{simp}}$, $\eta_{\hat{\boldsymbol{x}}_{\text{simp}},y}$, $\eta_{\hat{\boldsymbol{x}}_{\text{simp}}}$

end procedure

Algorithm 4 Generate Concessive Explanation

$\boldsymbol{x}$: the feature values of an instance in the test set
N: the number of features in the dataset
τ: a threshold for selecting the concessive feature values
$\boldsymbol{\beta}_y^{\boldsymbol{x}}$: coefficients corresponding to the feature values in $\boldsymbol{x}$ from the classifier of class y
procedure GENERATECONCESSIVE($\boldsymbol{x}$, N, τ, $\boldsymbol{\beta}_y^{\boldsymbol{x}}$)
    ▷ get the indices and corresponding coefficients of feature values with negative coefficients
    $\boldsymbol{Index}^{neg} \leftarrow \emptyset$, $\boldsymbol{\beta}_y^{\boldsymbol{x}_{neg}} \leftarrow \emptyset$
    for $i \leftarrow 1$ to N **do**
        if $\beta_{y,i}^{\boldsymbol{x}} < 0$ **then**
            ▷ collect indices of feature values with negative coefficients
            $\boldsymbol{Index}^{neg} \leftarrow$ append($\boldsymbol{Index}^{neg}$, i)
            ▷ collect negative coefficients
            $\boldsymbol{\beta}_y^{\boldsymbol{x}_{neg}} \leftarrow$ append($\boldsymbol{\beta}_y^{\boldsymbol{x}_{neg}}$, $\beta_{y,i}^{\boldsymbol{x}}$)
        end if
    end for
    ▷ sort $\boldsymbol{Index}^{neg}$ in ascending order of the negative coefficients
    $\boldsymbol{Index}^{\textit{neg-sorted}} \leftarrow$ sort($\boldsymbol{Index}^{neg}$, $\boldsymbol{\beta}_y^{\boldsymbol{x}_{neg}}$, ascend)
    ▷ get the feature value with the most negative coefficient
    $\hat{\boldsymbol{x}}_{\text{cc}} \leftarrow$ {get-feature-value($Index_1^{\textit{neg-sorted}}$, $\boldsymbol{x}$)}
    ▷ get the feature values whose coefficients $\geq \tau \times$[the most negative coefficient]
    for $i \leftarrow 2$ to $||\boldsymbol{Index}^{\textit{neg-sorted}}||$ **do**
        if $|\beta_{y,i}^{\boldsymbol{x}_{neg}}| \geq |\tau \times \beta_{y,1}^{\boldsymbol{x}_{neg}}|$ **then**
            $\hat{\boldsymbol{x}}_{\text{cc}} \leftarrow$ append($\hat{\boldsymbol{x}}_{\text{cc}}$, get-feature-value($Index_i^{\textit{neg-sorted}}$, $\boldsymbol{x}$))
        else
            break
        end if
    end for
    return $\hat{\boldsymbol{x}}_{\text{cc}}$
end procedure

Algorithm 5 Generate Counterfactual Explanation

$\boldsymbol{x}$: the feature values of an instance in the test set
$f_{\boldsymbol{\beta}}$: the logistic regression model
y': an alternative class ($\neq y$) for the counterfactual
feature-values: the list of feature values in the dataset $\{x_{1,1}, \ldots, x_{1,m_1}, \ldots, x_{N,1}, \ldots, x_{N,m_N}\}$
procedure GENERATECOUNTERFACTUAL($\boldsymbol{x}$, $f_{\boldsymbol{\beta}}$, y', *feature-values*)
    ▷ for each feature, compute the difference between the coefficient for each feature value not in $\boldsymbol{x}$ and the coefficient of the feature value in $\boldsymbol{x}$ based on the classifier of y'
    $\boldsymbol{\delta}_{y'} \leftarrow$ compute-diff-coeff($\boldsymbol{x}$, $\boldsymbol{\beta}_{y'}$, *feature-values*)
    ▷ sort the features in descending order of their maximum positive impact on y', and for each feature, sort the values in ascending order of their positive impact on y'
    $\boldsymbol{x}_{order} \leftarrow$ sort-feature-values-positive($\boldsymbol{x}$, $\boldsymbol{\delta}_{y'}$, *feature-values*)
    $\boldsymbol{x}_{new} \leftarrow \boldsymbol{x}$
    $\hat{\boldsymbol{x}}_{\text{cf}} \leftarrow \emptyset$ ▷ the counterfactual feature values
    ▷ replace a current feature value with a different one until the outcome switches to y'
    for x_j in $\boldsymbol{x}_{order}$ **do**
        $\boldsymbol{x}_{new} \leftarrow$ replace-feature-value($\boldsymbol{x}_{new}$, x_j)
        if $f_{\boldsymbol{\beta}}(\boldsymbol{x}_{new}) = y'$ **then**
            ▷ find the feature values in $\boldsymbol{x}_{new}$ that are different from those in $\boldsymbol{x}$
            $\hat{\boldsymbol{x}}_{\text{cf}} \leftarrow$ get-different-values($\boldsymbol{x}_{new}$, $\boldsymbol{x}$)
            break
        end if
    end for
    ▷ if the value of a feature in $\hat{\boldsymbol{x}}_{\text{cf}}$ is not the highest impact one, add the higher impact values of that feature to $\hat{\boldsymbol{x}}_{\text{cf}}$
    $\hat{\boldsymbol{x}}_{\text{cf}} \leftarrow$ append($\hat{\boldsymbol{x}}_{\text{cf}}$, get-higher-impact-feature-values($\hat{\boldsymbol{x}}_{\text{cf}}$, $\boldsymbol{x}_{order}$))
    return $\hat{\boldsymbol{x}}_{\text{cf}}$
end procedure

B The Car Evaluation Dataset

This dataset, sourced from (Dua and Graff, 2017), has 1728 instances and four classes – unacceptable, acceptable, good and very good, with 70% of the instances (1210 cars) being unacceptable. In line with our previous work (Maruf et al., 2023), we decided to generate a balanced binary classification dataset.[6] This was done by (i) merging the instances from three classes ('acceptable', 'good' and 'very good') into one class called 'acceptable', which comprises 518 instances; and (ii) randomly removing 692 instances from the unacceptable class, which yields 518 unacceptable instances. We then split these data into 80% training and 20% test sets using proportional sampling (the final class breakdown of the training and test sets appears in Table 8).

Table 8: Breakdown of classes for the training and test sets in the Car Evaluation dataset.

Partition	Unacceptable	Acceptable	Total
Training	416	412	828
Test	102	106	208
Total	518	518	1036

[6] Recall that our algorithms rely on the values of the coefficients generated by a logistic regression model, hence they also apply to unbalanced datasets — a cost-sensitive logistic regressor (Zhang et al., 2022) can be used for such datasets.

C Screenshots from the experiment

Background

Artificial Intelligence (AI) systems are used to generate predictions in different domains, such as health, finance and industry. For example, the AI system used in this study predicts whether a particular car is acceptable or unacceptable to a potential customer.

We are developing a computer system that automatically generates explanations for the predictions made by this AI system. The objective of our study is to find out which types of explanations people find useful in order to understand and act on the predictions of the AI system. We would appreciate your help in making this determination.

The car sales domain

Pretend that you are a car dealer who is offered cars for sale by different manufacturers. You need to determine whether you will be able to sell these cars to your customer base. If so, you would deem these cars acceptable, otherwise they would be unacceptable. To help you make these decisions, you are trialing a state-of-the-art AI system that predicts whether a car is **acceptable** or **unacceptable**. The AI system makes these predictions based on the decisions made by your customers in the past and the six car features in the table below. The accuracy of the AI system in predicting the acceptability of a car is 96%.

*Car features and their possible values from left (make a car more **acceptable** to your customers) to right (make a car more **unacceptable** to your customers).*

Feature	Possible values			
	More acceptable			More unacceptable
Buying price	Low	Medium	High	Very high
Maintenance cost	Low	Medium	High	Very high
Number of doors	Five	Four	Three	Two
Seating capacity	More than four	Four		Two
Size of luggage boot	Big	Medium		Small
Safety rating	High	Medium		Low

Which of the following features are **important to you as a car dealer** to determine the acceptability of a car? Select all that apply.

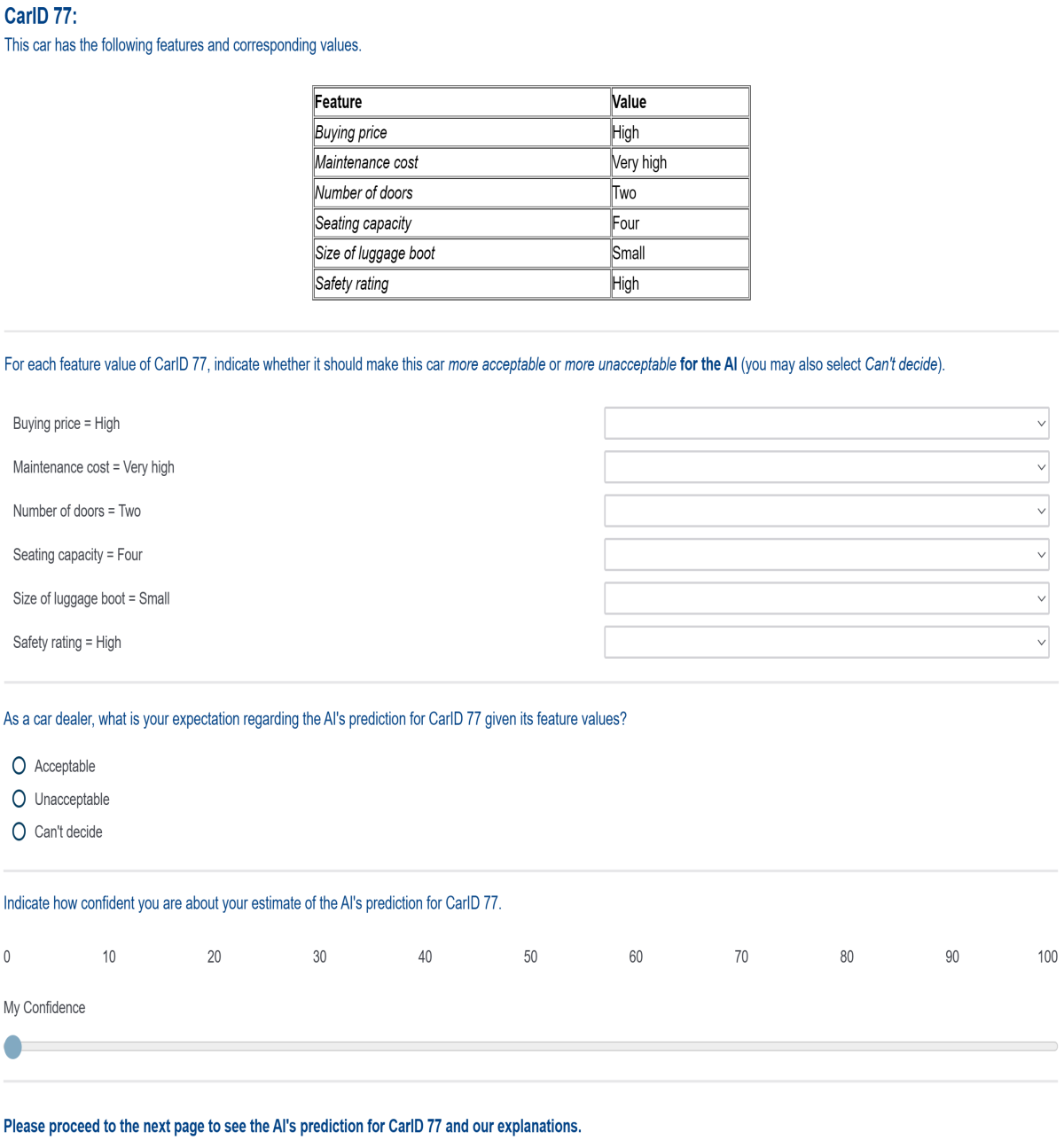
Buying price	Maintenance cost	Number of doors	Seating capacity	Size of luggage boot	Safety rating	None of these
☐	☐	☐	☐	☐	☐	☐

AI systems make predictions based on trends and patterns they have learned from large amounts of data. Therefore, the reasoning of AI systems may differ from our intuitions, which are normally based on our personal experience. In addition, for each situation, an AI system considers the importance of a feature value relative to other feature values, and hence may determine that some feature values have a higher importance in some situations and a lower importance in other situations. For example, if a car has a seating capacity of four people, having a low buying price may be deemed very important by the AI system. In contrast, the AI system may consider the buying price to be less important if the car has a seating capacity of only two people.

Going forward, please bear in mind that our generated explanations are based on the reasoning of our AI system, and may **not** reflect what you consider important for the acceptability or unacceptability of a car.

Before we describe the main experiment, we want to establish a baseline of your expectations regarding the AI's predictions (initially, these expectations are likely to be based on your opinions as a car dealer). To do this, we will show you the feature values of a **test car** and ask your expectation about whether an AI should deem this car acceptable or unacceptable, and which feature values should be considered important for this decision. Your answers will **not** affect our perceptions about you.

Figure 1: Background information; narrative immersion for the survey; features and feature values of a car; description of the reasoning of AI systems; preamble to the experiment.



CarID 77:

This car has the following features and corresponding values.

Feature	Value
Buying price	High
Maintenance cost	Very high
Number of doors	Two
Seating capacity	Four
Size of luggage boot	Small
Safety rating	High

For each feature value of CarID 77, indicate whether it should make this car *more acceptable* or *more unacceptable* **for the AI** (you may also select *Can't decide*).

Buying price = High

Maintenance cost = Very high

Number of doors = Two

Seating capacity = Four

Size of luggage boot = Small

Safety rating = High

As a car dealer, what is your expectation regarding the AI's prediction for CarID 77 given its feature values?

- Acceptable
- Unacceptable
- Can't decide

Indicate how confident you are about your estimate of the AI's prediction for CarID 77.

0 10 20 30 40 50 60 70 80 90 100

My Confidence

Please proceed to the next page to see the AI's prediction for CarID 77 and our explanations.

Figure 2: First page of a car in the main survey: background information about the car; question about whether the feature values of the car should make it more (un)acceptable for the AI; question about estimating the AI's prediction and indicating the confidence level if the estimated outcome is 'acceptable' or 'unacceptable'.

CarID 77:

This car has the following features and corresponding values.

Feature	Value
Buying price	High
Maintenance cost	Very high
Number of doors	Two
Seating capacity	Four
Size of luggage boot	Small
Safety rating	High

Based on the feature values of CarID 77, our AI system deems it ***unacceptable***.

Below you will see three explanations generated by our system. Please note that these explanations have been generated in advance, and are **not** tailored to your expectations of the feature values. Also, recall that for each situation, an AI system considers the importance of a feature value relative to other feature values, and hence may determine that some feature values have a higher importance in some situations and a lower importance in other situations. Feature values that are not so important may be omitted from an explanation.

With reference to Explanations A, B and C, indicate the extent to which you agree with the statements below **in your role of car dealer**.

	Explanation A The AI system deems this car ***unacceptable*** mainly because it has • a *very high maintenance cost* and • a *small luggage boot*.							**Explanation B** The AI system deems this car ***unacceptable*** mainly because it has • a *very high maintenance cost* and • a *small luggage boot*. In fact, *75 out of 100 cars* that have a *very high maintenance cost* and a *small luggage boot* are deemed ***unacceptable*** by the AI system.							**Explanation C** Even though this car has a *high safety rating* and a *seating capacity of four people*, the AI system still deems this car ***unacceptable***, mainly because it has • a *very high maintenance cost* and • a *small luggage boot*. However, if this car had a *low or medium maintenance cost*, then the AI system would deem it ***acceptable***.						
	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree	Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
This explanation helps me understand the reasoning of the AI system.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
This explanation has misleading, contradictory or irrelevant information.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
This explanation is complete (it is not missing information).	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Based on this explanation, I would ***not accept*** this car.	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Figure 3: Second page of a car in the main survey (top section): background information about the car (repeated); model prediction; simple explanation (A), unifying explanation (B) and conservative explanation (C) for this car; rating scales for explanatory attributes.

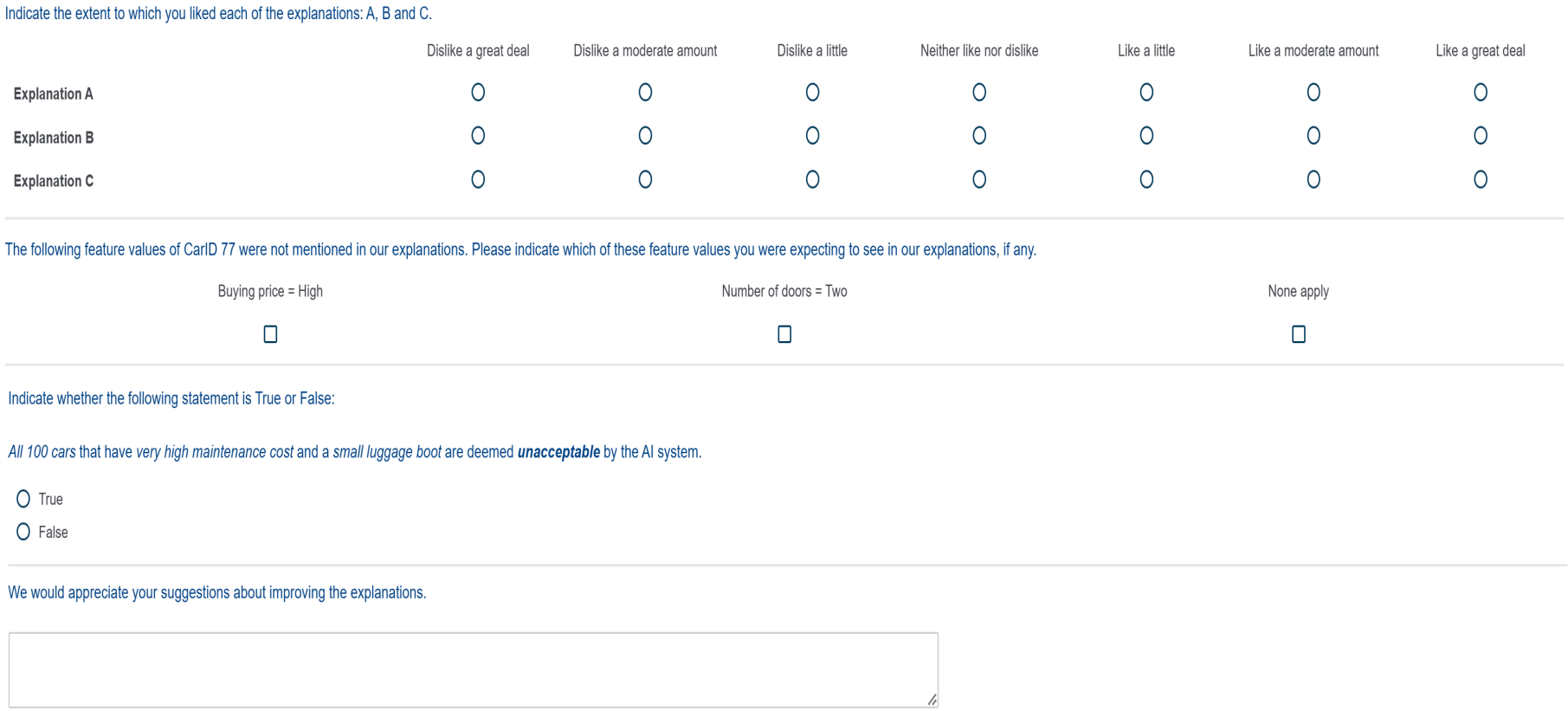


Figure 4: Second page of a car in the main survey (bottom section): rating scales for how much each explanation is liked; user expectations about feature values omitted from the explanations; attention question; request for suggestions.

Table 9: Comparison between ratings of explanation types: mean (standard deviation) of ratings, and statistical significance (Wilcoxon signed-rank test); a lower score is better for Misleading/Contradictory/Irrelevant, and a higher score is better for the other attributes; statistically significant results are **boldfaced**, and trends ($0.05 < p\text{-}value < 0.1$) are *italicised*.

	Mean (standard deviation)			Statistical Significance		
Attribute	**Simple**	**Conservative**	**Unifying**	Simple vs Conservative	Simple vs Unifying	Unifying vs Conservative
Complete	3.71 (1.72)	5.02 (1.85)	4.78 (1.79)	**6.73E-11**	**5.46E-11**	*0.084*
Misleading/Contradictory/Irrelevant	2.12 (1.37)	2.30 (1.52)	2.14 (1.39)	1	1	1
Understand AI's reasoning	4.43 (1.72)	5.64 (1.37)	5.58 (1.36)	**5.06E-13**	**1.08E-14**	1
Entice to act	5.13 (1.56)	5.55 (1.54)	5.59 (1.48)	**8.56E-04**	**1.31E-05**	1
Liked by users	3.40 (1.63)	5.21 (1.81)	5.18 (1.52)	**3.58E-13**	**3.30E-15**	1

Table 10: Effect of the acceptance status of a car on ratings of explanation types: mean (standard deviation) of ratings, and statistical significance (Wilcoxon signed-rank test); a lower score is better for Misleading/Contradictory/Irrelevant, and a higher score is better for the other attributes; statistically significant results are **boldfaced**, and trends ($0.05 < p\text{-}value < 0.05$) are *italicised*.

		Mean (standard deviation)			Statistical Significance		
Attribute	**Acceptance Status**	**Simple**	**Conservative**	**Unifying**	Simple vs Conservative	Simple vs Unifying	Unifying vs Conservative
Complete	Acceptable	4.01 (1.62)	5.21 (1.84)	5.21 (1.60)	**1.89E-04**	**9.37E-06**	1
	Unacceptable	3.40 (1.77)	4.82 (1.84)	4.35 (1.86)	**3.14E-06**	**6.25E-05**	*0.057*
Misleading/ Contradictory/Irrelevant	Acceptable	2.06 (1.19)	2.14 (1.38)	2.14 (1.42)	1	1	1
	Unacceptable	2.18 (1.52)	2.46 (1.64)	2.15 (1.36)	1	1	0.607
Understand AI's reasoning	Acceptable	4.72 (1.54)	5.90 (1.08)	5.91 (0.87)	**3.14E-06**	**8.06E-08**	1
	Unacceptable	4.14 (1.85)	5.38 (1.58)	5.24 (1.66)	**1.39E-06**	**1.25E-06**	1
Entice to act	Acceptable	5.06 (1.52)	5.54 (1.62)	5.76 (1.36)	**0.020**	**1.63E-05**	1
	Unacceptable	5.20 (1.61)	5.56 (1.46)	5.42 (1.60)	0.337	1	1
Liked by users	Acceptable	3.80 (1.50)	5.31 (1.65)	5.50 (1.29)	**9.45E-06**	**1.65E-09**	1
	Unacceptable	3.00 (1.66)	5.10 (1.96)	4.85 (1.66)	**2.81E-07**	**5.57E-09**	1

D Experimental results

Table 9 displays the means and standard deviations of the users' ratings of the three explanation types with respect to the four explanatory attributes and the extent to which an explanation was liked, and the statistical significance of the results (Wilcoxon signed-rank test). Table 10 displays the same ratings broken down according to the acceptance status of a car. Table 11 shows the features expected by users that were omitted from conservative explanations for each car scenario.

Table 11: Number of users who expected to see a feature that was omitted from our explanations for each scenario; a feature that was mentioned in our explanations for that scenario is denoted by "–".

Car #	**Car16**	**Car53**	**Car77**	**Car80**
Feature / Outcome	accept	accept	unaccept	unaccept
Buying price	–	–	30	32
Maintenance cost	–	–	–	12
Number of doors	8	14	12	6
Seating capacity	–	–	–	–
Luggage boot size	15	13	–	6
Safety rating	–	17	–	–