Explaining Decision-Tree Predictions by Addressing Potential Conflicts between Predictions and Plausible Expectations

Sameen Maruf[†] **Ingrid Zukerman**[†] **Ehud Reiter**[‡] **Gholamreza Haffari**[†]
[†]Dept. of Data Science and AI, Faculty of IT, Monash University, Victoria, Australia
[‡]Dept. of Computing Science, University of Aberdeen, Scotland, UK
[†]{firstname.lastname}@monash.edu [‡]e.reiter@abdn.ac.uk

Abstract

We offer an approach to explain Decision Tree (DT) predictions by addressing potential conflicts between aspects of these predictions and plausible expectations licensed by background information. We define four types of conflicts, operationalize their identification, and specify explanatory schemas that address them. Our human evaluation focused on the effect of explanations on users' understanding of a DT's reasoning and their willingness to act on its predictions. The results show that (1) explanations that address potential conflicts are considered at least as good as baseline explanations that just follow a DT path; and (2) the conflict-based explanations are deemed especially valuable when users' expectations disagree with the DT's predictions.

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.

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. Nonetheless, it is necessary to explain transparent models because (1) large datasets may not always be available, as is the case in our evaluation datasets (§ 4.1); (2) it is common practice to clarify the outcomes of opaque models by approximating them with transparent models (§ 2); and (3) even if these transparent models are understandable by ML experts, they may still be unclear to practitioners.

In this paper, we generate textual explanations of predictions made by a particular transparent model: DT. Our explanations address potential conflicts between aspects of these predictions and plausible expectations licensed by background information (i.e., expectations that "make sense" in light of this information). Specifically, we identify four types of conflicts whereby events that appeared unlikely or likely on the basis of background information happened or did not happen respectively; we then specify schemas that address these conflicts (§ 3).

We generated explanations for two datasets: *Telecom* and *Nursery*. In Telecom, a DT predicts whether a customer will churn (leave) or stay with the company based on their profile (e.g., whether they have a phone service and what are their monthly charges); 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). The bottom part of Table 1 illustrates an explanation generated for an instance in the Nursery dataset. The explanation addresses a potential conflict between (a) a plausible expectation that a child with a good childcare situation is likely to be *Wait listed*, and (b) the DT's prediction that the child will be *Priority accepted*.

Our human evaluation of the explanations generated for the two datasets (§ 4) considers users' overall preferences for different explanation types, and the effect of explanations on two explanatory goals: users' understanding of the DT's reasoning, and

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 114–127,
Aberdeen, Scotland, UK, 20-24 September 2021.

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: Explanation for the prediction of an instance in the Nursery dataset (bottom part); features used in the prediction and their values (top part).

their willingness to act on its predictions.[1] In addition, users rated the explanations on completeness, and on the presence of extraneous information.

The main findings of our user study are: (1) explanations that address potential conflicts are generally considered at least as good as baseline explanations that just follow a DT path; and (2) the conflict-based explanations are deemed especially valuable when users' expectations disagree with DT predictions. We stress that these 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.

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. Recently, Krause and Vossen (2020) identified additional triggers that should be addressed in explanations.

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 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; Štrumbelj and Kononenko, 2014; 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 was found by ranking all the paths leading to alternative outcomes according to their distance from the factual explanation. 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 the tree structure as binary meta-features, and minimizing an $L1$-like metric to retrieve the shortest statement. However, these works 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.

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 only those in (Biran and McKeown, 2017) use vague language to convey strength of evidence.

The approach described in this paper complements explanations by addressing both unexpected predictions and unexpected effects of features, 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.

Finally, and more broadly, expectation-theory posits that the surprisingness of an event may stem

[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 the system, and helping debug a system (Reiter, 2019).

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 our research, we employ a probabilistic formulation to identify potential conflicts between plausible expectations and aspects of DT predictions.

3 Justifying DT predictions

In this work, we explain the outcome predicted by a DT for sample 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 the features and values of a Nursery instance;[2] the DT then classifies this instance into one of three classes: *Reject*, *Wait list* and *Priority accept*.

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 variables; and (3) we identify these conflicts by comparing aspects of a DT prediction with plausible expectations derived from probabilistic relations.

3.1 Potential Conflicts

First, we define *potential conflicts*, and their building blocks: *plausible expectations* and *aspects of a DT prediction*. 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 and to the impact of a value j of feature x_i, denoted $x_{i,j}$. They are derived from prior and posterior probabilities of outcomes by means of relations R1-R3 and associated constraints (Table 2).

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

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

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

where *Prior*(c) is the prior probability of 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 $\mathcal{C}'$ is an alternative class with the highest *Posterior* probability. The posterior probability of a class c is calculated from training data for each feature value $x_{i,j}$. If it is high, it licenses an expectation for $x_{i,j}$ to yield class c; and if it is low, the expectation is for $x_{i,j}$ *not* to yield class c. For example, if according to the 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), it is plausible to expect a child with such parents' employment *not* to be *Priority accepted*.[3]

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*$(x_{i,j}, \mathcal{C})$. *Impact* 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 class predicted by a DT (R4), or when a feature value that was expected to have an impact does not (R5).[4]

R4. *Plausible* outcome $\neq$ *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 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. However, the determination of the *Impact* of a variable in R5 depends on the model, e.g., as seen above, variable impact for DTs is determined by path membership.

The values of relations R1-R3 are obtained from discretized probabilistic relations (§ 3.1.1).

3.1.1 Discretizing probabilistic relations

To generate explanations that use language to communicate relative probabilities, we harness the research in (Elsaesser and Henrion, 1989), which

[2] Sample features for the evaluation datasets and their values appear in Table 4; the DT feature values for the Nursery dataset are described in Table 10, Appendix A.

[3] Our formalism assumes that users are aware of the prior and posterior probabilities of outcomes (they were given this information in our evaluation, § 4.2), and employs these probabilities as the basis for explaining DT predictions. Hence, it differs from probabilistic models, such as Bayesian Networks or Naïve Bayes, which use probabilities to infer outcomes.

[4] Biran and McKeown (2017) consider situations where a variable 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.

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

Table 2: Definition of potential conflicts (explanations appear in Tables 1 and 3): $\mathcal{C}$ denotes the *Predicted* class, and $\mathcal{C}'$ denotes an alternative class that has the highest *Posterior* probability (*Post* is shorthand for *Posterior*); the colours of (in)equalities match those in Figure 1; text in Column 4 indicates surprise about the plausible outcome in Column 3, and text in Column 6 expresses surprise about the plausible impact of $x_{i,j}$ in Column 5.

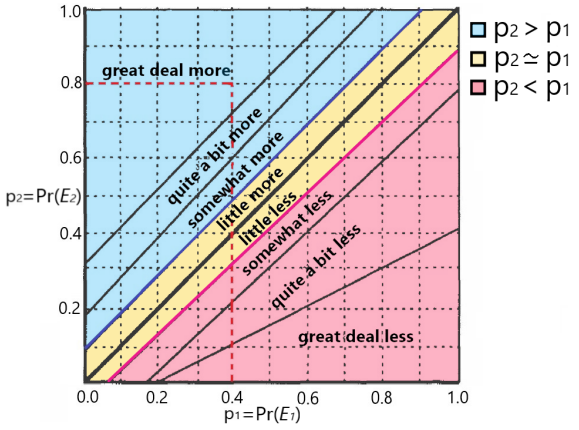

Figure 1: Verbal mapping of relative probabilities.

maps probability differences into verbal expressions. 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 $p_2 = 0.8$ (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.

3.1.2 Identifying Potential Conflicts

Table 2 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* derived from the relations defining the conflict (Column 2); Column 4 shows the actual *Predicted* class $\mathcal{C}$; Column 5 displays the plausible expected *impact* of $x_{i,j}$ — a feature value that satisfies the relations defining a conflict (Column 2) is always expected to have an impact; 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 illustrated with examples from the Nursery dataset.

***Plausible*$\neg\mathcal{C}$/*Predict*$\mathcal{C}$** (top segment in Table 2). 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). 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 3): 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]

[5] As seen in Table 10, Appendix A, the term "critical childcare" indicates high insecurity in obtaining this service.

Schema	Sample Generated Explanations for the Nursery dataset
Conflict-based (outcome only): *Plausible*$\neg\mathcal{C}$*/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: ***Path*** + $x^*_{i,j}$ + $\mathcal{C}$	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*$\mathcal{C}$*-*$x_{i,j}$*NoImpact*	
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 + ***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 and impact of feature value): *Plausible*$\mathcal{C}'$*/Predict*$\mathcal{C}$*-*$x_{i,j}$*NoImpact*	
Preamble: $x^*_{i,j}$ + R3 + $\mathcal{C}'$ + $\mathcal{C}$	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 + ***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%).
Basic (no conflict): counterpart of *Plausible*$\mathcal{C}'$*/Predict*$\mathcal{C}$*-*$x_{i,j}$*NoImpact*	
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%).

Table 3: Schemas that address three of the potential conflicts defined in Table 2 and Basic schema (our baseline), with sample explanations; relative probabilities are described in Figure 1; the selection of a *pivot feature* value is described in § 3.2; font denotes **feature values** and ***features*** in the DT path, and *Classes*.

Plausible*$\mathcal{C}$*/Predict*$\mathcal{C}$*-*$x_{i,j}$*NoImpact (bottom of second segment in Table 2). 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 the posterior probability of class $\mathcal{C}$ in light of at least one other 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 3): 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 ≠ actual *Impact*).

Plausible*$\mathcal{C}'$*/Predict*$\mathcal{C}$** (third segment in Table 2). Here, an alternative outcome $\mathcal{C}'$ 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*$\mathcal{C}'$*/Predict*$\mathcal{C}$ goes further, nominating a potential alternative class $\mathcal{C}'$. The expectation for $\mathcal{C}'$ is plausible because its posterior probability is higher than its prior (R2) and the posterior of $\mathcal{C}$ (R3), and $\mathcal{C}'$ 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 3): In the Nursery dataset, children with good *current childcare* are more likely to get *Wait listed* than *Priority accepted* (R3: *Posterior*($\mathcal{C}'$) > *Posterior*($\mathcal{C}$)). However, a particular child with certain feature values and good *current childcare* gets *Priority accepted* (R4: *Plausible* outcome $\mathcal{C}'$ ≠ *Predicted* class $\mathcal{C}$).

3.2 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.[6] 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 (§ 4.3.3).

[6] In the future, we will consider higher-dimensional spaces, which may require addressing several features with conflicts or adopting a different strategy, e.g., an interactive approach.

3.2.1 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¬C/PredictC* and *PlausibleC/PredictC-$x_{i,j}$NoImpact*, we choose the $x^*_{i,j}$ with the maximum absolute difference between *Posterior*$(\mathcal{C}|\, x^*_{i,j})$ and *Prior*$(\mathcal{C})$ for the *Predicted* class $\mathcal{C}$. For the *PlausibleC′/PredictC* variants, we select the $x^*_{i,j}$ with the maximum difference between *Posterior*$(\mathcal{C}'|\, x^*_{i,j})$ and *Posterior*$(\mathcal{C}|\, x^*_{i,j})$.

3.2.2 Realizing explanations

A Conflict-based explanation has 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. Table 3 displays schemas that address three potential conflicts, and one Basic schema (which is our baseline), together with sample explanations; an explanation that illustrates *PlausibleC′/PredictC* "vanilla" appears in Table 1 (the schema for this potential conflict is [*Preamble*: $x^*_{i,j}$ + R3 + $\mathcal{C}'$ + $\mathcal{C}$; *Resolution*: ***Path*** + $x^*_{i,j}$ + $\mathcal{C}$]). Since the focus of our research is on content selection, the explanations are realized by means of domain-independent programmable templates.

The ***Preamble*** presents probabilistic relations that license plausible expectations. The preambles of *Plausible¬C/PredictC* and *PlausibleC/PredictC-$x_{i,j}$NoImpact* describe relation R1; and those of the *PlausibleC′/PredictC* 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 4), 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*.

4 Empirical Evaluation

Our evaluation considers two main questions: (Q1) How do Conflict-based explanations compare to Basic baseline explanations? (Q2) Which types of Conflict-based explanations are preferred to Basic explanations, if any?

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
internet service:	Fiber optic,	DSL,	no
online security:	yes,	NA (no internet service),	no
tenure (months with company):	1 month,		72 months
monthly charges:	\$19,		\$117

Table 4: Classes, sample features (in the presentation order used in our explanations) and values in the evaluation datasets; the feature values in the Nursery DT are described in Table 10, Appendix A.

Next, we describe our datasets and classifier, followed by our experimental design and results.[7]

4.1 Datasets

We used two datasets, which were pre-processed as described in Appendix A: *Nursery* (Olave et al., 1989), which has 12630 instances and three classes; and *Telecom*, which has 3302 instances and two classes. These datasets were chosen due to their diverse character, and the differences in number and types of features and predicted classes. Both 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. It 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).[8] 78% of the Nursery test samples and all the Telecom test samples had at least one potential conflict.

4.2 Experiment Design

Our experiment starts with a demographic questionnaire followed by the body of the survey.

The body of the survey begins with a narrative immersion, where participants are 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 cus-

[7] We have addressed the recommendations for human evaluation in (Howcroft et al., 2020). The experiment and data are available at https://doi.org/10.26180/15147462.

[8] 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.

tomers will churn (leave) or stay (Telecom). The participants are then given a brief account of how an AI makes predictions, and shown the features and values that are input to the AI (illustrated in Table 4) — a screenshot of the introductory narrative for the Nursery dataset appears in Figure 2, Appendix D. Next, a sequence of scenarios is presented in random order, each pertaining to a different family/customer — a screenshot of a Nursery scenario appears in Figure 3, Appendix D. Between scenarios, a short version of the Matching Familiar Figures Test (MFFT) (Cairns and Cammock, 1978) is shown as a filler.

Scenario description. We chose scenarios with the strongest available potential conflict (using a procedure similar to that in § 3.2.1), and diverse pivot and explanatory variables. Scenarios without conflicts were excluded from our evaluation, as they warrant only a Basic explanation. To ensure that all the potential conflicts in Table 2 are represented, we chose eight Nursery scenarios (four each for *Wait list* and *Priority accept*)[9] and ten Telecom scenarios (five each for *Stay* and *Churn*).

Each scenario begins by showing a set of features such as those in Table 4, together with their values for a particular family/customer and the *Prior* and *Posterior* probabilities of the possible classes. Users are then asked to make an educated guess about the predicted class, after which they are shown the prediction made by the DT.

Next, users are given two side-by-side explanations for this prediction: Conflict-based versus Basic. The selection of a side (left or right) for an explanation type is randomized between scenarios, but all the participants see the same side-by-side configuration for a given scenario.

Users' views about explanations. Users are 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 misleading/contradictory/irrelevant information, as well as the understandability of the AI's reasoning and their willingness to act on the prediction on the basis of an explanation (exact statements appear in the screenshot in Figure 3, Appendix D). The first three attributes come from Hoffman *et al.*'s (2018) *Explanation Satisfaction Scale*, and the third and fourth attributes are our explanatory goals (§ 1). Participants are also asked which explanation(s) they prefer, if any.

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

Question	Option	Nursery	Telecom
Gender	Male / Female	12 / 28	25 / 17
Age	18-34 years old	33	37
Ethnicity	Asian / Caucasian	17 / 17	28 / 4
English proficiency	Medium / High	5 / 36	5 / 37
Education	Bachelor / Master	13 / 13	14 / 22
ML expertise	Low / Med-High	27 / 14	18 / 24
Domain familiarity	Yes / No	9 / 32	31 / 11

Table 5: Descriptive statistics: for gender, age, ethnicity and education, we present the options that had most participants; domain familiarity was self-rated.

To detect unreliable responses, we inserted an attention question, where we asked users to indicate whether a neutral statement about the background information in the scenario was true or false.

Participant cohorts. To avoid participant fatigue, we conducted a separate experiment for each dataset — details appear in Appendix B. The surveys were implemented in the Qualtrics survey software, and conducted on SONA.

We obtained a total of 83 valid responses out of 109 — 41 for Nursery and 42 for Telecom (responses were validated based on the answers to the attention questions and the total time spent on the experiment). Table 5 shows the statistics for the Nursery and the Telecom cohorts.

4.3 Results

To answer Q1, we compared Conflict-based explanations with Basic ones for each dataset in terms of the four explanatory attributes mentioned above, and user preferences (§ 4.3.1). We also analyzed the influence of various independent variables on users' ratings of Conflict-based explanations compared to Basic ones (§ 4.3.2). To answer Q2, we analyzed how individual Conflict-based explanations compare to their Basic counterparts (§ 4.3.3).

Statistical significance for the ratings of the four attributes for Conflict-based versus Basic explanations was obtained using Wilcoxon signed-rank test; a one- and two-proportion Z-test was respectively used for the proportion of preference counts within one population and between two populations. Statistical significances were adjusted with Holm-Bonferroni correction for multiple comparisons (Holm, 1979).

4.3.1 Conflict-based vs Basic explanations

Our results show that for the Nursery dataset (top of Table 6), Conflict-based explanations are deemed

Attribute	Conflict-based Mean (SD)	Basic Mean (SD)	Stat. Sig.
Nursery			
Complete	3.43 (0.97)	3.00 (0.98)	< 0.001
Misleading...	2.72 (1.00)	2.55 (0.89)	< 0.05
Understandable	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
Misleading...	3.00 (1.14)	2.81 (1.05)	–
Understandable	3.49 (0.92)	3.33 (0.87)	–
Willingness to act	3.16 (0.99)	3.09 (0.94)	–

Table 6: Comparison between explanation types: scores and statistical significances (Wilcoxon signed-rank test); a lower score is better for Misleading... , and a higher score is better for the other attributes.

	Count					χ^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 7: Preference for an explanation type: χ^2 statistic and statistical significances (one-proportion Z-test) calculated from clear preferences for Conflict-based/ Basic explanations.

significantly more complete, understandable and enticing to act on a DT's prediction than Basic explanations. However, Conflict-based explanations are also deemed more misleading/contradictory/irrelevant than Basic explanations. For Telecom (bottom of Table 6), Conflict-based explanations are considered significantly more complete than Basic explanations, but equivalent for the other three attributes.

In terms of preferences, for both datasets, the majority of users prefer Conflict-based explanations to Basic ones (Table 7). However, the two datasets differ 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 7), with a higher percentage of users preferring the Conflict-based explanations for the Nursery dataset.

4.3.2 Influence of independent variables

Our experiment has several independent variables, including predicted outcome, pivot feature, explanation length and (dis)agreement between an expected and a predicted class. 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 Basic explanations (Table 9).

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≠Exp	3.48 (0.99)	2.90 (0.99)	< 0.01
Misleading...	Pred=Exp	2.80 (1.03)	2.54 (0.90)	< 0.05
	Pred≠Exp	2.57 (0.92)	2.57 (0.86)	–
Understand-able	Pred=Exp	3.61 (1.07)	3.20 (0.99)	< 0.01
	Pred≠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≠Exp	3.40 (1.12)	2.87 (0.98)	< 0.01
Telecom				
Complete	Pred=Exp	3.18 (0.97)	2.99 (0.95)	–
	Pred≠Exp	3.35 (1.04)	2.72 (1.01)	< 0.01
Misleading...	Pred=Exp	3.08 (1.14)	2.83 (1.05)	–
	Pred≠Exp	2.75 (1.10)	2.75 (1.08)	–
Understand-able	Pred=Exp	3.45 (0.90)	3.35 (0.86)	–
	Pred≠Exp	3.62 (0.98)	3.25 (0.93)	–
Willingness to act	Pred=Exp	3.14 (0.97)	3.17 (0.90)	–
	Pred≠Exp	3.25 (1.07)	2.83 (1.04)	< 0.05

Table 8: Effect of (dis)agreement between users' expectations and DT predictions: scores and statistical significances (Wilcoxon signed-rank test).

Interestingly, our analysis shows that (dis)agreement between users' expectations *according to their survey answers* and the class *Predicted* by the DT has a significant influence on the ratings of Conflict-based explanations compared to Basic ones (users' answers disagreed with a *Predicted* class when they selected a different class or *Can't Decide* – see options in Figure 3, Appendix D).

For the Nursery dataset, the general results obtained for Conflict-based versus Basic explanations hold for completeness, understandability and willingness to act on predictions for both agreement and disagreement between users' expectations and DT predictions (top of Table 8). However, Conflict-based explanations were deemed to contain more misleading/contradictory/irrelevant information than Basic ones only when users' expectations matched DT predictions. This suggests that the additional information provided by Conflict-based explanations is welcome when a prediction is *not* as expected.

For the Telecom dataset, Conflict-based explanations were considered more complete and enticing

Basic vs Conflict-based	Nursery							Telecom						
	Count					χ^2	Stat. Sig.	Count					χ^2	Stat. Sig.
	Conflict	Basic	Both	None	Total			Conflict	Basic	Both	None	Total		
Basic vs *Plausible¬C/PredictC*	33	12	3	14	62	9.80	< 0.01	46	21	2	15	84	9.33	< 0.01
Basic vs *PlausibleC/PredictC-$x_{i,j}$NoImp*	8	6	1	6	21	0.29	–	14	20	2	6	42	1.06	–
Basic vs *PlausibleC′/PredictC* "vanilla"	33	13	6	9	61	8.70	< 0.01	23	6	3	10	42	8.53	< 0.05
Basic vs *PlausibleC′/PredictC-$x_{i,j}$NoImp*	38	14	3	6	61	11.08	< 0.01	34	31	4	15	84	0.14	–

Table 9: Preference for individual explanation types: χ^2 statistic and statistical significances (one-proportion Z-test) calculated from clear preferences for Conflict-based/Basic explanations (*NoImp* is shorthand for *No Impact*).

to act only when users' expectations differed from DT predictions (bottom of Table 8).

In terms of preferences, 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, Table 12 in Appendix C). However, for Telecom, Conflict-based explanations were preferred only when users' expectations disagreed with DT predictions (*p-value* < 0.001).

4.3.3 Individual Conflict-based explanations

Our comparison between individual Conflict-based explanations and their Basic counterparts shows that a statistically significantly higher proportion of users preferred *Plausible¬C/PredictC* and *PlausibleC′/PredictC* "vanilla" to Basic explanations for both Nursery and Telecom (Table 9). But *PlausibleC′/PredictC-$x_{i,j}$NoImpact* was preferred to its Basic counterpart only for the Nursery dataset, where it had the largest margin. Finally, *PlausibleC/PredictC-$x_{i,j}$NoImpact*, which addresses a conflict with respect to variable impact only, was deemed equivalent to its Basic counterpart for both datasets. However, according to (Biran and McKeown, 2017), users were more satisfied with explanations about unexpected variable impacts than no explanation. This suggests that further studies are required to determine the conditions for explaining unexpected variable impacts.

The results in Table 9 indicate that if a DT prediction has several qualifying conflicts, they should be prioritized in the following order: *Plausible¬C/PredictC* $\succ$ *PlausibleC′/PredictC* "vanilla" $\succ$ *PlausibleC′/PredictC-$x_{i,j}$NoImpact*.

5 Conclusion

Our approach for explaining DT predictions addresses potential conflicts between aspects of these predictions and plausible expectations licensed by background information. To this effect it operationalizes the identification of four types of conflicts, and specifies schemas for generating explanations that address these conflicts. Our approach is model agnostic, except for the determination of the actual impact of a variable, which is readily available in most ML models.

Our evaluation on the Nursery and Telecom datasets shows that (1) explanations addressing potential conflicts between DT predictions and plausible expectations from background information are considered at least as good as baseline explanations; and (2) the 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.

Our approach has the following limitations, which we propose to address in the future: (1) it does not perform feature selection to reduce long paths in a DT; (2) Conflict-based explanations address only one pivot feature; and (3) the explanations omit information about DT accuracy for particular instances.

Our evaluation has the following limitations: (1) we cannot divorce length from explanation type, as Conflict-based explanations are about twice as long as Basic ones; (2) the cohorts for the two datasets had different demographics, so, given the size of our population, it is not possible to attribute differences in our results for each dataset to domain or demographic differences; and (3) we could not recruit participants with relevant experience, but in light of our narrative immersion and the general accessibility of the concepts in the explanations, we believe that our results are informative.

Acknowledgments

This research was supported in part by grant DP190100006 from the Australian Research Council. We thank Marko Bohanec, one of the creators of the Nursery dataset, for helping us understand the features and their values. We also thank the anonymous reviewers for their helpful comments.

References

Or Biran and Kathleen McKeown. 2017. Human-centric justification of Machine Learning predictions. In *IJCAI'17*, pages 1461–1467, Melbourne, Australia.

Ed Cairns and Tommy Cammock. 1978. Development of a more reliable version of the Matching Familiar Figures test. *Developmental Psychology*, 14(5):555–560.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable Machine Learning. Arxiv:1702.08608.

Christopher Elsaesser and Max Henrion. 1989. Verbal expressions for probability updates: How much more probable is "much more probable"? In *UAI'89*, pages 319–330, Windsor, Canada.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The WEKA workbench. In *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques (Fourth ed.)"*. Morgan Kaufmann Publishers, San Francisco, California.

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. Arxiv:1812.04608.

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

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

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *INLG2020*, pages 169–182, Dublin, Ireland.

Laurent Itti and Pierre Baldi. 2009. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306.

Kevin B. Korb, Richard McConachy, and Ingrid Zukerman. 1997. A cognitive model of argumentation. In *CogSci 1997*, pages 400–405, Stanford, California.

Lea Krause and Piek Vossen. 2020. When to explain: Identifying explanation triggers in human-agent interaction. In *NL4XAI'2020*, pages 55–60, Dublin, Ireland.

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

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *NIPS'17*, pages 4768–4777, Long Beach, California.

Manuel Olave, Vladislav Rajkovic, and Marko Bohanec. 1989. An application for admission in public school systems. In I.Th.M. Snellen, W.B.H.J. van de Donk, and J.-P. Baquiast, editors, *Expert Systems in Public Administration*, chapter 10, pages 145–160. Elsevier.

Andrew Ortony and Derek Partridge. 1987. Surprisingness and expectation failure: What's the difference? In *IJCAI'87*, page 106–108, Milan, Italy.

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

Ehud Reiter. 2019. Natural language generation challenges for explainable AI. In *NL4XAI'2019*, pages 3–7, Tokyo, Japan.

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD'16*, pages 1135–1144, San Francisco, California.

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI-18*, pages 1527–1535, New Orleans, Louisiana.

Kacper Sokol and Peter Flach. 2020a. LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. Arxiv:2005.01427.

Kacper Sokol and Peter Flach. 2020b. One explanation does not fit all: The promise of interactive explanations for Machine Learning transparency. Arxiv:2001.09734.

Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martin Pereira. 2020. Generation and evaluation of factual and counterfactual explanations for decision trees and fuzzy rule-based classifiers. In *WCCI*, pages 1–8, Glasgow, Scotland.

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

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665.

Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. 2018. Contrastive explanations with local Foil Trees. In *WHI'2018*, pages 41–46, Stockholm, Sweden.

Ingrid Zukerman and Richard McConachy. 1993. Generating concise discourse that addresses a user's inferences. In *IJCAI'93*, pages 1202–1207, Chambery, France.

A Datasets

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 10: Description of feature values in the Nursery DT; 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.

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 domain to those in Table 4, and the names of the retained classes to *Reject*, *Wait list* and *Priority accept*. The recoded feature values are described in Table 10.

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. 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).

Table 11 shows final classes in the two datasets and the breakdown of the training/test sets.

Partition	**Nursery**				**Telecom**		
	Reject	*Wait list*	*Priority accept*	**Total**	*Stay*	*Churn*	**Total**
Training	3485	3414	3205	10104	1596	1057	2653
Testing	835	852	839	2526	390	259	649
Total	4320	4266	4044	12630	1986	1316	3302

Table 11: Breakdown of classes for the training set and the test set for the Nursery and Telecom datasets.

B Experiment Design

The scenarios studied in this paper compare Conflict-based explanations with Basic explanations for two datasets. However, our experiment contains additional scenarios, which compare two Conflict-based explanations. 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 behave 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* $= 0.388$).

C Results

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

Table 12: Preferences broken up by (dis)agreement between users' expectations and DT predictions: χ^2 statistic and statistical significances (one-proportion Z-test) calculated from clear preferences for Conflict-based/Basic explanations.

D Screenshots from the Nursery survey

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 2: 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.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
This explanation helps me understand the reasoning of the AI system.	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
This explanation has misleading, contradictory or irrelevant information.	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
This explanation is complete (it is not missing information).	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯
Based on 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 3: Background information about the Nicholson family scenario; question about the expected outcome; model prediction (displayed after an outcome has been selected); *PlausibleC′/PredictC* explanation "vanilla" (A) and Basic explanation (B) for this scenario; attention question; preferences for explanations; features that determine expectations; request for suggestions.