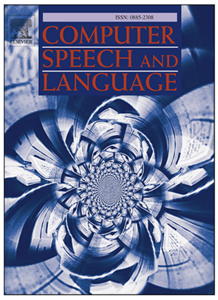
# Journal Pre-proof




Influence of context on users' views about explanations for decision-tree predictions

Sameen Maruf, Ingrid Zukerman, Ehud Reiter, Gholamreza Haffari

PII: S0885-2308(23)00002-5
DOI: https://doi.org/10.1016/j.csl.2023.101483
Reference: YCSLA 101483

To appear in: *Computer Speech & Language*

Received date : 15 April 2022
Revised date : 19 December 2022
Accepted date : 3 January 2023

Please cite this article as: S. Maruf, I. Zukerman, E. Reiter et al., Influence of context on users' views about explanations for decision-tree predictions. *Computer Speech & Language* (2023), doi: https://doi.org/10.1016/j.csl.2023.101483.

# Influence of Context on Users' Views about Explanations for Decision-Tree Predictions*

Sameen Maruf[a], Ingrid Zukerman[a,*], Ehud Reiter[b], Gholamreza Haffari[a]

[a]*Department of Data Science and AI, Monash University, VIC 3800, Australia*
[b]*Department of Computing Science, University of Aberdeen, Scotland AB24 3UE, UK*

**Abstract**

We consider the influence of two types of contextual information, *background information available to users* and *users' goals*, on users' views and preferences regarding textual explanations generated for the outcomes predicted by Decision Trees (DTs). To investigate the influence of background information, we generate contrastive explanations that address potential conflicts between aspects of DT predictions and plausible expectations licensed by background information. We define four types of conflicts, operationalize their identification, and specify explanatory schemas that address them. To investigate the influence of users' goals, we employ an interactive setting where given a goal and an initial explanation for a predicted outcome, users select follow-up questions, and assess the explanations that answer these questions. Here, we offer algorithms to generate explanations that address six types of follow-up questions.

The main result from both user studies is that explanations which have a contrastive aspect about a predicted class are generally preferred by users. In addition, the results from the first study indicate that these explanations are deemed especially valuable when users' expectations differ from predicted outcomes; and the results from the second study indicate that contrastive explanations which describe how to change a predicted outcome are particularly well regarded in terms of helping users' achieve this goal, and they are also popular in terms of helping users' achieve other goals.

*Keywords:* explainable AI, generating textual explanations, taking context into account, contrastive explanations, decision trees.

*This paper significantly expands the paper entitled "Explaining Decision-Tree Predictions by Addressing Potential Conflicts between Predictions and Plausible Expectations", co-authored by Sameen Maruf, Ingrid Zukerman, Ehud Reiter and Gholamreza Haffari, published in *INLG'2021 Proceedings – International Conference on Natural Language Generation*, pp. 114-127, Aberdeen, Scotland, UK, URL: https://aclanthology.org/2021.inlg-1.12.

*Corresponding author
*Email addresses:* `sameen.maruf@monash.edu` (Sameen Maruf), `ingrid.zukerman@monash.edu` (Ingrid Zukerman), `e.reiter@abdn.ac.uk` (Ehud Reiter), `gholamreza.haffari@monash.edu` (Gholamreza Haffari)

## 1. Introduction

Machine Learning (ML) models have become increasingly accurate in recent times, leading to their widespread adoption by decision makers in a variety of vital domains, including healthcare, defense and energy. This underscores the need for explanations of the outcomes of these models that support decision making by practitioners.

The research in explaining complex ML models can be broadly classified into two categories: (a) generating *post-hoc* explanations to explain specific outcomes of a model (Biran and McKeown, 2017; Ribeiro et al., 2016, 2018), and (b) explaining an entire model (Bastani et al., 2017; Lakkaraju et al., 2017). In this work, we focus on the first type of explanations, aimed at non-expert end users, such as decision makers and people affected by the outcomes of a model.

ML models may be classified into transparent and opaque models based on their interpretability (Doshi-Velez and Kim, 2017). Transparent models are "interpretable by a Machine Learning expert or a statistician" (Biran and McKeown, 2017). These models, e.g., Decision Trees (DTs), decision rules and linear models, are built on the basis of interpretable features, which are typically obtained through feature engineering. Transparent models are often less accurate than opaque models, in particular neural networks, provided large training datasets are available. However, large training datasets are not always available, as is the case in our evaluation datasets (Section 4.2). In addition, it is common practice to clarify the outcomes of opaque models by approximating them with transparent models (Section 2). Finally, even if these transparent models are understandable by ML experts, they may still be unclear to lay practitioners and end users, thus motivating us (and several others) to explain the outcomes of transparent models.

In this paper, we consider the influence of two types of contextual information, *background information available to users* and *users' goals*, on users' views and preferences regarding textual explanations generated for the outcomes of a particular transparent ML model: DT. We developed algorithms to generate different types of explanations, and conducted user studies to evaluate these explanations and assess the influence of these types of contextual information on users' views about the explanations. Our explanation-generation algorithms constitute a step towards explaining predictions of tree-based ensembles, such as Random Forests and AdaBoost, and the insights obtained from our studies generalize to other transparent models, such as decision rules and logistic regressors.

We now provide an overview of each type of contextual information, including datasets, evaluation and main findings.

***Background information.*** To investigate the influence of background information on users' views regarding explanations, we generated contrastive explanations that address potential conflicts between aspects of DT predictions and plausible expectations licensed by background information (i.e., expectations that "make sense" in light of this information). Specifically, we identified four

| **Feature** | **Value** |
|---|---|
| *Parents' employment*: | Challenging |
| *Current childcare*: | Good |
| *Child's health*: | Average |

| |
|---|
| From the data, one might expect that children with **good** ***current childcare*** will be a great deal more likely to get *Wait listed* than to get a *Priority acceptance* (54% vs 11%). However, the AI system has learned from the data that among children with **challenging** ***parents' employment*** and **average** ***health***, those with **good** ***current childcare*** are almost certain to get a *Priority acceptance* (close to 100%). |

Table 1: Features used in the prediction for an instance in the Nursery dataset, and explanation that addresses a potential conflict licensed by background information – the feature value that prompts this expectation appears in red; font denotes ***features***, **feature values** and *classes*.

types of conflicts whereby events that appeared unlikely or likely on the basis of background information happened or did not happen respectively, and then specified schemas for explanations that address these conflicts (Section 3.1).

*Datasets.* Explanations were generated for two datasets: *Nursery* and *Telecom* (Section 4.2). In Nursery, a DT predicts the acceptance status of a child to a childcare center on the basis of the circumstances of the child and their family (e.g., how satisfactory are the current childcare arrangements and how demanding is the parents' employment). In Telecom, a DT predicts whether a customer will churn (leave) or stay with a telecommunications company based on their profile (e.g., how long the customer has been with the company and what are their monthly charges). Table 1 illustrates an explanation generated for an outcome predicted for an instance in the Nursery dataset. The explanation addresses a potential conflict between a plausible expectation that a child with good *current childcare* is likely to be *Wait listed*, and the DT's prediction that the child will be *Priority accepted*.

*Evaluation.* We conducted a user study to evaluate the generated explanations in terms of completeness and presence of extraneous information, and also in terms of their ability to achieve two goals: enable users to understand the AI's reasoning for the predicted outcome, and motivate them to act on the AI's predictions.[1]

*Main findings.* The main findings of this study are: (1) explanations that address potential conflicts are generally considered at least as good as basic explanations that just follow a path in a DT in terms of completeness, helping users understand the AI's reasoning and enticing them to act on the predictions; and (2) Conflict-based explanations are deemed especially valuable when the outcome expected by users disagrees with DT predictions. We stress that these

[1] The participants in our study were told that they have an AI, but they were not informed about the specifics of the ML model. Other explanatory objectives include enhancing trust in an ML system and helping debug the system (Reiter, 2019).

| Feature | Value |
|---|---|
| *Age*: | 45.5 |
| *Daily cigarette consumption*: | 0 |
| *HDL cholesterol*: | Optimal |

**Follow-up question**: Which factor changes will result in the same prediction (*low risk of a coronary event*) for me?

If nothing else changes in your circumstances, the following would result in the same prediction (*low risk of a coronary event*) for you:

- any changes in one of these factors: ***weight status***, ***daily alcohol intake***, ***blood pressure***, ***total cholesterol***, ***triglycerides*** and ***diabetes***; or
- your ***HDL cholesterol*** changes from **optimal** to **borderline**.

[Information about ***HDL cholesterol*** and factors that affect it may be found here.]

Table 2: Features used in the prediction for an instance in the Busselton dataset, follow-up question, and explanation that addresses this question; font denotes ***features***, **feature values** and *classes*.

findings pertain to explanations that address conflicts due to *plausible expectations* from background information — we do *not* claim that these explanations address *actual* user expectations.

***Users' goals.*** To investigate the influence of users' goals on their views regarding explanations, we employed an interactive setting where given a goal (*understand the AI's reasoning for the predicted outcome*, *change the predicted outcome* or *retain the predicted outcome*) and an initial basic explanation for a predicted outcome, users select follow-up questions, and assess the explanations that answer these questions. Specifically, we generated explanations that address six potential follow-up questions about predicted outcomes, e.g., "Which factors in the data are used by the AI system for its predictions?" and "Which factor changes will result in a specific different prediction for me?" (Section 3.2).

*Dataset.* Explanations were generated for the *Busselton* dataset (Section 4.2), where a DT predicts whether a person is at a high or low risk of coronary heart disease (CHD) based on demographic, medical and lifestyle information (e.g., how old they are and how much they smoke). Table 2 illustrates an explanation that addresses a question preferred by many users when the goal is to retain a predicted outcome (*low risk of a coronary event*): "Which factor changes will result in the same prediction for me?".

*Evaluation.* We conducted a user study to determine which follow-up questions are selected for different goals, and to evaluate the explanations that answer these questions in terms of their ability to address the questions, their usefulness for a specified goal, and whether additional information was needed to achieve this goal. In addition, like for the first study, users rated the explanations on completeness and on the presence of extraneous information.

*Main findings.* The main findings of this study are: (1) there is some overlap between the follow-up questions that were selected for all the goals, but there are enough differences to warrant tailoring explanations to users' goals; and (2) the follow-up question about changes that lead to a specific prediction that differs from the actual prediction is the most selected question for all the goals, and its associated *transfactual* explanation is not only highly rated in terms of usefulness for the goal of *changing the predicted outcome*, but also well regarded in terms of usefulness for the other goals.[2]

This paper is organized as follows. In Section 2, we discuss research on generating explanations for predictions made by ML models and related work on explanations that address users' reasoning and on interactive explanations. In Section 3, we present our approach to generate explanations that consider background information and address follow-up questions. Section 4 describes our datasets and experimental design. Our results appear in Section 5, followed by discussion and concluding remarks in Section 6.

## 2. Related work

In 1990-2000, explanations derived from knowledge bases were enhanced by addressing aspects of users' reasoning. Specifically, Zukerman and McConachy (1993) and Horacek (1997) considered potential inferences from explanations, omitting easily inferable information and addressing erroneous inferences; Korb et al. (1997) took into account reasoning fallacies when explaining the reasoning of Bayesian Networks; and Stone (2000) generated instructions from which users could draw appropriate inferences about actions to take.

A parallel line of work focused on interactive explanations. Moore and Paris (1993) introduced a system that reasons about the intentions behind utterances and the rhetorical relations between them, and uses this information to respond to users' follow-up questions. Cawsey (1993)'s system used interactions with users to update its initial assumptions about the users' knowledge, thus enabling the system to plan and present explanations incrementally. Zukerman et al. (1999) offered the following actions to interrogate explanations generated for Bayesian Networks: select a proposition to be explained, request to argue for/against a proposition in an explanation, explain the effect of a proposition on the goal (what about), include/exclude a proposition, and consider a hypothetical change in the belief in a proposition (what if). The last two actions lead to counterfactual arguments.

Current research on explanation generation focuses on explaining the predictions made by ML models – a sub-field called *Explainable AI* (*XAI*). 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. A common first step in explaining the predictions of neural networks is to build

[2]Hoffman and Klein (2017) and Hoffman et al. (2017) distinguish between counterfactual explanations, which pertain to past events that did not take place, and transfactual explanations, which pertain to changes that affect the future.

a *local surrogate explainer model* that uses a transparent model to approximate the neighbourhood of an instance of interest. Linear regression (Ribeiro et al., 2016; Lundberg and Lee, 2017), decision rules (Ribeiro et al., 2018) and DTs (van der Waa et al., 2018; Guidotti et al., 2019; Sokol and Flach, 2020a) have been employed for this purpose.

A DT's prediction is generally explained by tracing the path from the root to a predicted outcome (Guidotti et al., 2019; Stepin et al., 2020). Recently, researchers have generated class-contrastive counterfactual explanations to enhance the explanations of DT predictions. Stepin et al. (2020) generated explanations that have a factual and a counterfactual component; the former is the DT trace, while the latter is the DT path that leads to an alternative outcome and has the shortest bitwise XOR-based distance to the DT trace. However, they do not determine when a counterfactual enhancement is required. The need for an enhancement was studied in (Biran and McKeown, 2017) — they identified and addressed unexpected effects of individual features on predictions made by logistic regression. However, they did not consider unexpected predictions.

The recent XAI research described above focuses on static explanations. A promising direction for future research is to allow users to *interactively* explore why a model predicted a particular outcome (Abdul et al., 2018). Cheng et al. (2019) found that their interactive interface, which allowed users to modify the value of features and see the impact of this change on the prediction of a linear regressor (what if), increased users' objective and self-reported understanding of the ML model compared to a static interface, which did not allow such changes. Sokol and Flach (2020b) studied counterfactual explanations for DTs in an interactive system where users could change or remove features, or request an explanation for a hypothetical instance. Counterfactual explanations were generated by representing a tree structure as binary meta-features, and selecting the shortest statement that minimizes an $L1$-like metric compared to the DT trace.

Reiter (2019) argued that good explanations must be written for a specific purpose and audience, have a narrative structure, and use vague language to communicate uncertainty. The explanations generated in (Sokol and Flach, 2020b) and (Biran and McKeown, 2017) have a narrative structure, and those in (Biran and McKeown, 2017) use vague language to convey strength of evidence. A different perspective is offered by expectation theory, which posits that the surprisingness of an event may stem from a discrepancy between the state of the world and propositions that are deducible from presented information (Ortony and Partridge, 1987). Itti and Baldi (2009) offer a Bayesian formulation of the influence of surprisingness on visual attention shifts in terms of the difference between prior and posterior probabilities. In the first part of this research, we employ a probabilistic formulation to identify potential conflicts between plausible expectations and aspects of DT predictions. Our approach complements explanations by addressing both unexpected predictions and unexpected effects of feature values, thereby enhancing their narrative structure. In addition, we leverage the work of Elsaesser and Henrion (1989) to address Reiter's desideratum of using vague language to convey probabilities.

Based on insights from psychology, Miller (2019) argued that the explanatory process is best thought of as a conversation. In line with this, Weld and Bansal (2019) envisioned an interactive explanation system that presents users with an initial explanation, and supports several follow-up questions to further this conversation. A question-driven framework for interactive explanations was also advocated in (Liao et al., 2020). To this effect, they developed an XAI question bank comprising nine categories, each of which contains prototypical questions that represent users' requirements from explanations. However, these questions were explored through practitioners who design interfaces for end users, not the end users themselves. In addition, Liao et al. (2020) posited that different goals may prompt users to want answers for different types of questions. In the second part of this research, we consider a subset of the categories in Liao et al.'s XAI question bank that pertains to the reasoning of an ML model, and investigate its relevance to different users' goals through an interactive question-driven setting.

## 3. Justifying DT predictions

In this section, we explain the outcomes predicted by a DT for particular instances, where an instance comprises a set of *features*, each associated with a *value*, and an outcome is a *discrete class.* For example, the top of Table 1 shows features and values used by a DT to make a prediction of *Priority acceptance* for a particular Nursery instance (the other classes are *Reject* and *Wait list*) — Table C.20 contains a detailed description of the feature values in the Nursery dataset; Table 7 displays the features and associated values in our evaluation datasets.

As mentioned in Section 1, in this work, we investigate the influence of two types of contextual information on users' views about textual explanations for DT predictions: (1) background information available to users, and (2) users' goals. For the former, we generate one-shot explanations that address potential expectations licensed by background information that are violated by a predicted outcome and/or the impact of a feature value (Section 3.1). For the latter, given an initial explanation for a DT's prediction, we consider several types of follow-up questions that may help users achieve particular goals, and generate explanations for each type of question (Section 3.2). The main difference between the explanations generated to investigate the two types of contextual information is that in the former, the part of the explanation that addresses an expectation violation is wrapped around a basic baseline explanation that just follows a DT path, while in the latter, a basic explanation is presented first, and we provide stand-alone explanations that address individual follow-up questions.

### 3.1. Influence of background information

Like Biran and McKeown's (2017) approach, ours hinges on identifying discrepancies, but it differs from their approach in that (1) we propose *addressing* *potential conflicts* as a guiding principle for selecting content that complements explanations of DT predictions; (2) these conflicts pertain to predicted outcomes and to the impact of feature values; and (3) we identify these conflicts by comparing aspects of a DT prediction with plausible expectations derived from background information. Thus, our conflict-based explanations are contrastive with respect to the predicted outcome and/or the impact of feature values.

#### *3.1.1. Potential Conflicts*

First, we define *plausible expectations* and *aspects of a DT prediction*, which are the building blocks of *potential conflicts*. We then specify language-based probabilistic relations that are the basis for plausible expectations, and describe the identification of potential conflicts.

***Plausible expectations*** pertain to the outcome predicted by a DT and to the impact of a value $j$ of feature $x_i$, denoted $x_{i,j}$. They are derived from the prior and posterior probabilities of outcomes by means of relations R1-R3 and associated constraints (Table 3) — a feature value satisfying any of these relations and associated constraints is expected to have an impact.

R1. $Posterior(\mathcal{C}|\, x_{i,j})$ vs $Prior(\mathcal{C})$

R2. $Posterior(\mathcal{C}_{max}|\, x_{i,j})$ vs $Prior(\mathcal{C}_{max})$

R3. $Posterior(\mathcal{C}_{max}|\, x_{i,j})$ vs $Posterior(\mathcal{C}|\, x_{i,j})$

where $Prior(c)$ is the prior probability of a class $c$, $Posterior(c|x_{i,j})$ is the probability of class $c$ given feature value $x_{i,j}$, $\mathcal{C}$ is the class predicted by a DT, and $\mathcal{C}_{max}$ is an alternative class with the highest *Posterior* probability. Our formalism assumes that users are aware of the probabilities in R1-R3 (they were given this information in our evaluation, Section 4.3.1).

The posterior probability of a class $c$ is calculated from training data for each feature value $x_{i,j}$. If it is high, it may license an expectation for $x_{i,j}$ to yield class $c$, and if it is low, the expectation may be for $x_{i,j}$ to *not* result in class $c$ (and to yield a class different from $c$). For example, according to the Nursery data, children with ordinary *parents' employment* have a lower probability of getting a *Priority acceptance* to the childcare center than children in the general population (R1), and the probability that children with ordinary *parents' employment* will get *Priority accepted* is lower than the probability that they will not. Hence, it is plausible to expect a child with such parents not to be *Priority accepted*.

***Aspects of a DT Prediction*** pertain to the class $\mathcal{C}$ *Predicted* by the DT, and the *Impact* of feature value $x_{i,j}$ on this class, denoted $Impact(x_{i,j}, \mathcal{C})$. The *Predicted* class $\mathcal{C}$ is determined by the features and their values in the current DT path, which may or may not include $x_{i,j}$. $Impact(x_{i,j}, \mathcal{C})$ is TRUE if $x_{i,j}$ influences the *Predicted* class $\mathcal{C}$ — for a DT, this happens when $x_{i,j}$ is in the path to $\mathcal{C}$; *Impact* is FALSE otherwise.

***A potential conflict*** takes place when an expected outcome differs from the

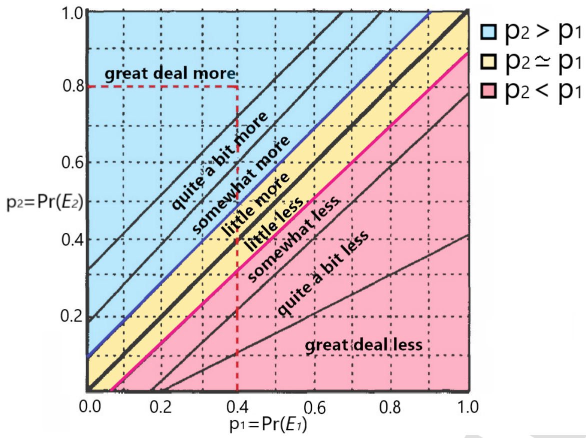



Figure 1: Verbal mapping of relative probabilities.

class predicted by a DT (R4), or when a feature value that was expected to have an impact does not (R5).[3]

R4. *Plausible* outcome from $x_{i,j} \neq$ *Predicted* class $\mathcal{C}$

R5. *Plausible* impact of $x_{i,j} \neq$ *Impact*$(x_{i,j}, \mathcal{C})$

In our example, a potential conflict ensues because, contrary to the expectation, the class *Predicted* for the child is *Priority accept* (R4).

It is worth noting that the only relation that depends on the model is R5, where the *Impact* of a feature value for DTs is determined by path membership. Relations R1-R3 and R4 are model agnostic: R1-R3 depend on probabilities obtained from the data, and R4 depends on R1-R3 and the *Predicted* class. The values of relations R1-R3 are obtained from discretized probabilistic relations described as follows.

***Discretizing probabilistic relations.*** To generate explanations that use language to communicate relative probabilities, we harness the research of Elsaesser and Henrion (1989), which maps probability differences into verbal expressions.[4] Figure 1 depicts their empirically derived phrase-selection function, which achieved a 72% accuracy compared to people's actual usage. For example, if the probability of event $E_1$ is $p_1 = 0.4$, and that of event $E_2$ is $p_2 = 0.8$

[3]Biran and McKeown (2017) consider situations where a feature may be expected to have a high or a low impact. But in a probabilistic formulation, expecting an event with low probability is tantamount to expecting this event *not* to happen with high probability.

[4]There is more recent research on verbalizing absolute probabilities (Wintle et al. (2019) and citations therein), but to the best of our knowledge, the work of Elsaesser and Henrion (1989) is the only one that considers changes in probabilities.

(dashed red lines in Figure 1), the phrase "$E_2$ is *a great deal more* likely than $E_1$" is selected.

Following a small pilot study to validate these expressions for our explanations, we merged the intermediate expressions "somewhat more/less" and "quite a bit more/less" in Figure 1 into simply "more/less". The resultant six-phrase mapping is used to define the wording for relations R1-R3.

***Identifying Potential Conflicts.*** Table 3 displays the potential conflicts addressed by our explanations. Each segment represents a potential conflict, with the surprises boxed in red. Column 1 shows the name of the conflict, Column 2 displays the relations that license plausible expectations for an outcome and for the impact of feature value $x_{i,j}$ (the colour-coded relations are computed as specified in Figure 1, while the constraints are calculated using point probabilities); Column 3 presents the *Plausible* expected outcome from $x_{i,j}$ derived from the relations and constraints in Column 2; Column 4 shows the actual *Predicted* class $\mathcal{C}$ based on the values of the features in the current DT path; Column 5 displays the *Plausible* expected impact of $x_{i,j}$, which is always TRUE; and Column 6 shows the actual *Impact*$(x_{i,j}, \mathcal{C})$. Relation R4 is calculated by comparing the values of Columns 3 and 4, and Relation R5 is obtained from Columns 5 and 6.

We now describe each conflict, and illustrate it with examples from the Nursery dataset.

***Plausible¬$\mathcal{C}$/Predict$\mathcal{C}$*** (first segment in Table 3). This conflict arises when it is plausible to expect that in light of $x_{i,j}$, class $\mathcal{C}$ will not happen (Column 3), but surprisingly, $\mathcal{C}$ is *Predicted* (Column 4). The expectation is plausible because the posterior probability of class $\mathcal{C}$ given $x_{i,j}$ is less than or equal to its prior probability (R1), and also lower than the posterior probability of $\neg\mathcal{C}$ (Column 2), where $\neg\mathcal{C}$ denotes all the classes other than $\mathcal{C}$. For this conflict, we only examined the case where *Impact*$(x_{i,j}, \mathcal{C})$ = TRUE, i.e., $x_{i,j}$ is in the DT path. The FALSE case was disregarded, as the ensuing potential conflict seemed weak. However, for completeness, this case should be revisited in the future.
**Example** (full text in Table 4): In the Nursery dataset, children with critical *current childcare* are less likely to be *Wait listed* than applicants overall (R1: *Posterior* $<$ *Prior*). However, in the context of other information about a particular child, having critical *current childcare* gets them *Wait listed* (R4: *Plausible* outcome $\neg\mathcal{C} \neq$ *Predicted* class $\mathcal{C}$).[5]

***Plausible$\mathcal{C}$/Predict$\mathcal{C}$-$x_{i,j}$NoImpact*** (second segment in Table 3). This conflict occurs when a feature value $x_{i,j}$ is expected to have an impact (Column 5), but it has no effect on the *Predicted* class, i.e., it is not in the DT path (Column 6). The expectation for $x_{i,j}$ to have an impact arises when the posterior probability of class $\mathcal{C}$ in light of $x_{i,j}$ is higher than its prior probability (R1) and the posterior probabilities of all the other classes, and it is also higher than

[5]As seen in Table C.20, the term "critical childcare" indicates high insecurity in obtaining this service.

| | | R4 | | R5 | |
|---|---|---|---|---|---|
| **Conflict name** | **Relations licensing plausible expectations** | ***Plausible* outcome from $x_{i,j}$** | ***Predicted* class** | ***Plausible* impact of $x_{i,j}$** | ***Impact*$(x_{i,j}, \mathcal{C})$** |
| *Plausible*$\neg\mathcal{C}$ / *Predict*$\mathcal{C}$ | R1: $Posterior(\mathcal{C} \mid x_{i,j}) <, \simeq Prior(\mathcal{C})$<br>$Posterior(\mathcal{C} \mid x_{i,j}) < Posterior(\neg\mathcal{C} \mid x_{i,j})$ | $\neg\mathcal{C}$ | $\boxed{\mathcal{C}}$ | TRUE | TRUE |
| *Plausible*$\mathcal{C}$ / *Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact* | R1: $Posterior(\mathcal{C} \mid x_{i,j}) > Prior(\mathcal{C})$<br>$\forall C_k \neq \mathcal{C}\ Posterior(\mathcal{C} \mid x_{i,j}) > Posterior(C_k \mid x_{i,j})$<br>$\exists x_{m,n}\quad Posterior(\mathcal{C} \mid x_{i,j}) > Posterior(\mathcal{C} \mid x_{m,n})$ | $\mathcal{C}$ | $\mathcal{C}$ | TRUE | $\boxed{\text{FALSE}}$ |
| *Plausible*$\mathcal{C}_{max}$ / *Predict*$\mathcal{C}$ "vanilla" | R1: $Posterior(\mathcal{C} \mid x_{i,j}) <, \simeq Prior(\mathcal{C})$<br>R2: $Posterior(\mathcal{C}_{max} \mid x_{i,j}) > Prior(\mathcal{C}_{max})$<br>R3: $Posterior(\mathcal{C}_{max} \mid x_{i,j}) > Posterior(\mathcal{C} \mid x_{i,j})$ | $\mathcal{C}_{max}$ | $\boxed{\mathcal{C}}$ | TRUE | TRUE |
| *Plausible*$\mathcal{C}_{max}$ / *Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact* | $\forall C_k \neq \mathcal{C}_{max}\ Posterior(\mathcal{C}_{max} \mid x_{i,j}) > Posterior(C_k \mid x_{i,j})$ | | | | $\boxed{\text{FALSE}}$ |

Table 3: Definition of potential conflicts (explanations appear in Tables 1, 4 and A.17): $\mathcal{C}$ denotes the *Predicted* class based on the values of the features in the current DT path, and $\mathcal{C}_{max}$ denotes an alternative class that has the highest *Posterior* probability; the colours of (in)equalities match those in Figure 1; $\boxed{\text{text}}$ in Column 4 indicates surprise about the plausible outcome from $x_{i,j}$ in Column 3, and $\boxed{\text{text}}$ in Column 6 expresses surprise about the plausible impact of $x_{i,j}$ in Column 5.

the posterior probability of class $\mathcal{C}$ in light of at least one feature value in the current DT path — $x_{i,j}$ cannot be the "weakest" among the mentioned features (Column 2). Here, the plausible expectation for class $\mathcal{C}$ matches the DT's prediction, i.e., there is no conflict about the expected outcome.

**Example** (full text in Table 4): In the Nursery dataset, children with challenging *parents' employment* are more likely to get *Priority accepted* than the general population (R1: *Posterior* $>$ *Prior*), but *parents' employment* is not in the DT path (R5: *Plausible* impact of $x_{i,j} \neq$ actual *Impact*$(x_{i,j}, \mathcal{C})$).

***Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$** (third segment in Table 3). Here, a particular alternative outcome $\mathcal{C}_{max}$ is a plausible expectation from $x_{i,j}$ (Column 3), but surprisingly, class $\mathcal{C}$ is *Predicted* (Column 4). This conflict resembles *Plausible*$\neg\mathcal{C}$/*Predict*$\mathcal{C}$ in that the posterior probability of class $\mathcal{C}$ in light of $x_{i,j}$ is relatively low, i.e., $\neg\mathcal{C}$ is plausible (R1). However, *Plausible*$\mathcal{C}_{max}$/*Predict*$\mathcal{C}$ goes further, nominating a potential alternative class $\mathcal{C}_{max}$.[6] The expectation for $\mathcal{C}_{max}$ is plausible because its posterior probability is higher than its prior (R2) and the posterior of $\mathcal{C}$ (R3), and $\mathcal{C}_{max}$ has the highest posterior probability among all the classes (Column 2). This conflict has two variants: "**vanilla**" – only the *Predicted* class is unexpected (top of the third segment); and $x_{i,j}$***NoImpact*** – both the *Predicted* class and the lack of impact of $x_{i,j}$ (Column 6) are unexpected (bottom of the third segment).

**Example of the first variant** (full text in Table 1; the second variant appears in Table 4): In the Nursery dataset, children with good *current childcare* are more likely to get *Wait listed* than *Priority accepted* (R3: *Posterior*$(\mathcal{C}_{max}) >$ *Posterior*$(\mathcal{C})$). However, a particular child with certain feature values and good *current childcare* gets *Priority accepted* (R4: *Plausible* outcome $\mathcal{C}_{max} \neq$ *Predicted* class $\mathcal{C}$).

### *3.1.2. Generating Conflict-based Explanations*

The inputs to the explanation generator are: an instance, a *Predicted* class and a set of conflicts. At present, our explanations address a potential conflict with respect to one feature value only.[7] Thus, for each conflict type, we first select a *pivot feature value* (denoted $x^*_{i,j}$), and then realize our explanation. We do not select a particular conflict type for an instance, as making this determination is one of the aims of our evaluation (Section 5.1).

***Selecting a pivot feature value.*** If several feature values qualify for a potential conflict type, we choose the strongest in terms of word mapping, e.g., "a great deal more" is stronger than "more". Ties are broken as follows: for *Plausible*$\neg\mathcal{C}$/*Predict*$\mathcal{C}$ and *Plausible*$\mathcal{C}$/*Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact*, we choose the $x^*_{i,j}$

[6] For a binary classification problem, one would expect that the same $x_{i,j}$ should qualify for both *Plausible*$\neg\mathcal{C}$/*Predict*$\mathcal{C}$ and *Plausible*$\mathcal{C}_{max}$/*Predict*$\mathcal{C}$. However, given the added constraints in *Plausible*$\mathcal{C}_{max}$/*Predict*$\mathcal{C}$ (Table 3), this is not always the case.

[7] In the future, we will consider higher-dimensional spaces, which may require addressing conflicts about several features.

| Schema | Sample explanations generated for the Nursery dataset |
|---|---|
| **Basic (no conflict): counterpart of *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$-$x_{i,j}$NoImp*** | |
| *DT-Path* + $\mathcal{C}$ | The AI system has learned from the data that children with **very critical *current childcare*** and **average *health*** are almost certain to get a *Priority acceptance* (close to 100%). |
| **Conflict-based (outcome only): *Plausible¬$\mathcal{C}$/Predict$\mathcal{C}$*** | |
| *Preamble*: $x^*_{i,j}$ + R1 + $\mathcal{C}$ | From the data, one might expect that children with **critical *current childcare*** will be less likely than applicants overall to get *Wait listed* (19% vs 34%). |
| *Resolution*: {*DT-Path*/$x^*_{i,j}$} + $x^*_{i,j}$ + $\mathcal{C}$ | However, the AI system has learned from the data that among children with **ordinary *parents' employment***, **somewhat problematic *social situation*** and **good *health***, those with **critical *current childcare*** are almost certain to get *Wait listed* (close to 100%). |
| **Conflict-based (impact of feature value only): *Plausible$\mathcal{C}$/Predict$\mathcal{C}$-$x_{i,j}$NoImp*** | |
| *Preamble*: $x^*_{i,j}$ + R1 + $\mathcal{C}$ | From the data, one might expect that children with **challenging *parents' employment*** will be more likely than applicants overall to get a *Priority acceptance* (46% vs 32%). |
| *Resolution*: $x^*_i$ + R5 + *DT-Path* + $\mathcal{C}$ | However, the AI system has learned from the data that the ***parents' employment*** has no effect on the outcome in this situation, and that children with **very critical *current childcare*** and **good *health*** are almost certain to get a *Priority acceptance* (close to 100%). |
| **Conflict-based (outcome & impact of feature value): *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$-$x_{i,j}$NoImp*** | |
| *Preamble*: $x^*_{i,j}$ + R3 + $\mathcal{C}_{max}$ + $\mathcal{C}$ | From the data, one might expect that children with **ordinary *parents' employment*** will be more likely to get *Wait listed* than to get a *Priority acceptance* (47% vs 19%). |
| *Resolution*: $x^*_i$ + R5 + *DT-Path* + $\mathcal{C}$ | However, the AI system has learned from the data that the ***parents' employment*** has no effect on the outcome in this situation, and that children with **very critical *current childcare*** and **average *health*** are almost certain to get a *Priority acceptance* (close to 100%). |

Table 4: Basic schema (our baseline) and schemas that address three of the potential conflicts defined in Table 3 (*NoImp* is shorthand for *No Impact*), with sample explanations for the Nursery dataset; relative probabilities are described in Figure 1, and the presentation of probabilities in brackets is in line with the findings in (Wintle et al., 2019); the selection of a ***pivot feature* value** is described in Section 3.1.2; font denotes ***features*, feature values** and *Classes*.

with the maximum absolute difference between $Posterior(\mathcal{C}|\, x_{i,j})$ and $Prior(\mathcal{C})$ for the *Predicted* class $\mathcal{C}$. For the *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$* variants, we select the $x^*_{i,j}$ with the maximum difference between $Posterior(\mathcal{C}_{max}|\, x_{i,j})$ and $Posterior(\mathcal{C}|\, x_{i,j})$.

***Realizing explanations.*** Explanations are represented by schemas (Table 4); the schemas for Conflict-based explanations have two main parts: *Preamble*, which presents a plausible expectation from the pivot feature value $x^*_{i,j}$, and *Resolution*, which describes how this expectation is thwarted.

The ***Preamble*** presents probabilistic relations that license plausible expectations. The preambles of *Plausible¬$\mathcal{C}$/Predict$\mathcal{C}$* and *Plausible$\mathcal{C}$/Predict$\mathcal{C}$-$x_{i,j}$NoImpact* describe relation R1; and those of the *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$* variants convey R3.

The ***Resolution*** has two components: (1) the feature values in the DT path that lead to the *Predicted* class $\mathcal{C}$, which also constitutes the Basic baseline explanation (Guidotti et al., 2019; Stepin et al., 2020); and (2) the impact of $x^*_{i,j}$, or lack thereof, in the context of the other feature values in the DT path. The features in the DT path are presented in a pre-established order (Table 7), except for $x^*_{i,j}$, whose placement is determined by the schemas: when $x^*_{i,j}$ is in the DT path, it appears right before the *Predicted* class; otherwise, the lack of impact of $x^*_i$ is announced at the start of the *Resolution*.

Table 4 displays schemas of explanations that address three potential conflicts, and one Basic schema (which is our baseline), together with sample explanations for the Nursery dataset; an explanation that illustrates *Plausible*$\mathcal{C}_{max}$/*Predict*$\mathcal{C}$ "vanilla" for the Nursery dataset appears in Table 1 (the schema for this potential conflict is [*Preamble*: $x^*_{i,j}$ + R3 + $\mathcal{C}_{max}$ + $\mathcal{C}$; *Resolution*: $\{$*DT-Path*/$x^*_{i,j}\}$ + $x^*_{i,j}$ + $\mathcal{C}$]; sample explanations for the Telecom dataset appear in Table A.17. Since the focus of our research is on content selection, the schemas are realized by means of domain-independent programmable templates (Table A.16).

### 3.2. Influence of users' goals

In this part of the work, we postulate that users' goals may influence their preferences and opinions of explanations for outcomes predicted by an ML model. To explore this idea, we consider three goals: *understand the AI's reasoning for a predicted outcome*, *change the predicted outcome* and *retain the predicted outcome*. After viewing an instance and an initial Basic explanation for a prediction, users are given one of these goals. They then choose follow-up questions to achieve this goal, and we generate an explanation to address each question.

The first two goals have been defined as explanatory goals in (Wachter et al., 2018). The goal of *understanding the AI's reasoning* is similar to the general goal of transparency in XAI (Felzmann et al., 2019), and is also considered in the evaluation of our first approach (Section 4.3.1). The goals of *changing* or *retaining a predicted outcome* pertain to the impact of ML predictions on end users, and unlike the first goal, they depend on the desirability of an outcome, i.e., people usually want to change an undesirable outcome to a desirable one, and retain a desirable outcome.

#### 3.2.1. Users' goals and follow-up questions

Most of the explanatory goals described in the literature, such as trust, effectiveness and persuasiveness (Tintarev and Masthoff, 2012; Nunes and Jannach, 2017), are from an explainer's perspective. In this work, we consider the perspective of a recipient of an explanation.

As mentioned above, in order to achieve a particular goal, users may want to ask follow-up questions. However, allowing open-ended questions may require additional interactions and may result in misunderstandings, which obfuscates the aim of this work. In addition, even if a question is understood, it may not be possible to generate an answer for it in the context of a particular ML model. To alleviate these problems, we selected six follow-up questions that cover the subset of question categories specific to explaining a model's reasoning in the XAI question bank in (Liao et al., 2020; Liao and Varshney, 2022):

- General Questions:

  *FactorsUsed?*: Which factors in the data are used by the AI system for its predictions?

  *FactorsNotUsed?*: Which factors in the data are not used by the AI system for its predictions?

- Profile-specific Questions:

  *WhyNot*$\mathcal{C}'$*?*: Why wasn't I given a specific different prediction?

  *HowtoGet*$\mathcal{C}'$*?*: Which factor changes will result in a specific different prediction for me?

  *HowtoStillGet*$\mathcal{C}$*?*: Which factor changes will result in the same prediction for me?

  *WhatIf-Change1Factor?*: What would be the prediction if one of the factors were to change for me? [Users are then asked to select one factor]

*FactorsUsed?* and *FactorsNotUsed?* are general questions about the workings of the ML model, which complement each other; *FactorsNotUsed?* is related to the $x_{i,j}$*NoImpact* variants in Section 3.1, but here it is presented as a general question about the features not used by the model at all. The remaining four questions are specific to a user's profile (an instance) and the predicted outcome $\mathcal{C}$, and are inspired by research on contrastive, counterfactual and transfactual explanations (Lipton, 1990; Miller, 2019; Verma et al., 2020; Stepin et al., 2021; Hoffman and Klein, 2017; Hoffman et al., 2017).[8] *WhyNot*$\mathcal{C}'$*?* and *HowtoGet*$\mathcal{C}'$*?* are class-contrastive questions, as they refer to a specific outcome $\mathcal{C}'$ that differs from the predicted one. The explanation for *WhyNot*$\mathcal{C}'$*?* is similar to the *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" variant in Section 3.1. The explanations for *HowtoGet*$\mathcal{C}'$*?*, *HowtoStillGet*$\mathcal{C}$*?* and *WhatIf-Change1Factor?* are transfactual (Hoffman and Klein, 2017; Hoffman et al., 2017), in the sense that they discuss prospective actions that might occur, rather than retrospective actions that did not take place, as is done in counterfactual explanations (Verma et al., 2020; Guidotti et al., 2019; Sokol and Flach, 2018; Poyiadzi et al., 2020). For *HowtoGet*$\mathcal{C}'$*?* and *HowtoStillGet*$\mathcal{C}$*?*, the explanation-generation algorithm determines the factors of interest, while for

[8] Most of the literature does not distinguish between counterfactuals and transfactuals, and refers to explanations of this type broadly as counterfactuals.

| |
|---|
| ***FactorsUsed?*:** Which factors in the data are used by the AI system to predict a person's risk of a coronary event? |
| In general, the following factors are used by the AI system to predict a person's risk of a coronary event: ***age***, ***gender***, ***weight status***, ***daily alcohol intake***, ***daily cigarette consumption***, ***total cholesterol*** and ***HDL cholesterol***. |
| ***FactorsNotUsed?*:** Which factors in the data are not used by the AI system to predict a person's risk of a coronary event? |
| The following factors do not improve the accuracy of the AI's predictions, and hence are not used by the AI system: ***blood pressure***, ***triglycerides*** and ***diabetes***. |

Table 5: Sample explanations generated for the two general questions for the Busselton dataset; font denotes ***features***.

*WhatIf-Change1Factor?*, the user selects one factor. It should be noted that for a multi-class classification problem, users should nominate the other class of interest $\mathcal{C}'$ for questions *WhyNot*$\mathcal{C}'$*?* and *HowtoGet*$\mathcal{C}'$*?*. In contrast, when we generate explanations for *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$, we nominate the class with the highest *Posterior* probability as the contrastive class (Section 3.1).

### *3.2.2. Generating explanations for follow-up questions*

The algorithm that generates the content of the explanations which answer follow-up questions depends on the underlying ML model (a DT in this research). The inputs to the algorithm are: an instance, a *Predicted* class and a DT.[9] Table 5 displays sample explanations generated for the general questions, and Table 6 contains sample explanations for the profile-specific questions with respect to an instance used in our evaluation. The schemas for these explanations are realized by means of programmable templates (Tables A.18 and A.19).

***Explanations for general questions.*** The explanation for *FactorsUsed?* lists the subset of features used by a DT for making its predictions, which is obtained by collating the features from all the DT paths. To answer *FactorsNotUsed?*, we simply remove the subset of features obtained for *FactorsUsed?* from the set of features in the dataset.

***Explanation for WhyNot*$\mathcal{C}'$*?*.** This explanation differs from the *Plausible-*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" variant in that users may select *WhyNot*$\mathcal{C}'$*?* for alternative classes $\mathcal{C}'$ for which the algorithm in Section 3.1 would not have postulated a potential conflict on the basis of background information.

To generate this explanation, we take the DT path that leads to the *Predicted* class $\mathcal{C}$ for the instance in question (our Basic explanation), and for each node in this path, we compute the probability of the other class $\mathcal{C}'$ from the DT, given the feature values up to and including this node. Intuitively, this tells us how

[9]Multiple splits on the same numeric feature in a DT path (*age* in our case) are merged. For example, for the DT in Figure B.8, we merge the two splits: age $\leq$ 60.5 and age $>$ 42.6, into 42.6 $<$ age $\leq$ 60.5, and generate the phrase 'between 43 and 60 years old' (Basic and *WhyNot*$\mathcal{C}'$*?* segments in Table 6).

| |
|---|
| **Instance:** |
| ***age***: 57.8, ***gender***: male, ***weight status***: overweight, ***daily alcohol intake***: 0, ***daily cigarette consumption***: 0, ***blood pressure***: normal-to-high, ***total cholesterol***: high, ***HDL cholesterol***: borderline, ***triglycerides***: borderline, ***diabetes***: no |
| **Prediction:** |
| *High risk of a coronary event* |
| **Basic explanation:** |
| This prediction was made because the AI system has learned from the data that **men** who are **between 43 and 60 years old**, have **high** ***total cholesterol*** and have **borderline** ***HDL cholesterol*** are at a *high risk of a coronary event.* |
| ***WhyNotC′?*:** Why wasn't I given a specific different prediction (*low risk of a coronary event*)? |
| The AI system has learned from the data that about 60% of **men** who are **between 43 and 60 years old** and have **borderline** ***HDL cholesterol*** are at a *low risk of a coronary event.* However, because you have **high** ***total cholesterol***, the AI system predicts that you are not at a *low risk of a coronary event.* |
| ***HowtoGetC′?*:** Which factor changes will result in a specific different prediction (*low risk of a coronary event*) for me? |
| If nothing else changes in your circumstances, the following would result in a different prediction (*low risk of a coronary event*) for you:<br>• your ***total cholesterol*** changes from **high** to any other value [**borderline**, **normal** or **low**]; or<br>• your ***HDL cholesterol*** changes from **borderline** to **optimal**. |
| ***HowtoStillGetC?*:** Which factor changes will result in the same prediction (*high risk of a coronary event*) for me? |
| If nothing else changes in your circumstances, the following would result in the same prediction (*high risk of a coronary event*) for you:<br>• any changes in one of these factors: ***weight status***, ***daily alcohol intake***, ***daily cigarette consumption***, ***blood pressure***, ***triglycerides*** and ***diabetes***; or<br>• your ***HDL cholesterol*** changes from **borderline** to **low**.<br>Also, if<br>• your ***daily cigarette consumption*** changes from **no cigarettes a day** to **more than 28 cigarettes a day**,<br>the prediction would remain the same, even if your ***HDL cholesterol*** changes from **borderline** to **optimal**. |
| ***WhatIf-Change1Factor?*:** What would be the prediction if one of the factors were to change for me? |
| User selects ***HDL cholesterol*** $\in$ *DT-Path* |
| If your ***HDL cholesterol*** changes from **borderline** to **low**, it would result in the same prediction for you (*high risk of a coronary event*), provided nothing else changes in your circumstances.<br>However, if your ***HDL cholesterol*** changes from **borderline** to **optimal**, it would result in a different prediction for you (*low risk of a coronary event*), provided nothing else changes in your circumstances. |
| User selects ***Daily cigarette consumption*** $\in DT$, $\notin$ *DT-Path* |
| If you **start smoking**, it would result in the same prediction for you (*high risk of a coronary event*), because ***daily cigarette consumption*** has no effect on the prediction in light of your ***age***, ***gender***, ***total cholesterol*** and ***HDL cholesterol***. |
| User selects ***Diabetes*** $\notin DT$ |
| If your **non-diabetic** status changes, it would result in the same prediction for you (*high risk of a coronary event*), because the AI system did not use ***diabetes*** to make predictions. |

Table 6: Sample Basic explanation and explanations that answer specific questions for an instance from the Busselton dataset; font denotes ***features***, **feature values** and *classes*. Text that points to external resources for features in *HowtoGetC′?*, *HowtoStillGetC?* and *WhatIf-Change1Factor?* has been omitted due to space constraints.

the probability of class $\mathcal{C}'$ changes when feature values are added in context. Next, for each node in the path, we compare the probability of class $\mathcal{C}'$ at this node to that at the previous node, and select the node with the largest drop in probability. For example, to generate the explanation in segment *WhyNot*$\mathcal{C}'$*?* in Table 6, we look at the DT path for the current instance (*age*: between 43 and 60 years, *HDL cholesterol*: borderline, *gender*: male and *total cholesterol*: high; the DT appears in Figure B.8), and find that the largest drop in the relative probability of the other class *low risk of a coronary event* occurs when the DT path splits on *total cholesterol* (the probability goes from 0.6 at *gender* to 0 at *total cholesterol*). The resultant explanation is that the user is *not* given the alternative prediction *low risk of a coronary event* because of his high *total cholesterol*.

***Explanation for HowtoGet*$\mathcal{C}'$*?.*** To obtain the list of feature changes that lead to a specific different prediction $\mathcal{C}'$, we look at the subset of paths in the DT that lead to this outcome. In case of a binary classification problem, as in our evaluation dataset (Section 4.2), this outcome is just the other possible class, while in case of a multi-class classification problem, it should be nominated by the user. In this work, we constrain the subset of paths that lead to $\mathcal{C}'$ by excluding paths which require the user's *age* or *gender* to be changed. This is done for the sake of brevity, and because these features usually cannot be changed, at least in the short term.

For each path that leads to $\mathcal{C}'$, we extract the set of feature values that differ from those in the current instance. If the same feature (or its value) is obtained from several paths, we combine them into one phrase, e.g., first item in segment *HowtoGet*$\mathcal{C}'$*?* in Table 6.

***Explanation for HowtoStillGet*$\mathcal{C}$*?.*** In contrast to *HowtoGet*$\mathcal{C}'$*?*, here we want to obtain the list of feature changes that retain the *Predicted* class $\mathcal{C}$. In the context of a DT, a user will get the same prediction given their profile, if they change values of individual features that are not in the current DT path or not in the DT (first item in segment *HowtoStillGet*$\mathcal{C}$*?* in Table 6). These features are obtained by removing the set of features in the current DT path (constituting our Basic explanation) from the set of features in the dataset.

A user could also get the same prediction for feature values that differ from those in the current DT path, e.g., second item in segment *HowtoStillGet*$\mathcal{C}$*?* in Table 6 (as for *HowtoGet*$\mathcal{C}'$*?*, several feature values obtained from several paths are combined into one phrase). However, it is possible that when the value of a feature in the DT path is changed, the alternative path taken contains features that were not in the previous DT path, and the values of these features may have to be changed in order to retain the *Predicted* class. An example of this can be seen in the last item in segment *HowtoStillGet*$\mathcal{C}$*?* in Table 6, where when *HDL cholesterol* (a feature in the current DT path) changes from 'borderline' to 'optimal', a feature previously not in the DT path (*daily cigarette consumption*) also needs to be changed in order to retain the predicted outcome. Both of these types of feature changes (second and third item in segment *HowtoStillGet*$\mathcal{C}$*?*

in Table 6) are extracted from the subset of paths in the DT that yield the *Predicted* class $\mathcal{C}$ by applying the procedure used for *HowtoGet*$\mathcal{C}'$*?*.

***Explanation for WhatIf-Change1Factor?.*** Here, we focus on a feature of interest to a user, and explain which changes to the value of this feature would lead to the same prediction $\mathcal{C}$ and which would lead to a specific different prediction $\mathcal{C}'$ (in case of a multi-class classification problem, we would have more than one class). If the feature of interest is in the current DT path (first option in *WhatIf-Change1Factor?* in Table 6), we first get the subset of paths that differ from the current DT path only in the value of the feature of interest, and then split this set based on whether the resultant prediction is the *Predicted* class $\mathcal{C}$ or a different class. Similarly to *HowtoGet*$\mathcal{C}'$*?* and *HowtoStillGet*$\mathcal{C}$*?*, multiple changes in the value of a feature that result in a particular prediction are combined into one phrase.

If the feature of interest is not in the current DT path or not in the DT, any change in its value will lead to the same prediction $\mathcal{C}$ (last two options in *WhatIf-Change1Factor?* in Table 6).

## 4. Experimental Setup

In this section, we describe our evaluation questions for each experiment (Section 4.1), and our datasets and classifier (Section 4.2), followed by our experimental design (Section 4.3).[10]

### *4.1. Evaluation questions*

Our evaluation for the first type of contextual information (Experiment I) looks at the influence of background information on users' views about explanations by considering two main questions:

Q1. How do Conflict-based explanations compare to Basic explanations and to each other in terms of completeness, presence of irrelevant/misleading/contradictory information, users' understanding of the AI's reasoning for a predicted outcome, their willingness to act on the prediction, and preferences?

Q2. Which independent variables influence users' views of the Conflict-based and Basic explanations?

Our evaluation for the second type of contextual information (Experiment II) looks at the influence of users' goals on their views about explanations, and considers three main questions:

Q1. How does the goal influence the selection of follow-up questions (FQs)? Specifically, (a) what are the most commonly selected FQs for a goal? and (b) do the selected FQs vary with the goal?

[10]We have addressed the recommendations for human evaluation in (Howcroft et al., 2020).

Q2. How does the goal influence users' views of the explanations generated for the six FQs and the Basic explanation in terms of completeness, presence of irrelevant/misleading/contradictory information, usefulness for the goal, and whether additional information is needed to achieve the goal?

Q3. Which independent variables influence users' views of the generated explanations?

As mentioned at the start of Section 3, the Conflict-based explanations in Experiment I contain the Basic explanation plus additional information. In contrast, the explanations that address FQs in Experiment II only convey the requested information.

### *4.2. Datasets*

We used two datasets for Experiment I, which were pre-processed as described in Appendix C.1: *Nursery* (Olave et al., 1989), which has 12630 instances and three classes; and *Telecom*, which has 3302 instances and two classes. As mentioned in Section 1, in Nursery, a DT predicts the acceptance status of a child to a childcare center on the basis of the circumstances of the child and their family; in Telecom, a DT predicts whether a customer will churn (leave) or stay with a telecommunications company based on their profile — the top two segments of Table 7 display the features of these datasets and their associated values. These datasets were chosen due to their diverse character, and the differences in number and types of features and predicted classes.

For Experiment II, we used the *Busselton* dataset (Knuiman et al., 1998), which was pre-processed as described in Appendix C.1, and has 2874 instances and two classes. This dataset contains demographic, medical and lifestyle information for a group of people, and information about whether they developed coronary heart disease (CHD) within ten years of the initial data collection (bottom segment of Table 7). The DT considers the first three types of information to predict whether a person is at a high or low risk of CHD. This dataset was chosen because we thought that the participants would be able to identify with the patients and their goals in light of predicted outcomes.

All three datasets were split into 80% training and 20% test sets using proportional sampling (we did not cross-validate, as average classifier accuracy is tangential to this research). We employed the J48 classifier (Quinlan, 1993) in WEKA (Frank et al., 2016) to learn DTs, which produced a DT with 47 nodes for the Nursery dataset (93% accuracy on the test set) and a DT with 41 nodes for Telecom (80% accuracy on the test set).[11] 78% of the Nursery test samples and all the Telecom test samples had at least one potential conflict. The Busselton dataset was imbalanced towards *low risk of a coronary event* (90%). Hence, we trained the DT using a cost-sensitive setting for imbalanced datasets, which

[11] Users are informed of a DT's overall accuracy, but not about its accuracy for individual predictions — in the future we will study the inclusion of this information in an explanation.

| **Nursery** | |
|---|---|
| *Classes*: | *Priority accept*, *Wait list*, *Reject* |
| ***parents' employment***: | challenging, somewhat difficult, ordinary |
| ***current childcare***: | very critical, critical, insufficient, sufficient, good |
| ***housing condition***: | inadequate, somewhat inadequate, adequate |
| ***social situation***: | problematic, somewhat problematic, unproblematic |
| ***child's health***: | poor, average, good |
| **Telecom** | |
| *Classes*: | *Stay*, *Churn (leave the company)* |
| ***senior citizen***: | yes, no |
| ***phone service***: | yes, no |
| ***multiple phone lines***: | yes, NA (no phone service), no |
| ***internet service***: | Fiber optic, DSL, no |
| ***online security***: | yes, NA (no internet service), no |
| ***tech support***: | yes, NA (no internet service), no |
| ***movie streaming***: | yes, NA (no internet service), no |
| ***paper billing***: | yes, no |
| ***tenure (months with company)***: | 1 $\cdots$ 72 |
| ***monthly charges ($)***: | 19 $\cdots$ 117 |
| **Busselton** | |
| *Classes*: | *Low risk of a coronary event*, *High risk of a coronary event* |
| ***age (in years)***: | 18 $\cdots$ 95 |
| ***gender***: | female, male |
| ***weight status***: | optimal, underweight, overweight, obese |
| ***daily alcohol intake (standard drinks)***: | 0 $\cdots$ 44 |
| ***daily cigarette consumption***: | 0 $\cdots$ 75 |
| ***blood pressure***: | optimal, normal-to-high, high |
| ***total cholesterol***: | low, normal, borderline, high |
| ***HDL cholesterol***: | optimal, borderline, low |
| ***triglycerides***: | low, normal, borderline, high |
| ***diabetes***: | no, yes |

Table 7: *Classes*, ***features*** (in the presentation order used in our explanations – ***age*** and ***gender*** are interchangeable) and their associated values in the evaluation datasets; the feature values in the Nursery dataset are described in Table C.20.

yielded a DT with 38 nodes (82% accuracy on the test set).[12] The DTs for the three datasets appear in Appendix B.

### *4.3. Experimental Design*

Both experiments started with a demographic questionnaire followed by the body of the survey, which consisted of the following components: an immersive narrative, a brief account of how an AI makes predictions plus the features

[12]Since we wanted the DT to produce credible results, and debugging a DT was not one of the goals given to users, we pruned two nodes which seemed unintuitive and had a very high inaccuracy for the minority class.

and values that were input to the AI (Table 7), and a sequence of scenarios presented in random order. The scenarios were based on our testsets, not on the subjects' own data. Each scenario began by showing a set of features from Table 7, together with their values for a particular family/customer (Experiment I) or patient (Experiment II). For each scenario, users were asked to make an educated guess about the outcome, and then they were shown the actual outcome followed by explanations, which were evaluated in terms of explanatory attributes. The attributes in common to both experiments are completeness of an explanation and presence of irrelevant/misleading/contradictory information, and come from the *Explanation Satisfaction Scale* in (Hoffman et al., 2018). The experiment-specific attributes are described in Sections 4.3.1 and 4.3.2. To detect unreliable responses, we inserted attention questions relevant to each scenario, which were True/False or multiple-choice.

We now provide details of the main body of the survey for each experiment (Sections 4.3.1 and 4.3.2), and describe the participant cohorts (Section 4.3.3).

#### *4.3.1. Experiment I – Influence of background information*

In the immersive narrative for Experiment I, participants were told that they are the director of a childcare center (Nursery) or the sales representative of a telecommunications company (Telecom), and that they have purchased an AI system to help them predict the acceptance status of prospective pupils (Nursery) or whether customers will churn (leave) or stay (Telecom) – Figure D.9 shows a screenshot of the narrative for the Nursery dataset. As mentioned above, users were then shown a sequence of scenarios. Between scenarios, a short version of the Matching Familiar Figures Test (MFFT) (Cairns and Cammock, 1978) was shown as a filler.

*Scenario description.* We chose scenarios with the strongest available potential conflict (using a procedure similar to that described in Section 3.1.2), and diverse pivot and explanatory variables. Scenarios without conflicts were excluded from our evaluation, as only a Basic explanation can be generated for them. To ensure that all the potential conflicts in Table 3 are represented, we chose eight Nursery scenarios (four each for *Wait list* and *Priority accept*)[13] and ten Telecom scenarios (five each for *Stay* and *Churn*).

As mentioned above, each scenario began by showing a set of features from Table 7, together with their values for a particular family/customer. We then showed the *Prior* and *Posterior* probabilities of the possible classes for these feature values. A screenshot of a Nursery scenario appears in Figure D.10.

*Users' views about explanations.* After users guessed the outcome, they were shown the prediction made by the DT, and were given two side-by-side explanations for this prediction: Conflict-based versus Basic. The selection of a side

[13] Examples for *Reject* were not presented, as there was only one reason to reject applicants: poor health.

(left or right) for an explanation type was randomized between scenarios, but all the participants saw the same side-by-side configuration for a given scenario.

Users were then asked to enter their level of agreement on a 5-point Likert scale ('Strongly disagree':1 to 'Strongly agree':5) with statements about four explanatory attributes: completeness of an explanation and presence of irrelevant/misleading/contradictory information, as well as users' understanding of the AI's reasoning for the predicted outcome and their willingness to act on the prediction on the basis of an explanation (exact statements appear in the screenshot in Figure D.10). The third and fourth attributes were used to determine the *post hoc* effect of our explanations on two common goals of explanations (Section 1) — the users were not told that the explanations were generated to help them achieve these goals. Participants were also asked which explanation(s) they preferred, if any.

#### *4.3.2. Experiment II – Influence of users' goals*

In the introduction to Experiment II, participants were told that a health consultancy has purchased an AI system that predicts whether a particular patient is at a high or low risk of a coronary event – a screenshot of the narrative appears in Figure D.11. Next, three profiles were presented in random order, each pertaining to a different patient. For each profile, we asked participants to pretend that they are the patient in the profile.[14]

*Profile description.* In realistic situations, users have their own goals. In particular, people would want to change undesirable outcomes and retain desirable ones. However, to ensure adequate representation of the three goals in our experiment, we provided users with goals. Owing to the length of the experiment, we chose one profile from the *low risk* class, and associated it with the goal *retain the predicted outcome*, and two profiles from the *high risk* class, associating them with the goals *understand the AI's reasoning for the predicted outcome* and *change the predicted outcome*. The profiles were chosen so that they yield diverse explanations and explanatory variables. However, having each goal associated with a different patient's profile poses a risk whereby the features of a profile could influence our findings (Section 5.2). In the future, we plan to address this issue by swapping the goals associated with the profiles and including additional profiles.

*Users' views about explanations.* Figure 2 depicts the workflow we employed for a profile (the screenshot in Figure D.12 illustrates the initial steps of our workflow). After users guessed the outcome for a particular profile, they were shown the prediction made by the DT, and given a Basic explanation for that prediction, followed by the goal associated with the profile — the goal was presented after the users had guessed the outcome so as not to preempt their expectations.

[14] Unlike Experiment I, here we did not use MFFTs between profiles because there were only three profiles.

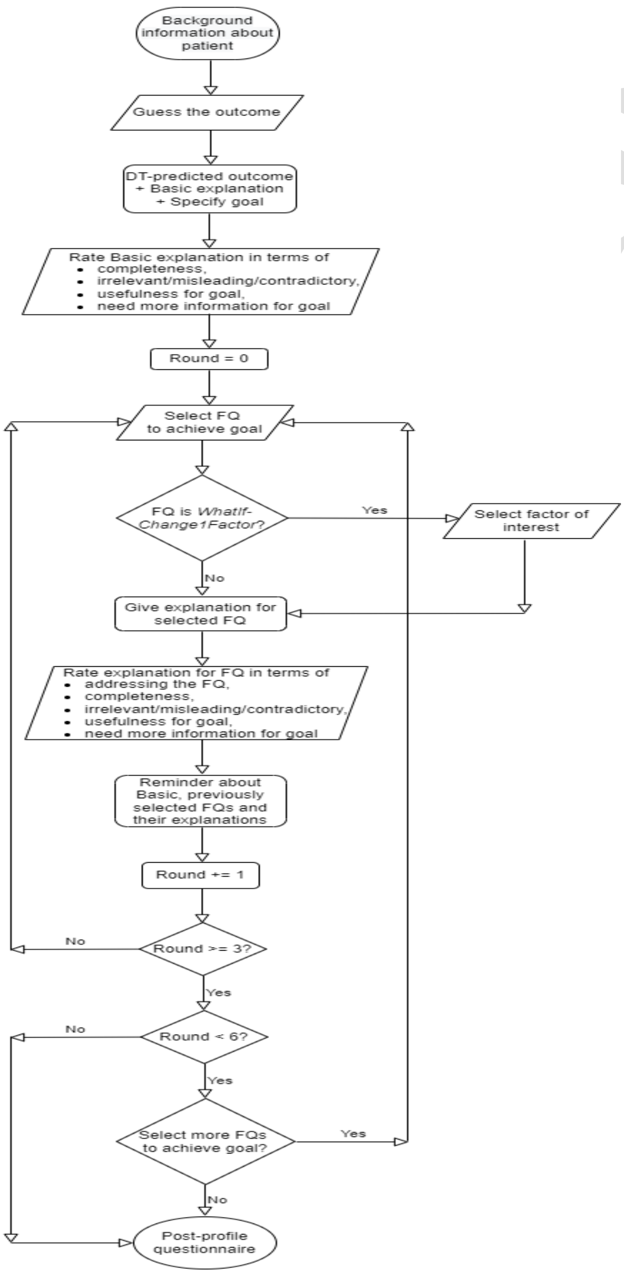



Figure 2: Workflow of a profile in Experiment II.

Users were then asked to enter their level of agreement on a 7-point Likert scale ('Strongly disagree':1 to 'Strongly agree':7) with statements about four explanatory attributes with respect to the Basic explanation:[15] completeness, presence of irrelevant/misleading/contradictory information, usefulness for their assigned goal and whether users needed more information to achieve their goal in light of this explanation (exact statements appear in the screenshot in Figure D.12). The first two attributes were also evaluated in Experiment I, while the third and fourth attributes are specific to the objective of this experiment.

Once users rated the Basic explanation, they were iteratively asked to select at least three FQs to help them achieve their assigned goal (bottom part of Figure D.12); if users selected *WhatIf-Change1Factor?*, they also had to choose the factor whose impact they were interested in, excluding *age* and *gender* (Figure 2). After a question was selected, we presented an answer. Users were then asked to enter their level of agreement on a 7-point scale with the statement "the explanation addresses the selected question", and to rate the explanation in terms of the four explanatory attributes they used to rate the Basic explanation. Before selecting another question, users were reminded of the Basic explanation, and of all the FQs they had selected so far and their answers.

After completing the three mandatory rounds of question selection, users could select more questions from the remaining FQs or proceed to the next patient profile. Before proceeding to the next profile, users were asked to enter their level of agreement on a 7-point Likert scale with two statements: "the explanations increased their confidence in the AI system" and "the explanations helped them achieve their goal". In addition, users were asked about the extent to which they could identify with the patient's profile ('Could not identify at all':1 to 'Identify a lot':5).

*4.3.3. Participant cohorts*

Both experiments were implemented in the Qualtrics survey software. Experiment I was conducted on SONA, while Experiment II was conducted on CloudResearch (Litman et al., 2017).[16] To avoid participant fatigue in Experiment I, we conducted a separate survey for each dataset – details appear in Appendix C.2.

Both experiments had about 76% valid responses — 83 out of 109 for Experiment I (41 for Nursery and 42 for Telecom), and 89 out of 116 for Experiment II. Responses were validated based on the answers to the attention questions and the total time spent on the experiment. Table 8 shows population statistics for the Nursery and Telecom cohorts, and Table 9 displays population statistics for

[15] In light of our experience from Experiment I, where extreme values of the ratings of explanatory attributes (1 and 5) were chosen only 11% of the time, we decided to expand the Likert scale for these attributes to 7 points for Experiment II, which is in line with recent best practice recommendations in (van der Lee et al., 2021).

[16] We chose a different platform for the second experiment to recruit users from a broader population, and to expedite the experiment. As seen in Tables 8 and 9, we obtained a different, but not necessarily broader, population.

| User Information | Option | No. of users | |
|---|---|---|---|
| | | Nursery 41 | Telecom 42 |
| Gender | Female / Male | 28 / 12 | 17 / 25 |
| Age | 25-34 years old / 18-24 years old | 20 / 13 | 19 / 18 |
| Ethnicity | Asian / Caucasian / Middle Eastern | 17 / 17 / 2 | 28 / 4 / 5 |
| Place of residence | Australia | 36 | 39 |
| English proficiency | High / Medium | 36 / 5 | 37 / 5 |
| Education | Master / Bachelor | 13 / 13 | 22 / 14 |
| ML expertise | Low / Medium-High | 27 / 14 | 18 / 24 |
| Domain familiarity | Yes / No | 9 / 32 | 31 / 11 |

Table 8: Descriptive statistics for Experiment I: for gender, age, ethnicity, place of residence, English proficiency and education, we present the options that had most participants; domain familiarity was self-rated (for Nursery, we asked users if they have/had a child in an early-education facility or if they have worked in such a facility, and for Telecom, we asked users to rate their familiarity with the operations of a telecommunications provider on a 5-point Likert scale — users were deemed familiar with the domain if they gave a rating of 3 or above).

| User Information | Option | No. of users 89 |
|---|---|---|
| Gender | Female / Male | 55 / 33 |
| Age | 25-34 years old / 35-44 years old | 36 / 27 |
| Ethnicity | Caucasian / African | 63 / 15 |
| Place of residence | North America | 88 |
| English proficiency | High | 88 |
| Education | Bachelor / Some college but no degree | 42 / 22 |
| ML expertise | Low / Medium | 41 / 40 |
| Risk of a coronary event | Somewhat / Slightly / Moderately concerned | 28 / 22 / 18 |

Table 9: Descriptive statistics for Experiment II: for all information items, we present the options that had most participants.

the Busselton cohort.

## 5. Experimental Results

In this section, we describe the analysis methodology and results for Experiment I (Section 5.1) and Experiment II (Section 5.2).

### *5.1. Experiment I – Influence of background information*

As mentioned in Section 4, for this experiment we address the following questions:

Q1. How do Conflict-based explanations compare to Basic explanations and to each other in terms of completeness, presence of irrelevant/misleading/contradictory information, users' understanding of the AI's reasoning for a predicted outcome, their willingness to act on the prediction, and preferences?

Q2. Which independent variables influence users' views of the Conflict-based and Basic explanations?

These questions are addressed as follows:

- *Q1.* For each dataset, we compare Conflict-based explanations with Basic ones, and compare between individual Conflict-based explanations, in terms of the four explanatory attributes (Section 5.1.1).[17] The comparison between Conflict-based explanations is indirect, as we only have ratings and preferences for Conflict-based versus Basic explanations. Nonetheless, we believe that such a comparison sheds light on the merit of individual Conflict-based explanations.
- *Q2.* We analyze the influence of (dis)agreement between a user-expected class and that predicted by a DT on users' views of Conflict-based explanations compared to Basic ones (Section 5.1.2). Our experiment had other independent variables, including predicted outcome, pivot feature and explanation length. The first two variables are scenario-specific, and hence offer no opportunities to draw generalizable conclusions. Regarding explanation length, Lombrozo (2016) reported that users generally prefer longer explanations, in particular when they include jargon. However, in our case, length is highly correlated with explanation type — Conflict-based explanations have 60 words on average in both Nursery and Telecom, and Basic explanations have 29 words. Hence, we cannot analyze length separately from explanation type. Nonetheless, our results suggest that length cannot be the only factor influencing users' views, as some types of Conflict-based explanations have similar preferences to their Basic counterparts (Table 12).

Statistical significance for the ratings of the four explanatory attributes for Conflict-based versus Basic explanations is obtained using Wilcoxon signed-rank test for paired data. When comparing between individual Conflict-based explanations for each attribute, we first obtain the statistical significance of the ratings using the Kruskal-Wallis test for more than two categories of unpaired data. In case of significance (*p-value* $< 0.05$), we follow up with pairwise comparisons between the Conflict-based explanation types using the Wilcoxon rank-sum test. A one- and two-proportion Z-test is respectively used for comparing the proportion of preference counts within one population and between two populations. Statistical significances are adjusted with Holm-Bonferroni (HB) correction for multiple comparisons (Holm, 1979).

[17] For both datasets, each Conflict-based explanation was evaluated on 1-4 scenarios depending on the representativeness of the conflict in question in the dataset, with most Conflict-based explanations appearing in two scenarios.

| Attribute | Conflict-based Mean (SD) | Basic Mean (SD) | Stat. Sig. |
|---|---|---|---|
| **Nursery** | | | |
| Complete | 3.43 (0.97) | 3.00 (0.98) | $< 0.001$ |
| Irrelevant/misleading/contradictory | 2.72 (1.00) | 2.55 (0.89) | $< 0.05$ |
| Understand the AI's reasoning | 3.61 (1.04) | 3.02 (1.03) | $< 0.001$ |
| Willingness to act | 3.56 (1.01) | 3.23 (1.01) | $< 0.001$ |
| **Telecom** | | | |
| Complete | 3.22 (0.99) | 2.93 (0.97) | $< 0.001$ |
| Irrelevant/misleading/contradictory | 3.00 (1.14) | 2.81 (1.05) | – |
| Understand the AI's reasoning | 3.49 (0.92) | 3.33 (0.87) | – |
| Willingness to act | 3.16 (0.99) | 3.09 (0.94) | – |

Table 10: Comparison between Conflict-based and Basic explanation types: scores and statistical significances (Wilcoxon signed-rank test); a lower score is better for Irrelevant/misleading/contradictory, and a higher score is better for the other attributes.

#### *5.1.1. Q1: Comparison of different explanation types*

In this section, we present our results for the comparison of the Conflict-based explanations with the Basic explanations in terms of the four explanatory attributes and users' preferences. We then analyze how individual Conflict-based explanations compare to each other.

***Conflict-based explanations versus Basic explanations.*** Our results show that for the Nursery dataset (top of Table 10), Conflict-based explanations were deemed significantly more complete, more helpful for understanding the AI's reasoning and more enticing to act on a DT's prediction than Basic explanations. However, Conflict-based explanations were also deemed to contain more irrelevant/misleading/contradictory information than Basic explanations (as shown in Section 5.1.2, this happens when predictions match users' expectations, as the additional information provided by Conflict-based explanations is likely deemed superfluous by the users in this case). For Telecom (bottom of Table 10), Conflict-based explanations were considered significantly more complete than Basic explanations, but equivalent for the other three attributes. For both datasets, we found a strong positive Spearman correlation between users' ratings for the goal of understanding the AI's reasoning and the goal of motivating users to act on a prediction (Nursery $\rho = 0.62$, Telecom $\rho = 0.64$, *p-value* $\ll 0.01$ for both).

In terms of preferences, for both datasets, most users preferred Conflict-based explanations to Basic ones (Table 11). However, the two datasets differed significantly in the proportions of preferences for Conflict-based explanations (two-proportion Z-test, *p-value* $< 0.05$; proportions calculated from the data in Table 11), with a higher percentage of users preferring the Conflict-based explanations for the Nursery dataset.

**Finding 1.** *Explanations that address potential conflicts are generally preferred*

| | Count | | | | | $\chi^2$ | Stat. Sig |
|---|---|---|---|---|---|---|---|
| | Conflict-based | Basic | Both | None | Total | | |
| **Nursery** | 112 | 45 | 13 | 35 | 205 | 28.59 | $< 0.001$ |
| **Telecom** | 117 | 78 | 11 | 46 | 252 | 7.80 | $< 0.01$ |

Table 11: Preference for an explanation type: $\chi^2$ statistic and statistical significances (one-proportion Z-test) calculated from clear preferences for Conflict-based/Basic explanations.

| **Basic vs Conflict-based** | Count | | | | |
|---|---|---|---|---|---|
| | Conflict-based | Basic | Both | None | Total |
| **Nursery** | | | | | |
| Basic vs *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ | 33 | 12 | 3 | 14 | 62 |
| Basic vs *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* | 8 | 6 | 1 | 6 | 21 |
| Basic vs *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" | 33 | 13 | 6 | 9 | 61 |
| Basic vs *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* | 38 | 14 | 3 | 6 | 61 |
| **Telecom** | | | | | |
| Basic vs *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ | 46 | 21 | 2 | 15 | 84 |
| Basic vs *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* | 14 | 20 | 2 | 6 | 42 |
| Basic vs *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" | 23 | 6 | 3 | 10 | 42 |
| Basic vs *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* | 34 | 31 | 4 | 15 | 84 |

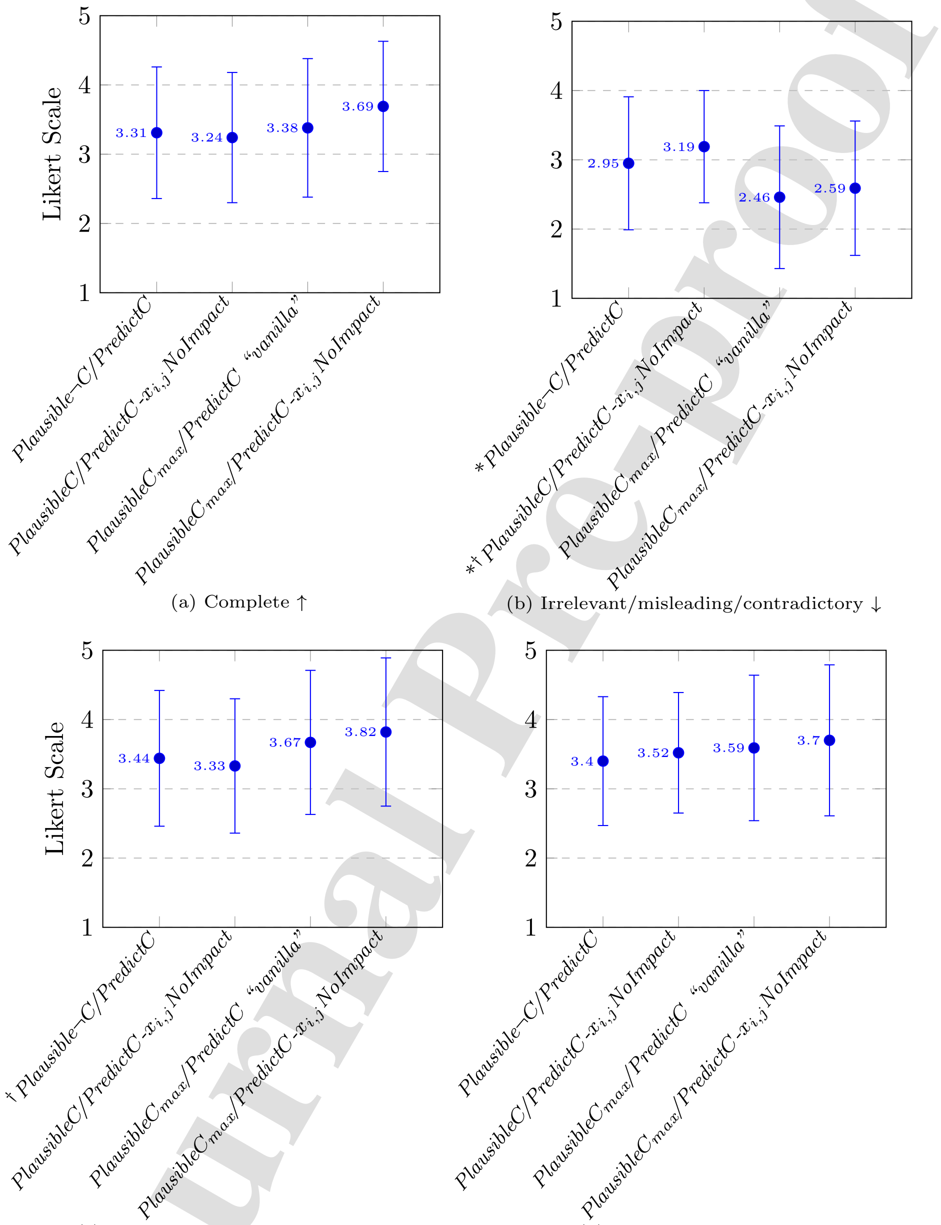
Table 12: Preference for individual Conflict-based explanations and their Basic counterparts.

*to Basic explanations, and are considered at least as good as Basic explanations for three of the four explanatory attributes.*

***Individual Conflict-based explanations.*** Here, we analyze how the individual Conflict-based explanations compare to each other in terms of the four explanatory attributes and users' preferences.

For the Nursery dataset, we found a significant difference between individual Conflict-based explanations for presence of irrelevant/misleading/contradictory information and for users' understanding of the AI's reasoning (Kruskal-Wallis test, *p-value* $< 0.01, 0.05$ respectively). Specifically, in terms of irrelevant/misleading/contradictory information, users deemed *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* worse than the two variants of *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$, and *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ worse than *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" (Figure 3b). In terms of understanding the AI's reasoning, the only difference was that users found *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* more helpful than *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ (Figure 3c). In contrast, for the Telecom dataset, we did not find significant differences between the ratings for the individual Conflict-based explanations for any of the explanatory attributes (Figure E.13).

Looking at preferences, *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ and *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" were preferred to their Basic counterparts for both datasets, while *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImpact* was preferred to the Basic explanation only for Nursery (Table 12). Comparing between Conflict-based explanations, for the Nursery dataset, there were no significant differences in the proportion




(a) Complete ↑

(b) Irrelevant/misleading/contradictory ↓

(c) Understand the AI's reasoning ↑

(d) Willingness to act ↑

Figure 3: Comparison between individual Conflict-based explanations for the Nursery dataset (sample sizes in *Total* column, Table 12): mean and standard deviation of ratings for the four explanatory attributes; ↑ / ↓ indicates that a higher / lower score is better for an attribute. Significant differences between an explanation type and *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" (Wilcoxon rank-sum test after HB correction) are denoted as * (*p-value* $< 0.05$), and significant differences between an explanation type and *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact* are denoted as † (*p-value* $< 0.05$).

of users who preferred individual Conflict-based explanations (two-proportion Z-test), despite the significant differences in users' ratings of two attributes for individual Conflict-based explanations (Figure 3). In contrast, for the Telecom dataset, a statistically significantly higher proportion of users preferred *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$ "vanilla" and *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$ to *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact* (two-proportion Z-test, *p-value* $< 0.05$, data in Table 12), even though there were no significant differences in attribute ratings for individual Conflict-based explanations. This points to a discrepancy between users' ratings of explanatory attributes and their overall preferences, which warrants further investigation.

Based on the analysis of the users' ratings (Figures 3 and E.13) and their preferences (Table 12), we propose the following finding.

**Finding 2.** *If a DT prediction has several qualifying conflicts, they should be prioritized in the following order:* Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$ "vanilla" $\succ$ Plausible$\neg\mathcal{C}$/Predict$\mathcal{C}$ $\succ$ Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$-$x_{i,j}$NoImpact.

For both datasets, we found *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$*NoImpact* to be lacking, which somewhat disagrees with the finding in (Biran and McKeown, 2017) whereby users were more satisfied with explanations about unexpected feature impacts than no explanation. This suggests that further studies are required to determine the conditions for explaining unexpected feature impacts when the outcome is expected.

*5.1.2. Q2: Influence of independent variables on users' views about explanations*

***(Dis)agreement between users' expectations and DT predictions.*** Our analysis shows that (dis)agreement between users' expectations (according to their survey answers) and the class predicted by a DT had a significant influence on their ratings for Conflict-based explanations compared to Basic ones (users' answers disagreed with a predicted class when they selected a different class or *Can't Decide* – options appear in Figure D.10).

For the Nursery dataset, the general results obtained for Conflict-based versus Basic explanations hold for completeness, users' understanding of the AI's reasoning and their willingness to act on predictions for both agreement and disagreement between users' expectations and DT predictions (top of Table 13). However, Conflict-based explanations were deemed more irrelevant/misleading/contradictory than Basic explanations only when users' expectations matched DT predictions (Conflict-based explanations were deemed equivalent to Basic ones if their expectations disagreed with DT predictions).

For the Telecom dataset, Conflict-based explanations were considered more complete and enticing to act only when users' expectations differed from DT predictions (bottom of Table 13).

In terms of preferences (Table E.22), most users preferred Conflict-based explanations to Basic ones for the Nursery dataset, regardless of the agreement between users' expectations and DT predictions (*p-value* $< 0.001$). However, for Telecom, Conflict-based explanations were preferred to Basic explanations only when users' expectations disagreed with DT predictions (*p-value* $< 0.001$).

| Attribute | Predict vs Expect | Conflict-based Mean (SD) | Basic Mean (SD) | Stat. Sig. |
|---|---|---|---|---|
| **Nursery** | | | | |
| Complete | Pred = Exp | 3.41 (0.96) | 3.04 (0.97) | $<0.01$ |
| | Pred $\neq$ Exp | 3.48 (0.99) | 2.90 (0.99) | $<0.01$ |
| Irrelevant/misleading/contradictory | Pred = Exp | 2.80 (1.03) | 2.54 (0.90) | $<0.05$ |
| | Pred $\neq$ Exp | 2.57 (0.92) | 2.57 (0.86) | – |
| Understand the AI's reasoning | Pred = Exp | 3.61 (1.07) | 3.20 (0.99) | $<0.01$ |
| | Pred $\neq$ Exp | 3.61 (0.97) | 2.66 (1.01) | $<0.001$ |
| Willingness to act | Pred = Exp | 3.64 (0.95) | 3.41 (0.98) | $<0.05$ |
| | Pred $\neq$ Exp | 3.40 (1.12) | 2.87 (0.98) | $<0.01$ |
| **Telecom** | | | | |
| Complete | Pred = Exp | 3.18 (0.97) | 2.99 (0.95) | – |
| | Pred $\neq$ Exp | 3.35 (1.04) | 2.72 (1.01) | $<0.01$ |
| Irrelevant/misleading/contradictory | Pred = Exp | 3.08 (1.14) | 2.83 (1.05) | – |
| | Pred $\neq$ Exp | 2.75 (1.10) | 2.75 (1.08) | – |
| Understand the AI's reasoning | Pred = Exp | 3.45 (0.90) | 3.35 (0.86) | – |
| | Pred $\neq$ Exp | 3.62 (0.98) | 3.25 (0.93) | – |
| Willingness to act | Pred = Exp | 3.14 (0.97) | 3.17 (0.90) | – |
| | Pred $\neq$ Exp | 3.25 (1.07) | 2.83 (1.04) | $<0.05$ |

Table 13: Effect of (dis)agreement between users' expectations and DT predictions: scores and statistical significances (Wilcoxon signed-rank test).

**Finding 3.** *Conflict-based explanations are deemed especially valuable when the outcome expected by users disagrees with DT predictions.*

### *5.2. Experiment II – Influence of users' goals*

As mentioned in Section 4, for this experiment we consider the following questions for the goals of *understanding the AI's reasoning*, *changing the predicted outcome* and *retaining the predicted outcome*:[18]

Q1. How does the goal influence the selection of follow-up questions (FQs)? Specifically, (a) what are the most commonly selected FQs for a goal? and (b) do the selected FQs vary with the goal?

Q2. How does the goal influence users' views of the explanations generated for the six FQs and the Basic explanation in terms of completeness, presence of irrelevant/misleading/contradictory information, usefulness for the goal, and whether additional information is needed to achieve the goal?

Q3. Which independent variables influence users' views of the generated explanations?

[18] Our analysis includes data for the initial Basic explanation and the explanations associated with the FQs selected in the three mandatory rounds, because a fourth FQ was selected in only 5% of the 267 data points (89 users attempting three goals).

These questions are addressed as follows.

***Q1.*** We apply the following algorithms.

- *Q1a.* We use the *Markov Chain3* (*MC3*) algorithm (Lin, 2010) to determine an aggregate ranking of FQs for a particular goal (Section 5.2.1). This algorithm constructs a transition probability matrix such that the probability of going from $\mathrm{FQ}_i$ to $\mathrm{FQ}_j$ is proportional to the number of users that gave $\mathrm{FQ}_j$ a better ranking than $\mathrm{FQ}_i$. That is, transition probabilities represent pairwise rankings, and the steady state transition-probability matrix represents the aggregate rankings of the different FQs — the higher the steady state probability the better the rank.

- *Q1b.* We employ *Rank Biased Overlap* (*RBO*) (Webber et al., 2010) to determine the extent of the overlap between the order in which the FQs were selected for different goals (Section 5.2.1). RBO assigns a weight to each ranked position, and computes the weighted similarity between two ordered lists; the result is in the $[0, 1]$ range, where 0 means disjoint ordered lists and 1 means identical ones. For each pair of goals, $G_1$ and $G_2$, we compute the RBO between the list of FQs selected by each user for $G_1$ and the list of FQs selected for $G_2$; we then average the RBO for all users to obtain the average overlap between FQs for the two goals.[19]

***Q2 and Q3.***

- *Q2.* For each goal, we compare between the eight explanation types (Basic, two types for *WhatIf-Change1Factor?*[20] and one type for each of the remaining five FQs) in terms of the four explanatory attributes (completeness, presence of irrelevant/misleading/contradictory information, usefulness for the goal and needing more information to achieve the goal; Section 5.2.2). It is worth noting that unlike Experiment I, the explanations generated for the FQs are presented after a Basic explanation (not in direct comparison with it), and answer specific questions. Nonetheless, we compare the ratings of follow-up explanations with those of their initial Basic explanation to set up a reference point for our results.

- *Q3.* We analyze the influence of three independent variables on each explanatory attribute (Section 5.2.3): (1) whether an explanation addresses

[19]An alternative is to compute the overlap between the selected FQs without considering the order in which they were selected by a user. However, this would not be an accurate reflection of the rankings obtained from the MC3 algorithm, because MC3 takes ordering into account.

[20]The two types for *WhatIf-Change1Factor?* are referred to as *InPath* (if the feature of interest is in the current DT path) and *NotInPath* (if the feature of interest is not in the current DT path or not in the DT). The latter type combines two explanations (last two options in the *WhatIf-Change1Factor?* segment in Table 6), because only 14% of the features nominated by the users who selected *WhatIf-Change1Factor?* were not in the DT (Table E.23).

| Goal | Aggregated Ranking |
|---|---|
| *Understand the AI's reasoning* | ***HowtoGetC′?***, ***FactorsNotUsed?***, ***WhyNotC′?***, *WhatIf-Change1Factor?*, *FactorsUsed?*, *HowtoStillGetC?* |
| *Change the predicted outcome* | ***HowtoGetC′?***, ***WhatIf-Change1Factor?***, ***FactorsUsed?***, *WhyNotC′?*, *FactorsNotUsed?*, *HowtoStillGetC?* |
| *Retain the predicted outcome* | ***HowtoStillGetC?***, ***HowtoGetC′?***, ***FactorsUsed?***, *WhatIf-Change1Factor?*, *FactorsNotUsed?*, *WhyNotC′?* |

Table 14: Aggregated ranking of FQs produced by the MC3 algorithm for each goal; the top-three questions are in ***boldface-italics***.

the selected question (only for FQs – 7-point Likert scale), (2) the selection round for the FQs (first, second, third), and (3) explanation length (short, medium, long).[21] In light of the results obtained in Experiment I, we also planned to analyze the impact of (dis)agreement between a user-expected and a DT-predicted class on the explanation ratings. However, unlike Experiment I, here only 13% of the cases had a disagreement between the expected and predicted class, so we excluded this variable from our analysis.

For the categorical independent variables with more than two categories, explanation type (eight categories) and explanation length (three categories), we first obtain the statistical significance of the ratings for an explanatory attribute using the Kruskal-Wallis test for unpaired data. In case of significance ($p\text{-}value < 0.05$), we follow up with pairwise comparisons between the different categories of a variable using the Wilcoxon rank-sum test. When analyzing the influence of the FQ-selection round, we perform pairwise comparisons between the three rounds using the Wilcoxon signed-rank test for paired data. Statistical significances are adjusted with Holm-Bonferroni (HB) correction for multiple comparisons (Holm, 1979). Finally, for the numerical independent variable that represents users' agreement with "the explanation addresses the selected question", we use Spearman correlation, as we are interested in the general trend of how the ratings given to the explanations vary with the extent of this agreement.

#### *5.2.1. Q1: Influence of users' goals on the selection of FQs*

***Q1a.*** Table 14 shows the ranking of the FQs produced by the MC3 algorithm for each goal (the top-three FQs appear in ***boldface-italics***). As seen in Table 14, *HowtoGetC′?* was highly ranked for all the goals, which indicates that people are generally curious about alternative outcomes, even if they are not directly relevant to their goals. Further, the top-three FQs for *understanding the*

[21] We converted explanation length to categories by taking the 33rd and 66th percentile of the lengths of the generated explanations. The explanations with 39 words or less (33rd percentile; 9 explanations) fall in the 'short' category, those with 40-49 words (33rd - 66th percentile; 17 explanations) fall in the 'medium' category, and the remaining explanations (50-109 words; 12 explanations) fall in the 'long' category.

| Goal Pair | | | Mean (SD) |
|---|---|---|---|
| *Understand the AI's reasoning* | – | *Change the predicted outcome* | 0.39 (0.26) |
| *Change the predicted outcome* | – | *Retain the predicted outcome* | 0.40 (0.27) |
| *Understand the AI's reasoning* | – | *Retain the predicted outcome* | 0.34 (0.25) |

Table 15: Overlap (order dependent) produced by RBO between the FQs selected by the users for each pair of goals.

*AI's reasoning for the predicted outcome* are about information that is complementary to the current situation, i.e., an alternative outcome and factors that were not used, which make up the Conflict-based explanations in Experiment I. In addition, *WhatIf-Change1Factor?* was highly ranked for the goal of *changing the predicted outcome*, while *WhyNot*$\mathcal{C}'$*?* was not among the top-ranked options for this goal. Finally, as one would expect, *HowtoStillGet*$\mathcal{C}$*?* was highly ranked for *retaining the predicted outcome*, but was of little import for the other goals.

***Q1b.*** Table 15 shows the average overlap produced by RBO between the FQs selected by the users for each pair of goals. As seen in Table 15, even though there is some overlap, there are enough differences to warrant tailoring explanations to users' goals.

**Finding 4.** *There is some overlap between the* FQs *selected for the three goals, with* HowtoGet$\mathcal{C}'$? *being the most selected question for all the goals.*

The results in Table 14 provide general guidelines for the explanation types to be included in explanations generated for particular goals. However, these results are based on users' FQ selections, and do not take into account whether the explanations that address these questions actually helped users achieve their goals. In the future, we plan to remedy this shortcoming by combining FQ selection order for a goal with the ratings given to the associated explanations, and the final ratings users gave for whether the explanations helped them achieve the goal (Section 4.3.2).

#### *5.2.2. Q2: Influence of users' goals on their views about explanations*

Overall, the ratings of the eight explanation types for the goal of *changing the predicted outcome* are quite variable, while the ratings for the other two goals are more stable. Specifically, the results of the Kruskal-Wallis test that compares the ratings of the explanation types for each explanatory attribute show that for the goal of *changing the predicted outcome*, there were significant differences in the ratings of the explanation types in terms of completeness, usefulness for the goal and need for additional information to achieve the goal (*p-value* $<$ 0.001); for the goal of *understanding the AI's reasoning*, there were significant differences only for completeness (*p-value* $< 0.05$); and for the goal of *retaining the predicted outcome*, all the explanation types were deemed equivalent for all the explanatory attributes (*p-value* $> 0.05$).

Figures 4 and 5 show the results of further analysis of the explanation ratings for the goals of *understanding the AI's reasoning and changing the predicted outcome — no further analysis was performed for retaining the outcome (Figure E.14 compares the explanation types for this goal).* The significant results are as follows. For the goal of *understanding the AI's reasoning*, only the explanation for *WhatIf-Change1Factor?-InPath* was deemed significantly more complete than the Basic explanation (Figure 4a). For the goal of *changing the predicted outcome*, both of the general explanation types (*FactorsUsed?* and *FactorsNotUsed?*) and three profile-specific types (*HowtoGet*$\mathcal{C}'$*?*, *HowtoStillGet*$\mathcal{C}$*?* and *WhatIf-Change1Factor?-NotInPath*) were deemed more complete than the Basic explanation (Figure 5a). In addition, users thought that the explanation for *HowtoGet*$\mathcal{C}'$*?* was more useful than the Basic explanation for this goal (Figure 5c), and they disagreed more with requiring additional information to achieve this goal for the explanations presented for *HowtoGet*$\mathcal{C}'$*?* and *HowtoStillGet*$\mathcal{C}$*?* than for the Basic explanation (Figure 5d). Comparing between the explanations that address the FQs for the goal of *changing the predicted outcome*, the explanations for *FactorsNotUsed?* and *WhatIf-Change1Factor?-NotInPath* were found to be less useful for this goal than *HowtoGet*$\mathcal{C}'$*?* (Figure 5c), and the explanation for *FactorsNotUsed?* was deemed worse than that for *HowtoStillGet*$\mathcal{C}$*?* in terms of requiring additional information (Figure 5d). The poor results of *FactorsNotUsed?* are intuitively appealing, as the goal is to *change the predicted outcome*, and this explanation type discusses features that are not in the DT.

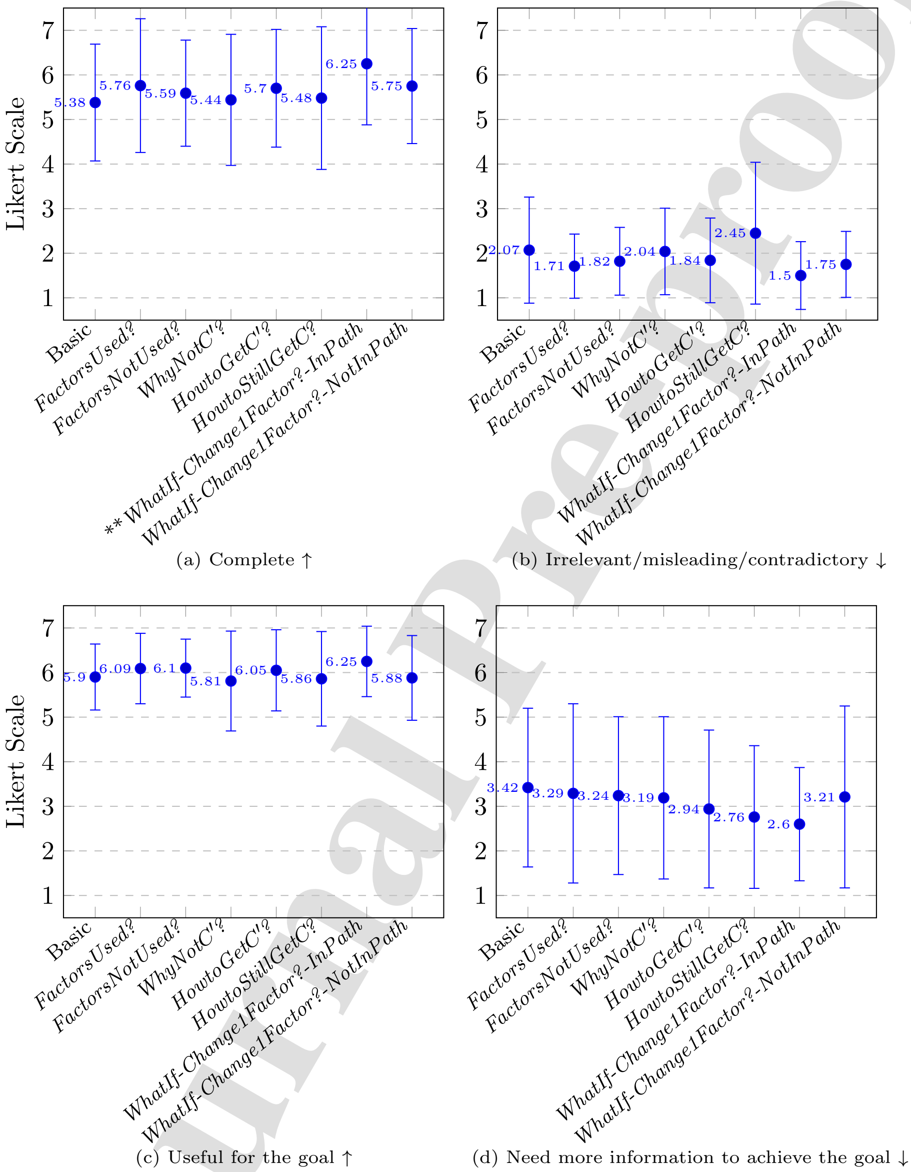
**Finding 5.** *The explanation that addresses* HowtoGet$\mathcal{C}'$? *is the most useful one for the goal of* changing the predicted outcome*; this explanation is also well regarded for this goal in terms of completeness, irrelevant/misleading/contradictory information (low rating) and need for additional information (low rating).*

**Finding 6.** *In general, all the explanation types are similarly regarded for the goals of* understanding the AI's reasoning *and* retaining the predicted outcome *in terms of all four explanatory attributes.*

*5.2.3. Q3: Influence of independent variables on users' views about explanations*

In this section, we discuss our findings about the influence of whether an explanation addresses a selected question, FQ-selection round and explanation length on users' ratings of the four explanatory attributes.

***Influence of whether an explanation addresses a selected question.*** Intuitively, one would expect an explanation that addresses a question selected by a user to be well regarded in terms of all the explanatory attributes. Indeed, we found a strong Spearman correlation between users' ratings of the extent to which an explanation addresses a selected question (97% of the explanations had a rating of 5 or higher) and their ratings of completeness ($\rho = 0.67$) and usefulness for the goal ($\rho = 0.63$), and a moderate negative correlation




(a) Complete ↑

(b) Irrelevant/misleading/contradictory ↓

(c) Useful for the goal ↑

(d) Need more information to achieve the goal ↓

Figure 4: Comparison between explanation types for *understanding the AI's reasoning for the predicted outcome* (sample sizes in Table E.24): mean and standard deviation of ratings for the four explanatory attributes; ↑ / ↓ indicates that a higher / lower score is better for an attribute. Statistically significant differences between our explanation types and the Basic explanation (Wilcoxon rank-sum test after HB correction) are denoted as ** (*p-value* $< 0.01$).

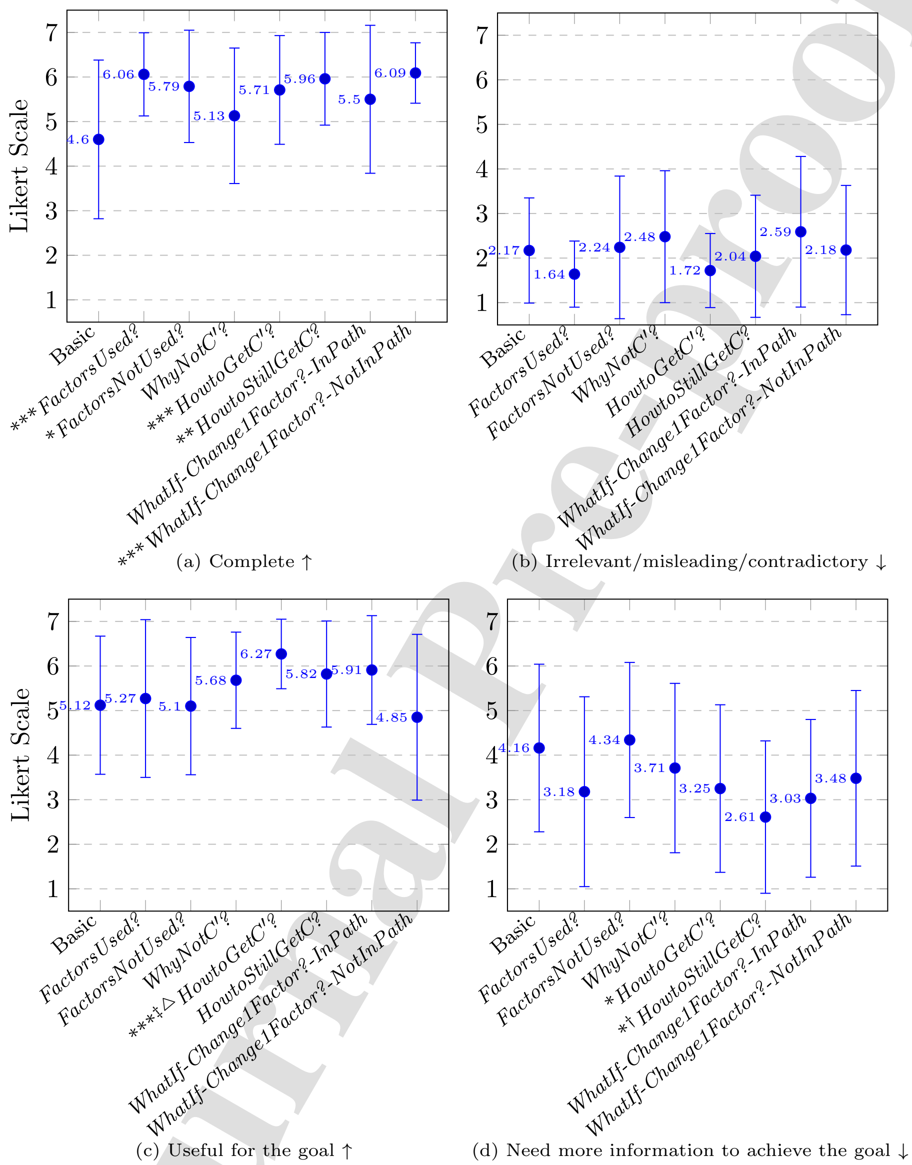



Figure 5: Comparison between explanation types for *changing the predicted outcome* (sample sizes in Table E.24): mean and standard deviation of ratings for the four explanatory attributes; ↑ / ↓ indicates that a higher / lower score is better for an attribute. Statistically significant differences between our explanation types and the Basic explanation (Wilcoxon rank-sum test after HB correction) are denoted as ***, **, * ($p$-*value* $< 0.001, 0.01, 0.05$ respectively), between an explanation type and *FactorsNotUsed?* are denoted as ‡, † ($p$-*value* $< 0.01, 0.05$ respectively), and between an explanation type and *WhatIf-Change1Factor?-NotInPath* are denoted as △ ($p$-*value* $< 0.01$).

between addressing the selected question and users' ratings pertaining to irrelevant/misleading/contradictory information ($\rho = -0.55$) and needing additional information to achieve the goal ($\rho = -0.43$) – all *p-values* $\ll 0.01$.

***Influence of* FQ*-selection round.*** We did not find a significant difference in the ratings given to the explanations for the selected FQs in any of the three follow-up rounds in terms of completeness, irrelevant/misleading/contradictory information and usefulness for the goal. In addition, there were no significant differences between the first and second round of FQs in terms of needing more information to achieve the goal. However, users' need for additional information after the third round of FQs was lower than after the first and second rounds (Wilcoxon signed-rank test, *p-value* $< 0.001, 0.05$ respectively).[22] These results indicate that users need more information than that provided in Basic explanations in order to achieve their goals, and that these requirement is largely satisfied with three additional explanations.

***Influence of explanation length.*** All the explanations for *FactorsUsed?* and *FactorsNotUsed?* were categorized as 'short', comprising 53% of the explanations in the 'short' category; and all the explanations for *HowtoStillGetC?* were categorized as 'long', comprising 40% of the explanations for the 'long' category. The explanations for the other FQs were distributed across the three length categories, and the Basic explanations were 'short' or 'medium'.

Overall, the ratings for explanations of different length categories differed significantly only with respect to needing more information to achieve the goal (Kruskal-Wallis test, *p-value* $< 0.05$), and only when there was a large difference in length category, i.e., small versus large (Wilcoxon rank-sum test, *p-value* $< 0.05$). That is, users thought that short explanations do not contain sufficient detail to help them achieve their goals — a finding that is consistent with Lombrozo's (2016) regarding users' preference for longer explanations.

**Finding 7.** *Users' ratings of* whether an explanation addresses a selected FQ *are positively correlated with their ratings for the explanation in terms of completeness and usefulness for the goal, and negatively correlated with their ratings for the other two attributes. The* FQ-selection round *and* explanation length *had an impact only on users' need for additional information to achieve a goal.*

## 6. Discussion and Conclusions

In this work, we have offered methodological and empirical contributions about the influence of two types of contextual information, viz background information available to users and users' goals, on users' views regarding textual explanations for DT predictions.

[22] We compared the 'need more information' third-round ratings of users who asked only three FQs to the ratings of users who asked four FQs, in order to investigate whether this finding is an artifact of our experimental setting. A Wilcoxon rank-sum test revealed that the ratings of both groups are equivalent, suggesting that this is not the case.

*Methodological Contributions.*

- *Influence of background information* – we generated contrastive explanations that address four types of potential conflicts between aspects of DT predictions and plausible expectations licensed by background information. To this effect, we operationalized the identification of these conflicts, and specified schemas for generating explanations that address them.
- *Influence of users' goals* – given an initial Basic explanation for a DT's prediction, we identified six types of follow-up questions, and generated explanations for each type of question. Here, we employed an interactive setting where users selected follow-up questions that helped them achieve a given goal.

  This interactive system is a step towards an explanatory dialogue system as envisioned by Lakkaraju et al. (2022), where users have the opportunity to engage with the system and ask follow-up questions that help them achieve their goals. Our system also addresses the following research challenges listed in (Verma et al., 2020): (1) transfactual explanations should be presented as discrete and sequential steps that inform users how to modify their current state; and (2) transfactual explanations should also inform users about what must not change.

We have focused on a particular transparent model (DT), as we believed that it would be a good starting point to explore the influence of two types of contextual information. However, the key ideas underpinning our algorithm for generating Conflict-based explanations are model agnostic, except for the determination of the actual impact of a feature value, which is readily available in most ML models, e.g., in linear and logistic regression, this information resides in the coefficients of the variables. The follow-up questions identified in Section 3.2.1 are also generic, and the explanations that answer these questions hinge on the identification of relevant features and feature values. For example, in order to answer question *HowtoGet$\mathcal{C}'$?*, we must identify combinations of features and values that lead to an alternative outcome, and to answer question *HowtoStillGet$\mathcal{C}$?*, we must identify combinations that lead to the predicted outcome. Singh et al. (2021) answer question *HowtoGet$\mathcal{C}'$?* by generating a *Partial Dependence Plot* for each feature, which shows the value at which a logistic regression model changes its decision, assuming that the values of other features remain the same. However, they do not look at combinations of feature values. The enumeration of all the combinations is model agnostic, but it is also exponential. An interesting avenue for future research involves pursuing promising combinations of feature values.

*Key findings.* The key findings obtained from our user studies are as follows.

- *Experiment I – Influence of background information* – we found that Conflict-based explanations are generally considered at least as good as the Basic baseline explanations in terms of completeness, enabling users' to understand the AI's reasoning, and enticing users to act on a DT's predictions; and that Conflict-based explanations are deemed especially valuable when users' expectations disagree with DT predictions. These insights are of practical import, since users' expectations are often not available to explanation systems, and Conflict-based explanations provide clear benefits, or at worst are neutral, regardless of the particulars of these expectations.

- *Experiment II – Influence of users' goals* – we found that the follow-up questions selected for the three goals in our study (*understand the AI's reasoning*, *change the predicted outcome* and *retain the predicted outcome*) have some overlap, and that *HowtoGet$\mathcal{C}'$?* is the most selected question for all the goals. The explanation that addresses *HowtoGet$\mathcal{C}'$?* is highly rated in terms of usefulness for the goal of *changing the predicted outcome*, and also well regarded in terms of the other explanatory attributes for this goal.

In summary, the results of our experiments indicate that explanations that have a contrastive aspect about the predicted class are generally preferred by users. This lends support to the argument in (Wachter et al., 2018) that contrastive explanations provide sufficient explanatory power for users to understand the predictions of an ML model, without understanding how the entire model works. Comparing between the explanations for the two class-contrastive questions in Experiment II, *Howto-Get$\mathcal{C}'$?*, which also has a transfactual aspect, was preferred to *WhyNot$\mathcal{C}'$?*. This finding aligns with long-standing research in philosophy, psychology and the social sciences which demonstrates that transfactuals (or counterfactuals) help users draw inferences about the relation between antecedent events (feature values) and outcomes (Byrne, 2007, 2019).

*Limitations and Future Work.* The main limitations of our approach are as follows.

- Our datasets have relatively few features, which reduces the need to address conflicts due to several features — a problem that must be considered in more complex domains. In addition, our DTs are quite concise, which minimizes the need to perform feature selection to shorten long DT paths. In fact, this problem arose only for explanations generated for *HowtoGet$\mathcal{C}'$?* and *HowtoStillGet$\mathcal{C}$?* in Experiment II, and it was alleviated by designating two attributes that should not be altered: *age* and *gender*. Recently, Hu et al. (2019) and Lin et al. (2020) proposed algorithms that generate succinct DTs, which mitigates the long-path problem. A potential avenue of future research could be to perform feature selection algorithmically (on succinct or full DTs), so as to mention only the features with high impact on a prediction, combined with the cost or practicality of a feature-value change for a particular user.

- Our explanations omit information about DT accuracy for particular instances. In the future, it is worth investigating the impact of including this information.

Our user studies have the following limitations.

- In Experiment II, each goal was associated with a different patient's profile. This poses a risk whereby the features of a profile could influence our findings (Experiment I had more scenarios and also two domains, thus reducing this risk). In the future, we plan to address this issue by swapping the goals associated with the profiles and including additional profiles.

- We could not recruit real users who would be personally engaged with the scenarios in the experiments. This is a general problem in evaluating NLG systems, which we tried to mitigate by having a narrative immersion at the start of our experiments.

Finally, the following results of our experiments warrant further investigation.

- The results of Experiment I reveal a discrepancy between users' ratings of explanatory attributes and their overall preferences, and also show some disagreement between users' views of explanations that consider a surprising impact of a variable without a surprising outcome ($\mathit{Plausible}\mathcal{C}/\mathit{Predict}\mathcal{C}$-$x_{i,j}\,\mathit{NoImpact}$) and the views reported in (Biran and McKeown, 2017).

- The results of Experiment II provide general guidelines for explanation types (FQs) to be included in explanations generated for particular goals. However, these results do not take into account whether the explanations that address these FQs helped users achieve their goals. To remedy this shortcoming, we propose to combine FQ selection-order for a goal with the ratings given to the associated explanations and the final ratings for whether the explanations helped users achieve their goal.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research was supported in part by grant DP190100006 from the Australian Research Council. Ethics approval for the user studies was obtained from Monash University Human Research Ethics Committee (ID-24208). We thank Marko Bohanec, one of the creators of the Nursery dataset, for helping us understand the features and their values. We are also grateful to the anonymous reviewers for their helpful comments.

## References


Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI2018), Montreal, Canada. pp. 1–18.

Bastani, O., Kim, C., Bastani, H., 2017. Interpreting blackbox models via model extraction. arXiv:1705.08504.

Biran, O., McKeown, K., 2017. Human-centric justification of Machine Learning predictions, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI2017), Melbourne, Australia. pp. 1461–1467.

Byrne, R.M.J., 2007. Précis of the rational imagination: How people create alternatives to reality. Behavioral and Brain Sciences 30, 439–453.

Byrne, R.M.J., 2019. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI2019), Macao, China. pp. 6276–6282.

Cairns, E., Cammock, T., 1978. Development of a more reliable version of the Matching Familiar Figures test. Developmental Psychology 14, 555–560.

Cawsey, A., 1993. Planning interactive explanations. International Journal of Man-Machine Studies 38, 169–199.

Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H., 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders, in: Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI2019), Glasgow, Scotland. pp. 1–12.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable Machine Learning. arXiv:1702.08608.

Elsaesser, C., Henrion, M., 1989. Verbal expressions for probability updates: How much more probable is "much more probable"?, in: Henrion, M., Shachter, R.D., Kanal, L.N., Lemmer, J.F. (Eds.), Uncertainty in Artificial Intelligence (UAI'89), Windsor, Canada. pp. 319–330.

Felzmann, H., Villaronga, E.F., Lutz, C., Tamò-Larrieux, A., 2019. Transparency you can trust: Transparency requirements for Artificial Intelligence between legal norms and contextual concerns. Big Data & Society 6, 1–14.

Frank, E., Hall, M.A., Witten, I.H., 2016. The WEKA Workbench – Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". 4 ed., Morgan Kaufmann Publishers, San Francisco, California.

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

Hoffman, R.R., Klein, G., 2017. Explaining explanation, Part 1: Theoretical foundations. IEEE Intelligent Systems 32, 68–73.

Hoffman, R.R., Mueller, S.T., Klein, G., 2017. Explaining explanation, Part 2: Empirical foundations. IEEE Intelligent Systems 32, 78–86.

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

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

Horacek, H., 1997. A model for adapting explanations to the user's likely inferences. User Modeling and User-Adapted Interaction 7, 1–55.

Howcroft, D.M., Belz, A., Clinciu, M.A., Gkatzia, D., Hasan, S.A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., Rieser, V., 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 (INLG2020), Dublin, Ireland. pp. 169–182.

Hu, X., Rudin, C., Seltzer, M., 2019. Optimal sparse decision trees, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (Eds.), Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Itti, L., Baldi, P., 2009. Bayesian surprise attracts human attention. Vision Research 49, 1295–1306.

Knuiman, M.W., Vu, H.T., Bartholomew, H., 1998. Multivariate risk estimation for coronary heart disease: the Busselton health study. Australian & New Zealand Journal of Public Health 22, 747–753.

Korb, K.B., McConachy, R., Zukerman, I., 1997. A cognitive model of argumentation, in: Proceedings of the 19th Annual Conference of the Cognitive Science Society (CogSci 1997), Stanford, California. pp. 400–405.

Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J., 2017. Interpretable and explorable approximations of black box models, in: SIGKDD2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Canada.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., Singh, S., 2022. Rethinking explainability as a dialogue: A practitioner's perspective. `arXiv:2202.01875`.

Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: Informing design practices for explainable AI user experiences, in: Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI2020), Honolulu, Hawaii. pp. 1–15.

Liao, Q.V., Varshney, K.R., 2022. Human-centered explainable AI (XAI): From algorithms to user experiences. `arXiv:2110.10790`.

Lin, J., Zhong, C., Hu, D., Rudin, C., Seltzer, M., 2020. Generalized and scalable optimal sparse decision trees, in: Daumé III, H., Singh, A. (Eds.), Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vienna, Austria. pp. 6150–6160.

Lin, S., 2010. Rank aggregation methods. WIREs Computational Statistics 2, 555–570.

Lipton, P., 1990. Contrastive explanation. Royal Institute of Philosophy Supplement 27, 247–266.

Litman, L., Robinson, J., Abberbock, T., 2017. Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods 49, 433–442.

Lombrozo, T., 2016. Explanatory preferences shape learning and inference. Trends in Cognitive Sciences 20, 748–759.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS2017), Long Beach, California. pp. 4768–4777.

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

Moore, J.D., Paris, C.L., 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. Computational Linguistics 19, 651–694.

Nunes, I., Jannach, D., 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction 27, 393–444.

Olave, M., Rajkovic, V., Bohanec, M., 1989. An application for admission in public school systems, in: Snellen, I., van de Donk, W., Baquiast, J.P. (Eds.), Expert Systems in Public Administration. Elsevier. chapter 10, pp. 145–160.

Ortony, A., Partridge, D., 1987. Surprisingness and expectation failure: What's the difference?, in: Proceedings of the 10th International Joint Conference on Artificial Intelligence (IJCAI87), Milan, Italy. pp. 106–108.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P., 2020. FACE: Feasible and actionable counterfactual explanations, in: Proceedings of the 3rd AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES'20), New York, New York. pp. 344–350.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, California.

Reiter, E., 2019. Natural language generation challenges for explainable AI, in: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI'2019), Tokyo, Japan. pp. 3–7.

Ribeiro, M.T., Singh, S., Guestrin, C., 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), San Francisco, California. pp. 1135–1144.

Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations, in: McIlraith, S.A., Weinberger, K.Q. (Eds.), Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, Louisiana. pp. 1527–1535.

Singh, R., Dourish, P., Howe, P., Miller, T., Sonenberg, L., Velloso, E., Vetere, F., 2021. Directive explanations for actionable explainability in machine learning applications. `arXiv:2021.02671`.

Sokol, K., Flach, P., 2018. Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI2018), Stockholm, Sweden. pp. 5868–5870.

Sokol, K., Flach, P., 2020a. LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. `arXiv:2005.01427`.

Sokol, K., Flach, P., 2020b. One explanation does not fit all: The promise of interactive explanations for Machine Learning transparency. `arXiv:2001.09734`.

Stepin, I., Alonso, J.M., Catala, A., Pereira, M., 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), Glasgow, Scotland. pp. 1–8.

Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M., 2021. A survey of contrastive and counterfactual explanation generation methods for explainable Artificial Intelligence. IEEE Access 9, 11974–12001.

Stone, M., 2000. Towards a computational account of knowledge, action and inference in instructions. Journal of Language and Computation 1, 231–246.

Tintarev, N., Masthoff, J., 2012. Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. User Modeling and User-Adapted Interaction 22, 399–439.

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

Verma, S., Dickerson, J.P., Hines, K., 2020. Counterfactual explanations for Machine Learning: A review. `arXiv:2010.10596`.

van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., Neerincx, M., 2018. Contrastive explanations with local Foil Trees, in: Proceedings of the ICML-18 Workshop on Human Interpretability in Machine Learning (WHI'18), Stockholm, Sweden. pp. 41–46.

Wachter, S., Mittelstadt, B., Russell, C., 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology 31, 842–887.

Webber, W., Moffat, A., Zobel, J., 2010. A similarity measure for indefinite rankings. ACM Transactions on Information Systems 28, 1–38.

Weld, D.S., Bansal, G., 2019. The challenge of crafting intelligible intelligence. Communications of the ACM 62, 70–79.

Wintle, B.C., Fraser, H., Wills, B.C., Nicholson, A.E., Fidler, F., 2019. Verbal probabilities: Very likely to be somewhat more confusing than numbers. PLoS ONE 14, e0213522.

Zukerman, I., McConachy, R., 1993. Generating concise discourse that addresses a user's inferences, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI93), Chambery, France. pp. 1202–1207.

Zukerman, I., McConachy, R., Korb, K., Pickett, D., 1999. Exploratory interaction with a Bayesian argumentation system, in: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI99), Stockholm, Sweden. pp. 1294–1299.

# Appendix A. Templates for the explanations generated in Experiment I, sample explanations generated for the Telecom dataset, and Templates for the explanations generated in Experiment II

| Schema | Template |
|---|---|
| **Basic (no conflict)** | |
| $DT\text{-}Path + \mathcal{C}$ | The AI system has learned from the data that [*dataset-members* with *DT-Path*] are [verbal-percent-leaf-prediction] to get [$\mathcal{C}$] ([percent-leaf-prediction]%). |
| **Conflict-based (outcome only): *Plausible*$\neg\mathcal{C}$*/Predict*$\mathcal{C}$** | |
| *Preamble*: $x^*_{i,j}$ + $\underline{\text{R1}}$ + $\mathcal{C}$ | From the data, one might expect that [*dataset-members* with $x^*_{i,j}$] will be [$\underline{\text{R1}}$] than [*dataset-members*] overall to get [$\mathcal{C}$] ([$Posterior(\mathcal{C}\|x^*_{i,j})$]% vs [$Prior(\mathcal{C})$]%). |
| *Resolution*: $\{DT\text{-}Path/x^*_{i,j}\} + x^*_{i,j} + \mathcal{C}$ | However, the AI system has learned from the data that among [*dataset-members* with $\{DT\text{-}Path/x^*_{i,j}\}$], those with [$x^*_{i,j}$] are [verbal-percent-leaf-prediction] to get [$\mathcal{C}$] ([percent-leaf-prediction]%). |
| **Conflict-based (impact of feature value only): *Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImp*** | |
| *Preamble*: $x^*_{i,j}$ + $\underline{\text{R1}}$ + $\mathcal{C}$ | From the data, one might expect that [*dataset-members* with $x^*_{i,j}$] will be [$\underline{\text{R1}}$] than [*dataset-members*] overall to get [$\mathcal{C}$] ([$Posterior(\mathcal{C}\|x^*_{i,j})$]% vs [$Prior(\mathcal{C})$]%). |
| *Resolution*: $x^*_i$ + R5 + $DT\text{-}Path + \mathcal{C}$ | However, the AI system has learned from the data that [$x^*_i$] has no effect on the outcome in this situation, and that [*dataset-members* with *DT-Path*] are [verbal-percent-leaf-prediction] to get [$\mathcal{C}$] ([percent-leaf-prediction]%). |
| **Conflict-based (outcome & impact of feature value): *Plausible*$\mathcal{C}_{max}$*/Predict*$\mathcal{C}$-$x_{i,j}$ *NoImp*** | |
| *Preamble*: $x^*_{i,j}$ + $\underline{\text{R3}}$ + $\mathcal{C}_{max} + \mathcal{C}$ | From the data, one might expect that [*dataset-members* with $x^*_{i,j}$] will be [$\underline{\text{R3}}$] to get [$\mathcal{C}_{max}$] than to get [$\mathcal{C}$] ([$Posterior(\mathcal{C}_{max}\|x^*_{i,j})$]% vs [$Posterior(C\|x^*_{i,j})$]%). |
| *Resolution*: $x^*_i$ + R5 + $DT\text{-}Path + \mathcal{C}$ | However, the AI system has learned from the data that [$x^*_i$] has no effect on the outcome in this situation, and that [*dataset-members* with *DT-Path*] are [verbal-percent-leaf-prediction] to get [$\mathcal{C}$] ([percent-leaf-prediction]%). |

Table A.16: Templates for the Basic schema (our baseline) and for schemas used in Experiment I to address three of the potential conflicts defined in Table 3 (*NoImp* is shorthand for *No Impact*); [*X*] indicates that *X* is being evaluated, *dataset-members* denotes nouns that refer to members of the dataset, and *DT-Path* denotes the features and values in the current path in the DT; probabilities are stated as percentages, and the presentation of probabilities in brackets is in line with the findings in (Wintle et al., 2019); the selection of a *pivot feature* value is described in Section 3.1.2.

| |
|---|
| **Basic (no conflict): counterpart of *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$-$x_{i,j}$ NoImp*** |
| The AI system has learned from the data that customers who have **no *online security*, *tenure with the company* less than 15 months** and ***monthly charges* greater than $69** are almost certain to *Churn* (close to 100%). |
| **Conflict-based (outcome only): *Plausible¬$\mathcal{C}$/Predict$\mathcal{C}$*** |
| From the data, one might expect that customers who **have *online security*** will be less likely to *Churn* than customers overall (21% vs 40%). |
| However, the AI system has learned from the data that among customers who have ***tenure with the company* less than 15 months** and ***monthly charges* greater than $81**, those **having *online security*** are almost certain to *Churn* (close to 100%). |
| **Conflict-based (impact of feature value only): *Plausible$\mathcal{C}$/Predict$\mathcal{C}$-$x_{i,j}$ NoImp*** |
| From the data, one might expect that customers who have **no *internet service*** will be more likely to *Stay with the company* than customers overall (89% vs 60%). |
| However, the AI system has learned from the data that ***internet service*** has no effect on the outcome in this situation, and that customers who have ***tenure with the company* greater than 5 months** and ***monthly charges* less than $69** are almost certain to *Stay with the company* (close to 100%). |
| **Conflict-based (outcome only): *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$* "vanilla"** |
| From the data, one might expect that customers who **have *monthly charges* greater than $69** will be more likely to *Churn* than to *Stay with the company* (54% vs 42%). |
| However, the AI system has learned from the data that among customers who have **Fiber optic *internet service*** and ***tenure with the company* greater than 53 months**, those **having *monthly charges* greater than $69** are almost certain to *Stay with the company* (close to 100%). |
| **Conflict-based (outcome & impact of feature value): *Plausible$\mathcal{C}_{max}$/Predict$\mathcal{C}$-$x_{i,j}$ NoImp*** |
| From the data, one might expect that customers who **have *tech support*** will be a great deal more likely to *Stay with the company* than to *Churn* (77% vs 23%). |
| However, the AI system has learned from the data that ***tech support*** has no effect on the outcome in this situation, and that customers who have **no *online security*, *tenure with the company* less than 15 months** and ***monthly charges* greater than $69** are almost certain to *Churn* (close to 100%). |

Table A.17: Sample explanations generated for the Telecom dataset (*NoImp* is shorthand for *No Impact*); multiple splits on the same numeric feature in a DT path (*tenure* and *monthly charges*) are merged; the presentation of probabilities in brackets is in line with the findings in (Wintle et al., 2019); the selection of a ***pivot feature* value** is described in Section 3.1.2; font denotes ***features***, **feature values** and *Classes*.

| |
|---|
| ***FactorsUsed?*:** Which factors in the data are used by the AI system to predict [task-definition]? |
| In general, the following factors are used by the AI system to predict [task-definition]: [$\{x_i \in DT\}$]. |
| ***FactorsNotUsed?*:** Which factors in the data are not used by the AI system to predict [task-definition]? |
| The following factors do not improve the accuracy of the AI's predictions, and hence are not used by the AI system: [$\{x_i\} - \{x_i \in DT\}$] . |

Table A.18: Templates for explanations generated for the general questions in Experiment II; [$X$] indicates that $X$ is being evaluated, $\{x_i\}$ denotes the set of features in the dataset, and $\{x_i \in DT\}$ denotes the set of features in the DT.

| |
|---|
| **Basic explanation:** |
| This prediction was made because the AI system has learned from the data that [*dataset-members* with $DT$-*Path*] are [$\mathcal{C}$]. |
| ***WhyNot$\mathcal{C}'$?*:** Why wasn't I given a specific different prediction? |
| The AI system has learned from the data that about [$\Pr(\mathcal{C}' \vert \{DT\text{-}Path/\{x_{\text{max-drop}}\}\})$]% of [*dataset-members* with $\{DT\text{-}Path/\{x_{\text{max-drop}}\}\}$] are [$\mathcal{C}'$]. However, because you have [$\{x_{\text{max-drop}}\}$], the AI system predicts that you are not [$\mathcal{C}'$]. |
| ***HowtoGet$\mathcal{C}'$?*:** Which factor changes will result in a specific different prediction ([$\mathcal{C}'$]) for me? |
| If nothing else changes in your circumstances, the following would result in a different prediction ([$\mathcal{C}'$]) for you:<br>• [list of $x_i$s and the change(s) in their values that result in $\mathcal{C}'$]. |
| ***HowtoStillGet$\mathcal{C}$?*:** Which factor changes will result in the same prediction ([$\mathcal{C}$]) for me? |
| If nothing else changes in your circumstances, the following would result in the same prediction ([$\mathcal{C}$])for you:<br>• any changes in one of these factors: [$\{x_i \notin DT\text{-}Path\}$]; or<br>• [list of $x_i$s and the change(s) in their values that result in $\mathcal{C}$].<br>Also,<br>• [change(s) in the values of $(x_i, x_k)$ that result in $\mathcal{C}$, where $x_i \in DT\text{-}Path$, $x_k \notin DT\text{-}Path$ and $x_k \in DT\text{-}Path'$ taken when the value of $x_i$ changes]. |
| ***WhatIf-Change1Factor?*:**<br>What would be the prediction if one of the factors were to change for me? |
| User selects $x_i \in DT\text{-}Path$ |
| If [$x_i$ with change(s) in its value that result in $\mathcal{C}$], it would result in the same prediction for you ([$\mathcal{C}$]), provided nothing else changes in your circumstances.<br>However, if [$x_i$ with change(s) in its value that result in $\mathcal{C}'$], it would result in a different prediction for you ([$\mathcal{C}'$]), provided nothing else changes in your circumstances. |
| User selects $x_j$ s.t. $x_j \in DT$ & $x_j \notin DT\text{-}Path$ |
| If [all changes in the value of $x_j$], it would result in the same prediction for you ([$\mathcal{C}$]), because [$x_j$] has no effect on the prediction in light of $\{x_i \in DT\text{-}Path\}$. |
| User selects $x_j \notin DT$ |
| If [all changes in the value of $x_j$], it would result in the same prediction for you ([$\mathcal{C}$]), because the AI system did not use [$x_j$] to make predictions. |

Table A.19: Templates for Basic explanations and for explanations generated for the profile-specific questions in Experiment II; [$X$] indicates that $X$ is being evaluated, $\{x_i\}$ denotes the set of features in the dataset, $\{x_i \in DT\}$ denotes the set of features in the DT, $DT$-*Path* denotes the current path in the DT, $\{x_i \in DT\text{-}Path\}$ denotes the set of features in $DT$-*Path*, $x_{\text{max-drop}}$ denotes the feature value with the greatest drop in the probability of $\mathcal{C}'$, and $\{x_{\text{max-drop}}\}$ denotes the feature values from $x_{\text{max-drop}}$ onward in $DT$-*Path*.

## Appendix B. Decision Trees learned for the Nursery, Telecom and Busselton datasets

```
health = good
|   current childcare = good: wait list
|   current childcare = sufficient: wait list
|   current childcare = insufficient
|   |   parents’ employment = ordinary: wait list
|   |   parents’ employment = somewhat difficult
|   |   |   social situation = unproblematic: wait list
|   |   |   social situation = somewhat problematic: wait list
|   |   |   social situation = problematic: priority accept
|   |   parents’ employment = challenging: priority accept
|   current childcare = critical
|   |   parents’ employment = ordinary
|   |   |   social situation = unproblematic: wait list
|   |   |   social situation = somewhat problematic: wait list
|   |   |   social situation = problematic: priority accept
|   |   parents’ employment = somewhat difficult: priority accept
|   |   parents’ employment = challenging: priority accept
|   current childcare = very critical: priority accept
health = average
|   current childcare = good
|   |   parents’ employment = ordinary: wait list
|   |   parents’ employment = somewhat difficult: wait list
|   |   parents’ employment = challenging: priority accept
|   current childcare = sufficient
|   |   parents’ employment = ordinary: wait list
|   |   parents’ employment = somewhat difficult: wait list
|   |   parents’ employment = challenging
|   |   |   housing condition = adequate: wait list
|   |   |   housing condition = somewhat inadequate: priority accept
|   |   |   housing condition = inadequate: priority accept
|   current childcare = insufficient
|   |   parents’ employment = ordinary: wait list
|   |   parents’ employment = somewhat difficult
|   |   |   housing condition = adequate: wait list
|   |   |   housing condition = somewhat inadequate: priority accept
|   |   |   housing condition = inadequate: priority accept
|   |   parents’ employment = challenging: priority accept
|   current childcare = critical
|   |   parents’ employment = ordinary
|   |   |   housing condition = adequate: wait list
|   |   |   housing condition = somewhat inadequate: priority accept
|   |   |   housing condition = inadequate: priority accept
|   |   parents’ employment = somewhat difficult: priority accept
|   |   parents’ employment = challenging: priority accept
|   current childcare = very critical: priority accept
health = poor: reject

Number of Leaves  :  33
Size of the tree  :  47
```

Figure B.6: DT for the Nursery dataset with recoded classes and features.

```
monthly charges <= 69.05
|   tenure <= 5
|   |   senior citizen = no
|   |   |   internet service = DSL
|   |   |   |   paper billing = no
|   |   |   |   |   phone service = no: churn
|   |   |   |   |   phone service = yes
|   |   |   |   |   |   gender = female: churn
|   |   |   |   |   |   gender = male: stay
|   |   |   |   paper billing = yes: stay
|   |   |   internet service = Fiber optic: churn
|   |   |   internet service = no: stay
|   |   senior citizen = yes: churn
|   tenure > 5: stay
monthly charges > 69.05
|   tenure <= 14
|   |   online security = no: churn
|   |   online security = yes
|   |   |   monthly charges <= 81.3: stay
|   |   |   monthly charges > 81.3: churn
|   |   online security = NA (no internet service): churn
|   tenure > 14
|   |   internet service = DSL: stay
|   |   internet service = Fiber optic
|   |   |   tenure <= 53
|   |   |   |   multiple phone lines = NA (no phone service): stay
|   |   |   |   multiple phone lines = no: stay
|   |   |   |   multiple phone lines = yes
|   |   |   |   |   tech support = no
|   |   |   |   |   |   paper billing = no
|   |   |   |   |   |   |   movie streaming = no
|   |   |   |   |   |   |   |   senior citizen = no: stay
|   |   |   |   |   |   |   |   senior citizen = yes: churn
|   |   |   |   |   |   |   movie streaming = yes: churn
|   |   |   |   |   |   |   movie streaming = NA (no internet service): churn
|   |   |   |   |   |   paper billing = yes: stay
|   |   |   |   |   tech support = yes: stay
|   |   |   |   |   tech support = NA (no internet service): stay
|   |   |   tenure > 53: stay
|   |   internet service = no: stay

Number of Leaves  :  24
Size of the tree  :  41
```

Figure B.7: DT for the Telecom dataset with recoded features.

```
age <= 60.5
|   age <= 42.6
|   |   smoke_amt <= 25: low risk
|   |   smoke_amt > 25
|   |   |   age <= 35.6: low risk
|   |   |   age > 35.6: high risk
|   age > 42.6
|   |   HDL-chol-cat = optimal
|   |   |   smoke_amt <= 28: low risk
|   |   |   smoke_amt > 28: high risk
|   |   HDL-chol-cat = borderline
|   |   |   gender = female: low risk
|   |   |   gender = male
|   |   |   |   Chol-cat = low: low risk
|   |   |   |   Chol-cat = normal: low risk
|   |   |   |   Chol-cat = borderline: low risk
|   |   |   |   Chol-cat = high: high risk
|   |   HDL-chol-cat = low: high risk
age > 60.5
|   age <= 69.1
|   |   gender = female
|   |   |   age <= 63.4: low risk
|   |   |   age > 63.4
|   |   |   |   weight-cat = underweight: high risk
|   |   |   |   weight-cat = optimal
|   |   |   |   |   smoke_amt <= 5
|   |   |   |   |   |   alc_amt <= 7: low risk
|   |   |   |   |   |   alc_amt > 7: high risk
|   |   |   |   |   smoke_amt > 5: high risk
|   |   |   |   weight-cat = overweight: low risk
|   |   |   |   weight-cat = obese: high risk
|   |   gender = male: high risk
|   age > 69.1: high risk

Number of leaves :  20
Size of the tree :  36
```

Figure B.8: Pruned DT for the Busselton dataset with recoded classes and features.

## Appendix C. Experimental Setup

### *Appendix C.1. Datasets*

***The Nursery dataset*** originally had five classes, three of which account for about 97% of the instances; we therefore removed the other two classes, which resulted in a balanced dataset with 12630 instances. The classes, features and feature values in the dataset were originally in Slovenian, and their English translation in (Olave et al., 1989) was somewhat peculiar. With the help of one of the authors of the original paper, we recoded the features and feature values in the Nursery dataset to those in Table 7, and the names of the retained classes to *Reject* (not recommended for admission), *Wait list* (can be admitted eventually) and *Priority accept* (should be given special priority for admission). The recoded feature values are described in Table C.20.

***The Telecom dataset*** had only two classes, *Stay* and *Churn*, but it was imbalanced towards *Stay* (73%). The DT had an accuracy of 79% when trained with a cost-sensitive setting for imbalanced datasets. This accuracy is comparable to those reported in Kaggle for several predictive models built for the Telecom dataset.

However, in order to avoid biasing participants' class expectations, we decided to even out the class distribution. To this effect, we retained only customers with a month-to-month contract, which had both outcomes, and randomly removed half of the incorrectly predicted cases. This yielded a more balanced dataset (60% *Stay*) and a slightly improved DT accuracy of 80% (trained without the cost-sensitive setting).

***The Busselton dataset*** had only two classes: whether someone will experience a CHD event or not within ten years of the initial data collection. We recoded these classes as *high risk of a coronary event* and *low risk of a coronary event* respectively. Originally, there were 4006 instances, but after removing instances with missing values, we were left with 2970 instances. There were 14 features in total, which we converted to 10 features by applying the following pre-processing steps:

1. Remove the redundant binary features of *smoker* and *drinker*.
2. Calculate Body Mass Index (BMI) from the *height* and *weight* features, and use the corresponding *weight category* as a feature,[23] instead of *height* and *weight*.
3. Create a categorical *blood pressure* feature corresponding to the numeric *systolic blood pressure* and *diastolic blood pressure* features.[24]
4. Convert the numeric features *total cholesterol*, *HDL cholesterol* and *triglycerides* to categorical features.[25]

[23] `https://www.betterhealth.vic.gov.au/tools/body-mass-index-calculator-for-adult`
[24] `https://www.mydr.com.au/blood-pressure-what-is-your-target/`
[25] `https://www.victorchang.edu.au/high-cholesterol`

| Feature value | Description |
|---|---|
| | ***Parents' employment*** |
| challenging | frequent relocations, transfers, long leaves of absence; parents are not employed in the school district and need to travel more than one hour for work. |
| somewhat difficult | hard working conditions that allow for an early retirement (e.g., miners, policemen, soldiers), night work, additional work engagements. |
| ordinary | normal condition. |
| | ***Current childcare*** |
| very critical | there is no possibility of childcare with family, and previous level of childcare was inadequate (child does not live with parents, problematic private care). |
| critical | there is no possibility of childcare with family, and previous level of care was less than adequate (frequent change of care, termination of care, alternate care by parents, occasional care). |
| insufficient | no possibility of childcare with family (both parents or single parent work full-time or are full-time students, no alternative care with relatives), but previous level of care was adequate (with own family, adequate private care, educational care organizations). |
| sufficient | childcare is possible with some relatives (healthy and unemployed grandparents living in the school district, other able-bodied and unemployed members of the household). |
| good | normal condition (childcare is possible in the family – father or mother unemployed and able to care). |
| | ***Housing condition*** |
| inadequate | subleased or emergency housing; cramped; has lack of sanitation facilities or water. |
| somewhat inadequate | subleased or cramped apartment. |
| adequate | normal condition. |
| | ***Social situation*** |
| problematic | inadequate educational ability of parents (gross neglect of education and care, violence); inadequate family relationships (serious conflicts between parents, between grandparents, between parents and grandparents, more severe forms of disturbance of parents or other family members); social and antisocial forms of restraining behavior by parents and other family members (alcoholism and other addictions, delinquency, quitting, etc.). |
| somewhat problematic | less than adequate educational ability of parents (uneven, inconsistent education, excessive difficulty or indulgence, neurotic reaction of parents); less than adequate family relationships (milder forms of parental personality disorders, privileged or neglected children, family conflicts). |
| unproblematic | normal condition. |
| | ***Child's health*** |
| poor | admission is not recommended due to the health conditions of the child. |
| average | the child has a mental or physical disorder that influences their admission status; the child's development is affected by health conditions of family members. |
| good | normal condition (healthy). |

Table C.20: Description of the feature values in the Nursery dataset; all the feature values for *current childcare*, *housing condition*, *social situation* and *child's health*, except the value defined as normal, require the opinion of relevant professional services.

| Partition | Nursery | | | | Telecom | | | Busselton | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Reject* | *Wait list* | *Priority accept* | **Total** | *Stay* | *Churn* | **Total** | *Low risk* | *High risk* | **Total** |
| Training | 3485 | 3414 | 3205 | 10104 | 1596 | 1057 | 2653 | 2082 | 219 | 2301 |
| Testing | 835 | 852 | 839 | 2526 | 390 | 259 | 649 | 519 | 54 | 573 |
| **Total** | 4320 | 4266 | 4044 | 12630 | 1986 | 1316 | 3302 | 2601 | 273 | 2874 |

Table C.21: Breakdown of classes for the training and test sets for the Nursery, Telecom and Busselton datasets.

In addition, we removed the following instances: (1) instances with outliers for the feature *alcohol amount* — to this effect, we used the default settings of the Interquartile range filter in WEKA (Frank et al., 2016); (2) five instances with the *blood pressure* category of 'severe hypertension'; and (3) duplicate instances obtained after converting some numerical features to categorical — this was done so as not to have the same instance in the training and test sets, which may lead to overfitting.

Table C.21 shows the final classes in our evaluation datasets and the breakdown of the training/test sets.

### *Appendix C.2. Experiment I – Influence of background information*

The analysis in this paper uses data from scenarios that compare Conflict-based explanations with Basic explanations. However, our experiment contains additional scenarios, which compare two Conflict-based explanations. We did not analyze these scenarios, because they involve only some of the explanation types.

To limit the duration of an experiment to less than 1 hour, the experiment for each dataset was split into two parts — each part was shown to a different group of participants.

- Each Nursery group was shown five scenarios that compare Conflict-based explanations with Basic explanations, and two scenarios that compare two Conflict-based explanations; two of the former scenarios were common to both Nursery groups.
- Each Telecom group was shown six scenarios that compare Conflict-based explanations with Basic explanations, and one scenario that compares two Conflict-based explanations; as for Nursery, two of the former scenarios were common to both groups.

The common scenarios were used to determine whether the two participant groups for a particular dataset behaved similarly. To this effect, we performed a two-proportion Z-test on preference for Conflict-based explanations in the common scenarios; we found no statistically significant differences between the preferences of the two Nursery groups (*p-value* $= 0.714$) or the preferences of the two Telecom groups (*p-value* $= 0.388$).

## Appendix D. Screenshots from our experiments

### *Appendix D.1. Experiment I (Nursery dataset)*

**Background**
We are developing a computer system that automatically generates explanations for predictions made by an Artificial Intelligence (AI) system. For example, say we have an AI system that predicts whether an applicant to a childcare centre will be accepted or rejected. Our explanation system generates several alternative explanations for this prediction.

The objective of this study is to find out which types of explanations people find useful in order to understand and accept the predictions of the AI system. We would appreciate your help in making this determination.

**About the survey**
In this survey, you will see seven situations together with some background information. We will present the outcome predicted by the AI system for each situation, and show you two alternative explanations for each outcome. You will then rate each explanation based on several criteria, such as clarity and completeness.

This experiment focuses on the childcare domain. We will first introduce you to this domain, and then we will give you an example of the questions you will get in the survey.

---

**The childcare domain**
You are the director of the Bilby Childcare Centre, a non-profit organisation whose aim is to serve all members of the community. Part of your job is to evaluate applications from the parents of prospective pupils. Evaluating these applications involves weighing the childcare needs of families across several factors, such as housing condition and health (see the table below), in order to accept children in most need of childcare. In the past, an admissions committee performed these assessments.

To make the admission process more efficient, you have purchased a state-of-the-art AI system that predicts the outcome of an application from the data considered by the committee and the decisions made by the committee in the past -- the possible outcomes are: **priority accept**, **wait-list** or **reject**. The accuracy of your AI system in predicting the committee's decisions is 93%.

**AI systems** make predictions based on trends and patterns they identify in the data. Therefore, they may determine that attributes that are relevant to some situations are not relevant to other situations. For example, if the family's current childcare arrangements are deemed 'sufficient', their housing condition may influence the AI system's prediction about the outcome of their application. In contrast, the AI system may not need to consider the housing condition, if the current childcare arrangements are deemed 'very critical'.

Each **applicant to the Bilby Childcare Centre** fills out an application form, which is transcribed into **five** factors that make sense to the AI system. The factors and their possible values are listed below in shades of red and blue. These colours will be used in the situations you will see in the survey.

| Factor | Possible values |
|---|---|
| *Parents' employment* | Challenging Somewhat difficult Ordinary |
| *Current childcare* | Very critical Critical Insufficient Sufficient Good |
| *Housing condition* | Inadequate Somewhat inadequate Adequate |
| *Social situation* | Problematic Somewhat problematic Unproblematic |
| *Health (of the child)* | Poor Average Good |

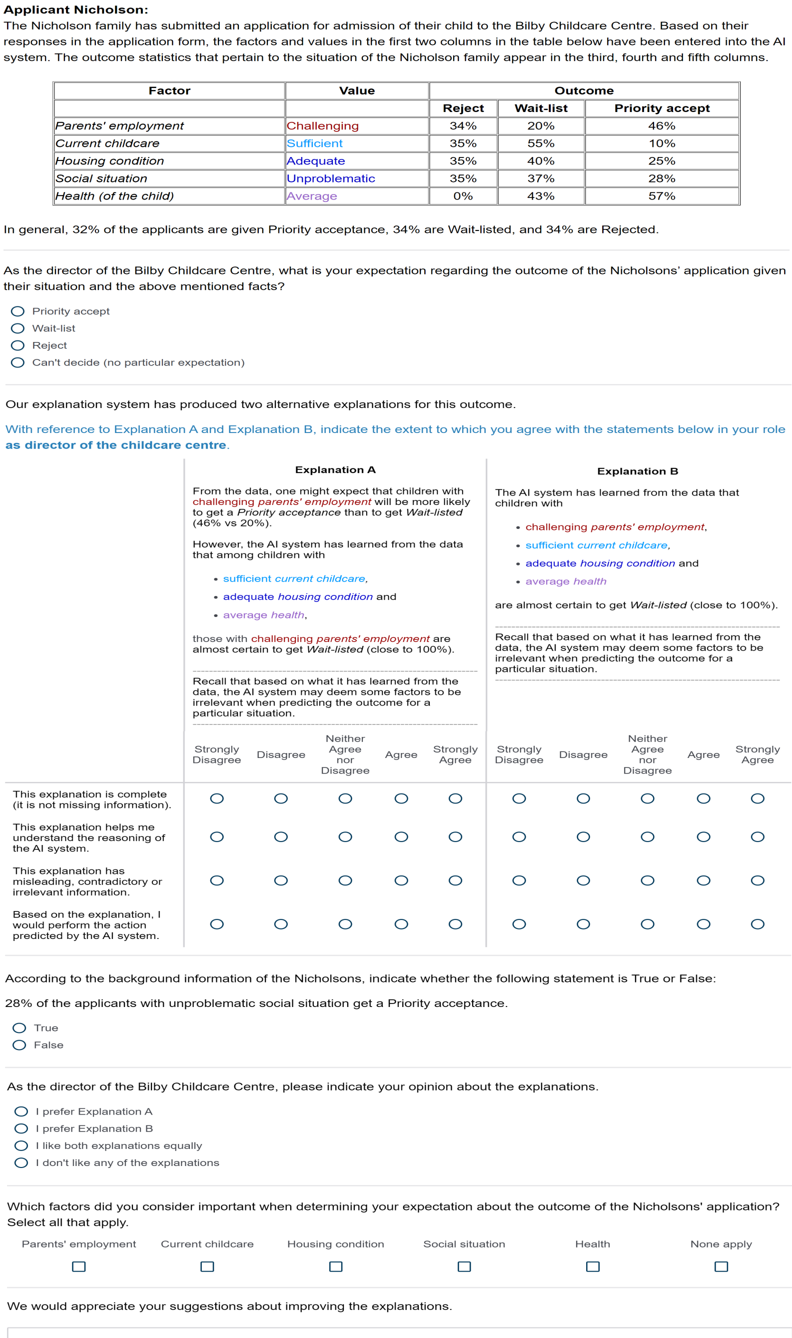
**Note**: since the Bilby Childcare Centre is a community service, it is not equipped to serve children with poor *health*. Therefore, children with poor *health* are rejected, even if their other factors would normally warrant acceptance.

In the following pages, you will see seven applications to the Bilby Childcare Centre. For each application, we will:

- present the above factors and their values, together with a few general facts regarding these factors -- the factors and their values are used by the AI system to make its predictions;
- ask you to make an educated guess about the outcome of the application;
- show you the prediction made by the AI system, together with two alternative explanations for this prediction; and
- ask you to rate these explanations along several criteria, such as clarity and completeness. Your ratings should be informed by your role **as the director of the childcare centre**.

Before we proceed, let's look at a sample application and the questions you will be asked.

Figure D.9: Narrative immersion for the Nursery survey.



**Applicant Nicholson:**
The Nicholson family has submitted an application for admission of their child to the Bilby Childcare Centre. Based on their responses in the application form, the factors and values in the first two columns in the table below have been entered into the AI system. The outcome statistics that pertain to the situation of the Nicholson family appear in the third, fourth and fifth columns.

| Factor | Value | Outcome | | |
|---|---|---|---|---|
| | | Reject | Wait-list | Priority accept |
| *Parents' employment* | Challenging | 34% | 20% | 46% |
| *Current childcare* | Sufficient | 35% | 55% | 10% |
| *Housing condition* | Adequate | 35% | 40% | 25% |
| *Social situation* | Unproblematic | 35% | 37% | 28% |
| *Health (of the child)* | Average | 0% | 43% | 57% |

In general, 32% of the applicants are given Priority acceptance, 34% are Wait-listed, and 34% are Rejected.

As the director of the Bilby Childcare Centre, what is your expectation regarding the outcome of the Nicholsons' application given their situation and the above mentioned facts?

- Priority accept
- Wait-list
- Reject
- Can't decide (no particular expectation)

Our explanation system has produced two alternative explanations for this outcome.

With reference to Explanation A and Explanation B, indicate the extent to which you agree with the statements below in your role **as director of the childcare centre**.

**Explanation A**

From the data, one might expect that children with challenging *parents' employment* will be more likely to get a *Priority acceptance* than to get *Wait-listed* (46% vs 20%).

However, the AI system has learned from the data that among children with

- sufficient *current childcare*,
- adequate *housing condition* and
- average *health*,

those with challenging *parents' employment* are almost certain to get *Wait-listed* (close to 100%).

Recall that based on what it has learned from the data, the AI system may deem some factors to be irrelevant when predicting the outcome for a particular situation.

**Explanation B**

The AI system has learned from the data that children with

- challenging *parents' employment*,
- sufficient *current childcare*,
- adequate *housing condition* and
- average *health*

are almost certain to get *Wait-listed* (close to 100%).

Recall that based on what it has learned from the data, the AI system may deem some factors to be irrelevant when predicting the outcome for a particular situation.

| | A: Strongly Disagree | A: Disagree | A: Neither Agree nor Disagree | A: Agree | A: Strongly Agree | B: Strongly Disagree | B: Disagree | B: Neither Agree nor Disagree | B: Agree | B: Strongly Agree |
|---|---|---|---|---|---|---|---|---|---|---|
| This explanation is complete (it is not missing information). | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| This explanation helps me understand the reasoning of the AI system. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| This explanation has misleading, contradictory or irrelevant information. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| Based on the explanation, I would perform the action predicted by the AI system. | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

According to the background information of the Nicholsons, indicate whether the following statement is True or False:

28% of the applicants with unproblematic social situation get a Priority acceptance.

- True
- False

As the director of the Bilby Childcare Centre, please indicate your opinion about the explanations.

- I prefer Explanation A
- I prefer Explanation B
- I like both explanations equally
- I don't like any of the explanations

Which factors did you consider important when determining your expectation about the outcome of the Nicholsons' application? Select all that apply.

| Parents' employment | Current childcare | Housing condition | Social situation | Health | None apply |
|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

We would appreciate your suggestions about improving the explanations.

Figure D.10: Background information about the Nicholson family; question about the expected outcome; model prediction (displayed after an outcome has been selected); $Plausible\mathcal{C}_{max}/Predict\mathcal{C}$ "vanilla" (A) and Basic (B) explanations for this scenario and rating scales for the explanations; attention question; preferences for explanations; features that determine expectation; request for suggestions.

### Appendix D.2. Experiment II (Busselton dataset)

**The domain**

A health consultancy has purchased a state-of-the-art AI system that predicts whether a particular patient is at a high or low risk of a coronary event. To make these predictions, the AI system takes into account different factors in a patient's profile, such as their age and cholesterol levels (see the table below). The accuracy of this AI system in predicting a patient's risk of a coronary event is 82%.

**AI systems** make predictions based on trends and patterns they identify in the data. Therefore, they may determine that factors that are relevant to some situations are not relevant to other situations. For example, if a person is more than 60 years old, their weight status may influence the AI system's prediction about their risk of a coronary event. In contrast, the AI system may not need to consider the weight status of people under 60 years of age.

The data from which our AI system has learned its patterns was obtained from **ten** personal, lifestyle and medical factors of previous patients. The same factors are obtained from new patients to predict whether they are at a high or low risk of a coronary event. These factors and their possible values are listed below in shades of red (more prone to a coronary event) and blue (less prone to a coronary event). These colours will be used in the situations you will see in the survey.

| **Personal and Lifestyle Factors** | **Possible values** |
|---|---|
| *Age* | 18 95 |
| *Gender* | Female Male |
| *Weight status based on Body Mass Index (BMI)* | Optimal Underweight Overweight Obese |
| *Daily alcohol intake (standard drinks)* | 0 44 |
| *Daily cigarette consumption* | 0 75 |
| **Medical Factors** | **Possible values** |
| *Blood pressure* | Optimal Normal-to-High High |
| *Total cholesterol* | Low Normal Borderline High |
| *HDL cholesterol* | Optimal Borderline Low |
| *Triglycerides* | Low Normal Borderline High |
| *Diabetes* | No Yes |

**Notes:**

- This dataset comes from the 1970s, and at that time people only had the option to choose from two genders.
- If you hover the mouse over the names of medical factors, you will see a brief description for each of them.
- If you hover the mouse over the values of *weight status*, *blood pressure*, *total cholesterol*, *HDL cholesterol* and *triglycerides*, you will see the range for each value.

**Disclaimer:**

The AI system developed for this study is a Machine Learning model that predicts the risk of a coronary event from data pertaining to **a particular population**. Although this system considers relevant medical factors, it may decide to ignore factors that don't improve the system's prediction accuracy **for this population** --- this decision is based on statistical considerations, **not on medical reasons**.

Figure D.11: Narrative immersion for the Busselton survey.

**PatientID 27:**

Assume that you are a *58 year old* *male* who is *overweight*, *does not drink* and *does not smoke*. However, you have *normal-to-high* *blood pressure*, *high* *total cholesterol*, *borderline* *HDL cholesterol* and *borderline* *triglycerides*. But on the upside, you are *not diabetic*.

Notes:

- If you hover the mouse over the *underlined values*, you will see their range.
- Click here to look at the glossary of all factors and their possible values for a patient's profile.

The AI system will predict whether you are at a *high or low risk of a coronary event*.

Before we proceed, please indicate your expectation regarding the outcome based on your profile.

- ○ High risk of coronary event
- ○ Low risk of coronary event
- ○ Can't decide

Based on your profile, our AI system predicts you to be at a **high risk of a coronary event**.

**Please read the following explanation carefully before you rate it. You will be asked about its content later on.**

This prediction was made because the AI system has learned from the data that *men* who

- are *between 43 and 60 years old*,
- have *high* *total cholesterol* and
- have *borderline* *HDL cholesterol*

are at a high risk of a coronary event.

Recall that based on what it has learned from the data, the AI system may deem some factors to be irrelevant when predicting the outcome for a particular patient profile.

For this profile, your objective is to *understand the reasoning behind the AI's prediction*.

Please indicate the extent to which you agree with the following statements about the above explanation:

| | Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree nor Disagree | Somewhat Agree | Agree | Strongly Agree |
|---|---|---|---|---|---|---|---|
| The explanation has irrelevant, misleading or contradictory information. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The explanation is complete (it is not missing information). | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| The explanation is useful for my objective of *understanding the reasoning of the AI*. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |
| In light of the explanation I have received for this patient's profile, I need more information to achieve my objective of *understanding the reasoning of the AI*. | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

Which of the following **factors** were deemed **relevant** by the AI system for **this patient's prediction**? Select all that apply.

| Age | Gender | Weight status | Daily alcohol intake | Daily cigarette consumption | Blood pressure | Total cholesterol | HDL cholesterol | Triglycerides | Diabetes |
|---|---|---|---|---|---|---|---|---|---|
| ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

In the following pages, you must **select at least three questions** (in sequence) that help you achieve the objective of *understanding the reasoning behind the AI's prediction* for your profile.

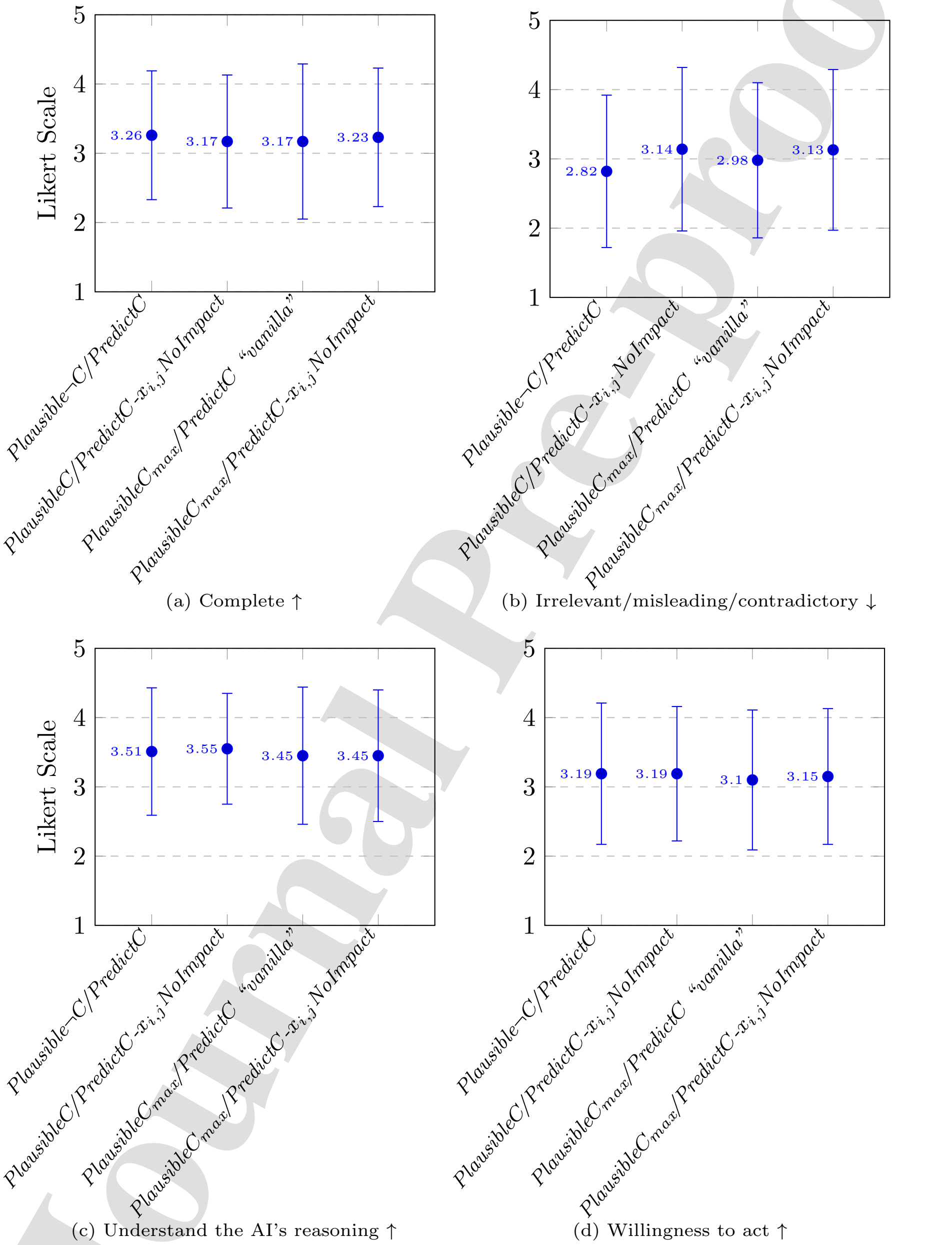
Our explanation system will provide an answer for each question you have selected.

After you have rated the answers for each of the three selected questions, you can ask more questions or proceed to the next patient profile.

Figure D.12: Background information about a patient; question about the expected outcome; model prediction (displayed after an outcome has been selected); Basic explanation for this patient; goal specified for this patient and rating scale for the explanation; attention question.

## Appendix E. Experimental Results

### *Appendix E.1. Experiment I*




(a) Complete ↑

(b) Irrelevant/misleading/contradictory ↓

(c) Understand the AI’s reasoning ↑

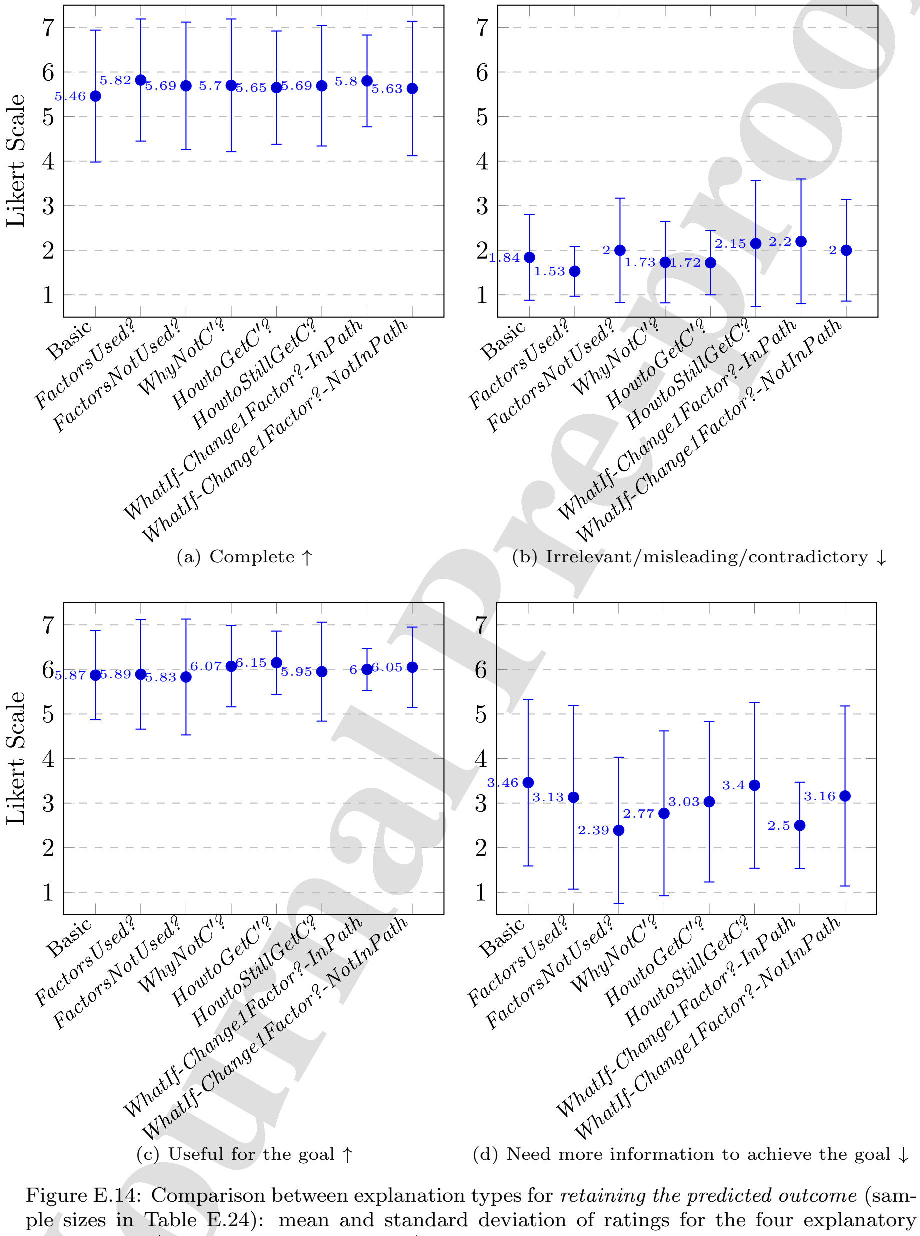
(d) Willingness to act ↑

Figure E.13: Comparison between individual Conflict-based explanations for the Telecom dataset (sample sizes in *Total* column, Table 12): mean and standard deviation of ratings for the four explanatory attributes; ↑ / ↓ indicates that a higher / lower score is better for an attribute.

| | Predict vs Expect | Count | | | | | $\chi^2$ | Stat. Sig. |
|---|---|---|---|---|---|---|---|---|
| | | Conflict-based | Basic | Both | None | Total | | |
| **Nursery** | Pred = Exp | 74 | 35 | 9 | 20 | 138 | 13.95 | $< 0.001$ |
| | Pred $\neq$ Exp | 38 | 10 | 4 | 15 | 67 | 16.33 | $< 0.001$ |
| **Telecom** | Pred = Exp | 78 | 72 | 8 | 34 | 192 | 0.24 | – |
| | Pred $\neq$ Exp | 39 | 6 | 3 | 12 | 60 | 24.20 | $< 0.001$ |

Table E.22: Preferences broken up by (dis)agreement between users' expectations and DT predictions: $\chi^2$ statistic and statistical significances (one-proportion Z-test) calculated from clear preferences for Conflict-based/Basic explanations.

*Appendix E.2. Experiment II*




(a) Complete ↑

(b) Irrelevant/misleading/contradictory ↓

(c) Useful for the goal ↑

(d) Need more information to achieve the goal ↓

Figure E.14: Comparison between explanation types for *retaining the predicted outcome* (sample sizes in Table E.24): mean and standard deviation of ratings for the four explanatory attributes; ↑ / ↓ indicates that a higher / lower score is better for an attribute. No significant differences were found between the explanation types from the Kruskal-Wallis test for any attribute.

| | *WhatIf-Change1Factor?* | | | |
|---|---|---|---|---|
| **Feature** | **In Path** | **Not in Path** | **Not in DT** | **Total** |
| ***weight status*** | 34 | 27 | 0 | 61 |
| ***daily alcohol intake*** | 0 | 27 | 0 | 27 |
| ***daily cigarette consumption*** | 5 | 5 | 0 | 10 |
| ***blood pressure*** | 0 | 0 | 9 | 9 |
| ***total cholesterol*** | 11 | 9 | 0 | 20 |
| ***HDL cholesterol*** | 14 | 4 | 0 | 18 |
| ***triglycerides*** | 0 | 0 | 3 | 3 |
| ***diabetes*** | 0 | 0 | 11 | 11 |
| **Total** | 64 | 72 | 23 | 159 |

Table E.23: *WhatIf-Change1Factor?*: Breakdown of features selected by the users according to the type of the explanation.

| | **Goal** | | |
|---|---|---|---|
| **Explanation type** | *Understand the AI's reasoning* | *Change the outcome* | *Retain the outcome* |
| **Basic** | 89 | 89 | 89 |
| ***FactorsUsed?*** | 34 | 33 | 38 |
| ***FactorsNotUsed?*** | 49 | 29 | 36 |
| ***WhyNotC'?*** | 48 | 31 | 30 |
| ***HowtoGetC'?*** | 63 | 79 | 60 |
| ***HowtoStillGetC?*** | 29 | 28 | 55 |
| ***WhatIf-Change1Factor?-InPath*** | 20 | 34 | 10 |
| ***WhatIf-Change1Factor?-NotInPath*** | 24 | 33 | 38 |

Table E.24: Number of times each FQ type was selected for each goal.

- Contrastive explanations about a predicted class are preferred by users
- They are deemed especially valuable when users' expectations differ from predictions
- Contrastive explanations having a transfactual aspect help users' achieve their goals

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: