Explaining Qualitative Preference Models

Wietske Visser and **Koen V. Hindriks** and **Catholijn M. Jonker**[1]

Abstract. We propose an explanation facility for a qualitative preference representation framework. We show how an explanation can be provided for qualitative, multi-criteria preferences based on the criteria that are used to decide preferences between outcomes. Such a facility provides an important tool for a user to understand how preferences are determined. We show that this facility can also be used by a user to inform the system about its preferences. Such a user-provided explanation can be used for updating and improving a preference model maintained by the system.

1 INTRODUCTION

A preference representation framework provides a tool for determining preferences between outcomes. That is, for any two outcomes it can determine whether one is strictly preferred to the other, both are equally preferred, or they are incomparable. In this paper, we discuss an additional facility, namely the *explanation* of preferences maintained by such a system. Explanation of preferences is useful and important in many cases, such as situations where a decision maker has to explain his decision to other actors; where a decision support system that is elicited from an expert has to explain its list of recommended options to a non-expert user; or where agents may give each other feedback on offers in negotiation, without revealing all their preferences [6]. Another reason to use explanation is to improve users' confidence in a system, since lack of confidence is an obstacle to acceptance and practical use of the system [7]. In these cases, it is not satisfactory to just present the preference model. Although this model does contain all information on which the preference is based, the format is not suitable for presentation to a user. First, the model is too technical for the average human user to interpret. Second, even experts may have trouble interpreting the model since it may be quite large, and hence it would be hard to quickly find the reason behind the preference.

Besides explaining someone's preferences to another party, explanation may also be used 'in reverse' during preference elicitation and updating. Here the idea is as follows. The user is not only asked to state his preference between two given outcomes, but also to explain this preference. This explanation can then be used to update the preference model in such a way that the explanation for the user's preference that would be generated by the updated model coincides with the explanation given by the user.

In this paper we propose an approach to generate explanations from a preference model and to use explanations to update a preference model. The preference models we consider are expressed in a particular preference representation framework called Qualitative Preference Systems (QPS) [8, 9]. QPS is a general framework for the representation of qualitative, multi-criteria preferences. In Section 2, we give a summary of the QPS framework. In Section 3 we propose a way to explain qualitative preferences by the deciding criteria, and discuss in particular how this can be implemented for QPS models. In Section 4 we discuss how such explanations, if given by the user of a system, can be used to update the system's current model of the user's preferences. We give detailed interaction diagrams that indicate when and how a QPS preference model should be altered. Section 5 concludes the paper.

[1] Interactive Intelligence Group, Delft University of Technology, The Netherlands, email: {Wietske.Visser, K.V.Hindriks, C.M.Jonker}@tudelft.nl

2 QUALITATIVE PREFERENCE SYSTEMS

The main aim of the Qualitative Preference System (QPS) framework [8, 9] is to determine preferences between *outcomes* in a purely *qualitative* way. Outcomes are defined as variable assignments that respect the constraints in a *knowledge base*. The preferences between outcomes are based on multiple *criteria*. Every criterion can be seen as a *reason* for preference, or as a preference from one particular *perspective*. We distinguish between simple and compound criteria. Simple criteria are based on a single variable. Multiple (simple) criteria can be combined in a compound criterion to determine an overall preference. QPS distinguishes between two kinds of compound criteria: cardinality criteria and lexicographic criteria. The subcriteria of a cardinality criterion all have equal priority, and preference is determined by a kind of voting mechanism that counts the number of subcriteria that support a certain preference and those that do not. In a lexicographic criterion, the subcriteria are ordered by priority and preference is determined by the subcriteria with the highest priority; lower priority subcriteria only influence the preference if the higher priority subcriteria are indifferent.

Definition 1. (Qualitative Preference System [8]) A *Qualitative Preference System (QPS)* is a tuple $\langle Var, Dom, K, \mathcal{C} \rangle$. *Var* is a finite set of *variables*. Every variable $X \in Var$ has a domain $Dom(X)$ of possible values. K (a *knowledge base*) is a set of constraints on the assignments of values to the variables in *Var*. An *outcome* α is an assignment of a value $x \in Dom(X)$ to every variable $X \in Var$, such that no constraints in K are violated. Ω denotes the set of all outcomes: $\Omega \subseteq \prod_{X \in Var} Dom(X)$. α_X denotes the value of variable X in outcome α. $\mathcal{C}$ is a finite rooted tree of criteria, where leaf nodes are simple criteria and other nodes are compound criteria. Child nodes of a compound criterion are called its subcriteria. The root of the tree is called the top criterion. Weak preference between outcomes by a criterion c is denoted by the relation $\succeq_c$. $\succ_c$ denotes the strict subrelation, $\approx_c$ the indifference subrelation. $\alpha \lambda_c \beta$ denotes that $\alpha \not\preceq_c \beta$ and $\beta \not\preceq_c \alpha$.

Definition 2. (Simple criterion [8]) A *simple criterion* c is a tuple $\langle X_c, \geq_c \rangle$, where $X_c \in Var$ is a variable, and $\geq_c$, a preference relation on the possible values of X_c, is a preorder on $Dom(X_c)$. $>_c$ is the strict

Table 1. Explanations

	lexicographic criterion c	goal-based cardinality criterion c
$\alpha \succ_c \beta$	any subcriterion $s \in C_c$ such that $\alpha \succ_s \beta$ and for all $s' \in C_c$: if $s' \rhd s$ then $\alpha \approx_{s'} \beta$ and if $s' \triangle s$ then $\alpha \succeq_{s'} \beta$ or there is a $s'' \in C_c (s'' \rhd_c s'$ and $\alpha \not\approx_{s''} \beta)$	the set of subgoals $g \in C_c$ such that $\alpha \succ_g \beta$
$\alpha \approx_c \beta$	for all subcriteria $s \in C_c$: $\alpha \approx_s \beta$	the set of subgoals $g \in C_c$ such that $\alpha \succ_g \beta$ plus the set of subgoals $g \in C_c$ such that $\beta \succ_g \alpha$
$\alpha \wedge_c \beta$	1: any subcriterion $s \in C_c$ such that $\alpha \wedge_s \beta$ and for all $s' \in C_c$: if $s' \rhd s$ then $\alpha \approx_{s'} \beta$ 2: any pair of subcriteria (s_1, s_2) where $s_1, s_2 \in C_c$ such that $\alpha \succ_{s_1} \beta$ and $\beta \succ_{s_2} \alpha$ and $s_1 \triangle_c s_2$ and for all $s' \in C_c$: if $s' \rhd_c s_1$ or $s' \rhd_c s_2$ then $\alpha \approx_{s'} \beta$	n/a

subrelation, $\doteq_c$ is the indifference subrelation. A simple criterion $c = \langle X_c, \geq_c \rangle$ *weakly prefers* an outcome α over an outcome β, denoted $\alpha \geq_c \beta$, iff $\alpha_{X_c} \geq_c \beta_{X_c}$.

Definition 3. (Goal [9]) A QPS *goal* is a simple criterion $\langle X, \geq \rangle$, where $X \in Var$ is a Boolean variable ($Dom(X) = \{\top, \bot\}$), and $\top \succ \bot$.

Definition 4. (Goal-based cardinality criterion [9]) A *goal-based cardinality criterion* c is a tuple $\langle C_c \rangle$ where C_c is a nonempty set of goals (the *subcriteria* or *subgoals* of c). A goal-based cardinality criterion $c = \langle C_c \rangle$ *weakly prefers* an outcome α over an outcome β, denoted $\alpha \geq_c \beta$, iff $|\{s \in C_c \mid \alpha \succ_s \beta\}| \geq |\{s \in C_c \mid \alpha \not\succeq_s \beta\}|$, or equivalently, iff $|\{s \in C_c \mid \alpha_{X_s} = \top\}| \geq |\{s \in C_c \mid \beta_{X_s} = \top\}|$.

Note that a goal-based cardinality criterion can only have goals as subcriteria. This is to guarantee transitivity of the preference relation induced by a cardinality criterion [8].

Definition 5. (Lexicographic criterion [8]) A *lexicographic criterion* c is a tuple $\langle C_c, \rhd_c \rangle$, where C_c is a nonempty set of criteria (the *subcriteria* of c) and $\rhd_c$, a *priority relation* among subcriteria, is a strict partial order (a transitive and asymmetric relation) on C_c. $s \triangle_c s'$ denotes that $s \not\rhd_c s'$ and $s' \not\rhd_c s$. A lexicographic criterion $c = \langle C_c, \rhd_c \rangle$ *weakly prefers* an outcome α over an outcome β, denoted $\alpha \geq_c \beta$, iff $\forall s \in C_c (\alpha \geq_s \beta \vee \exists s' \in C_c (\alpha \succ_{s'} \beta \wedge s' \rhd_c s))$.

3 EXPLAINING PREFERENCES

Ideally, any explanation given to a human user should be easily understandable by that user. Therefore, both the content and the format of the explanation matter. [6] distinguishes between two steps in explanation generation. First, the content of the explanation has to be selected. Next, a natural language explanation has to be generated. Like [6], we focus on the first step and only look at the content of an explanation. An example of natural language generation for evaluative arguments such as explanations can be found in [2].

We are not aware of any work on the explanation of preferences represented in a qualitative framework, but some work has been done on the explanation of (decisions based on) quantitative preferences. Klein and Shortliffe [5] presented strategies for automatically explaining decisions based on Multiattribute Value Theory (a quantitative preference representation framework). The explanations are based on the compellingness of objectives. Labreuche [6] presents a general framework for explaining the results of a multi-attribute preference model. He takes a quantitative approach where the utilities of the combined criteria are weighted and summed to obtain an overall utility. He develops a formal framework that justifies the selection of arguments (criteria) to be presented as explanation of a preference.

One of the main differences between quantitative and qualitative approaches to multi-criteria preference modelling is that quantitative approaches are compensatory, whereas their qualitative counterparts are not. In quantitative approaches, a low score on one criterion can be compensated by high scores on other criteria, even if the other criteria are less important, as long as the scores are high enough. In qualitative approaches, this is not possible. For example, if one outcome is preferred to another according to the highest priority subcriterion of a lexicographic criterion, it will also be preferred according to this lexicographic criterion, no matter what the preferences of the other subcriteria are. This allows us to precisely identify the criteria that are 'responsible' or 'deciding' for the overall preference. It is our intuition that these criteria also provide a natural explanation for the overall preference.

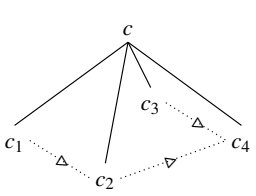
Explanations for preferences by QPS criteria

We now turn to the question how a preference between two outcomes by a QPS criterion can be explained. The answer to this question depends on the kind of criterion that is considered. Preferences by simple criteria (including goals) are self-explanatory, since they follow immediately from the specification of the simple criterion or goal. For example, a simple criterion c strictly prefers an outcome α to an outcome β because α's value of X_c is better than β's value of X_c. Similarly, a goal c strictly prefers an outcome α to an outcome β because α satisfies c but β does not. Of course, these facts may in turn require explanation. But since this would be explanation of knowledge (factual information about outcomes) rather than preferences, we do not discuss this topic here.

Preferences by compound criteria can be explained by the subcriteria that are *deciding* in the overall preference. Which subcriteria are deciding depends both on the kind of compound criterion (lexicographic or goal-based cardinality criterion) and on the kind of preference (strict, equal or incomparable). The deciding factor may be a single subcriterion, a pair, or even a set of multiple subcriteria that together determine the overall preference. In the following, we discuss the deciding subcriteria (and hence the explanations) for both kinds of compound criteria and for all kinds of preferences. An overview is given in Table 1.

Lexicographic criteria

Strict preference Suppose a lexicographic criterion c strictly prefers an outcome α over an outcome β ($\alpha \succ_c \beta$). The explanation of this preference is given by a subcriterion s that strictly prefers α to β ($\alpha \succ_s \beta$). But not just any subcriterion that strictly prefers α to β will do. First, every subcriterion s' with a higher priority than s ($s' \rhd_c s$) has to be indifferent: $\alpha \approx_{s'} \beta$, otherwise s would



$\alpha \approx_{c_1} \beta \approx_{c_1} \gamma \succ_{c_1} \delta$
$\alpha \succ_{c_2} \beta \approx_{c_2} \gamma \succ_{c_2} \delta$
$\alpha, \delta \succ_{c_3} \beta \approx_{c_3} \gamma;\ \alpha \wedge_{c_3} \delta$
$\gamma \approx_{c_4} \beta \succ_{c_4} \delta \succ_{c_4} \alpha$

Figure 1. Example lexicographic criterion

have been overruled by s'. Second, every subcriterion s' whose priority is incomparable to that of s ($s' \triangle_c s$) and which is not overruled ($\forall s'' \rhd_c s' : \alpha \approx_{s''} \beta$) has to agree with s or be indifferent ($\alpha \succeq_{s'} \beta$), otherwise s would not have decided the preference by c.

Example 1. Consider the lexicographic criterion c displayed in Figure 1. It has four subcriteria c_1, c_2, c_3, c_4 such that $c_1 \rhd_c c_2 \rhd_c c_4$ and $c_3 \rhd_c c_4$. The nature of the subcriteria is unspecified, but their preferences regarding four outcomes $\alpha, \beta, \gamma, \delta$ are given. The criterion c strictly prefers α over β: $\alpha \succ_c \beta$. The subcriteria that can explain this preference are c_2 and c_3. c_3 strictly prefers α over β, and is undominated. c_2 also strictly prefers α over β, and is dominated only by an indifferent criterion (c_1). Neither is 'contradicted' by a criterion with incomparable priority.

Equal preference A lexicographic criterion c is only indifferent between two outcomes α and β ($\alpha \approx_c \beta$) if all its subcriteria are indifferent between α and β. No single subcriterion is deciding in the overall preference, but all subcriteria contribute equally (note that priority does not matter, since indifferent criteria do not overrule lower priority criteria). This means that the explanation of the indifference is given by the fact that all subcriteria are indifferent.

Example 2. Consider again the lexicographic criterion c in Figure 1. c is indifferent between β and γ, because all subcriteria are indifferent between β and γ.

Incomparability If a lexicographic criterion c cannot compare between two outcomes α and β ($\alpha \wedge_c \beta$), this incomparability can have two possible reasons. First, the incomparability may result from a subcriterion s that cannot compare between α and β ($\alpha \wedge_s \beta$). Like in the case of strict preference, every subcriterion s' with a higher priority than s ($s' \rhd_c s$) has to be indifferent: $\alpha \approx_{s'} \beta$, otherwise s would have been overruled by s'.

Example 3. Consider again the lexicographic criterion c in Figure 1. c cannot compare between α and δ. This is due to subcriterion c_3, which cannot compare between α and δ, and which is not overruled by any other subcriterion. Therefore c_3 explains c's incomparability between α and δ.

Second, the incomparability may result from two conflicting subcriteria that do not overrule each other. That is, there is one subcriterion s_1 that strictly prefers α to β ($\alpha \succ_{s_1} \beta$), and all higher priority subcriteria are indifferent. There is also another subcriterion s_2 that strictly prefers β to α ($\beta \succ_{s_2} \alpha$), and all higher priority subcriteria are indifferent. Note that this also means that s_1 and s_2 have incomparable priorities, which means that neither overrules the other, so no preference can be determined. In this case, the subcriteria s_1 and s_2 together explain the incomparability.

Example 4. Consider again the lexicographic criterion c in Figure 1. c cannot compare between γ and δ. Subcriterion c_3 strictly prefers δ over γ, the other three subcriteria stricly prefer γ over δ. Not all subcriteria are suitable to explain the incomparability. c_4 is discarded because c_3 has higher priority. But also c_2 should not be used, even though it is incomparable in priority with c_3. This is because c_1 has higher priority and is not indifferent. This makes c_1 and c_3 the deciding criteria that are used as explanation.

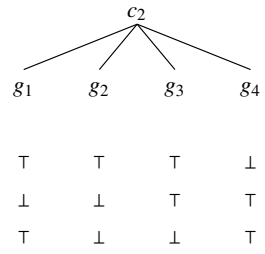


Figure 2. Example goal-based cardinality criterion

Goal-based cardinality criteria

Strict preference Suppose a goal-based cardinality criterion c strictly prefers an outcome α over an outcome β ($\alpha \succ_c \beta$). Then this is because the subgoals that α satisfies outnumber the subgoals that β satisfies. There may be subgoals that are satisfied by both α and β. They are counted on both sides, but do not influence the overall preference between α and β. Therefore, as an explanation of c's preference of α over β we only consider the subgoals g that α satisfies but β does not ($\alpha_{X_g} = \top$ and $\beta_{X_g} = \bot$, or equivalently, $\alpha \succ_g \beta$).

Example 5. Consider the goal-based cardinality criterion c_2 displayed in Figure 2. It has four goals g_1, g_2, g_3, g_4 as subcriteria. For three outcomes α, β, γ it is given whether they satisfy each of the four goals. The criterion c_2 strictly prefers α to β. The explanation of this preference is given by the goals g_1 and g_2 that α satisfies but β does not. Although α also satisfies goal g_3, this goal is not used in the explanation since it is also satisfied by β and hence is not deciding in the overall preference. Similarly, c_2's preference of α over γ can be explained by the goals g_2 and g_3.

Equal preference If a goal-based cardinality criterion c equally prefers two outcomes α and β ($\alpha \approx_c \beta$), this means that both outcomes satisfy the same number of subgoals of c. However, it does not necessarily mean that both outcomes satisfy the same goals. As explanation, we take the goals that α satisfies but β does not, and the set of goals that β satisfies but α does not. Both (disjoint) sets contain the same number of goals, which compensate for each other. This explains the indifference between the two outcomes.

Example 6. Consider again the goal-based cardinality criterion c_2 in Figure 2. c_2 is indifferent between β and γ. Both outcomes satisfy two goals, but one goal (g_4) is satisfied by both outcomes. Therefore the explanation of the indifference is given by g_3 (which is satisfied by β but not by γ) and g_1 (which is satisfied by γ but not by β).

4 USING EXPLANATION TO UPDATE A PREFERENCE MODEL

Before a preference model can be used in practice in a system, it has to be constructed or instantiated. Preference elicitation is likely to be an iterative process, and for this reason an existing preference model

should also be updateable. There are several ways of constructing and updating a preference model. In this paper we focus on the approach of guiding preference elicitation by asking the user particular questions and updating the preference model according to the answers. The advantages of this approach are that it provides an intuitive interaction with non-expert users and that preferences can be discovered during the process. In particular, we consider the case in which the user is asked not only to give his preference between two outcomes, but also to provide an explanation for this preference. This explanation can then be used to update the current preference model. If the user just provides his preference between outcomes, there may be many different ways in which the model could be updated to reflect this preference. The added value of additionally obtaining an explanation from the user is that it provides clues on how exactly the model should be updated, possibly after some further interaction involving targeted follow-up questions.

Updating a QPS model with explanations

We investigate how a system's current model of the user's preferences can be updated by engaging in a conversation with the user. Using explanations of preferences given by a user, the system can find out whether its current representation is accurate, and if not, where it has to be changed. Our approach allows for an initial model to be present that can be adapted by the user. The user can add preference information on his own initiative, or alternatively the system can ask the user to provide specific preferences (for example between two outcomes that are incomparable in its current model). In any case, if the preference given by the user does not match the preference that follows from the system's current model, the user is asked to provide an explanation. We assume that the user's explanation of his preference coincides with one of the explanations listed in Table 1. Depending on the user's answer and the nature of the top criterion (lexicographic or goal-based cardinality), the system can proceed by asking follow-up questions or updating its preference model in a particular way.

In the following, we discuss every situation in detail and provide interaction diagrams for each. We assume that the user has stated a preference between two outcomes that is not supported by the system's current preference model. It is important to distinguish between the current preference model maintained by the system, and the statements of the user. Since the interaction is designed to identify the elements of the model that need to be updated, the user's statements typically disagree with the current model. The interaction diagrams start with the system asking for an explanation for the given preference. The system's possible responses depend on the explanation given and the current preference model. More than one response may be applicable. In that case, the system should keep the interaction going until the preference model induces the given preference. When the process is finished, the updated preference model should not only model the preference given by the user, but also generate the same explanation for it.

Lexicographic criteria

Strict preference The interaction diagram for updating a preference model with a strict preference of an outcome α over an outcome β by a lexicographic criterion c is given in Figure 3. The explanation of such a preference is given by a subcriterion s of c that, according to the user, strictly prefers α to β. There can be different reasons why this subcriterion does not decide c's preference in the current preference model S.

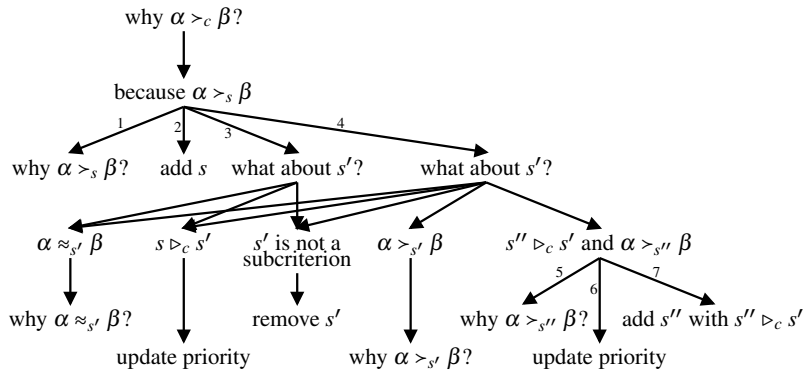
- First, s may not strictly prefer α to β according to S. In this case, the user is asked to explain this preference.
- Second, s may not be listed as a subcriterion of c in S. In this case, the system adds s to the set of subcriteria C_c.
- Third, according to S there may be another subcriterion s' that overrules s, i.e. that has higher priority but is not indifferent between α and β. In this case, the user is asked to clarify this issue, and may respond in several ways. (i) If the user states that s' actually is indifferent, he is asked for an explanation. (ii) If the user states that s actually has higher priority than s', the system updates the priority relation accordingly. (iii) If the user states that s' is not actually a subcriterion, the system removes s' from C_c.
- Fourth, according to S there may be another subcriterion s' that is not comparable in priority to s, does not weakly prefer α to β, and is not overruled. In this case, the user is asked to clarify this issue. The same responses by the user as in the previous case are possible, plus two more. (iv) If the user states that s' actually strictly prefers α to β, he is asked to give an explanation. (v) If the user states that there actually is another subcriterion s'' with higher priority that strictly prefers α to β, there are three options. If the preference does not follow from S, then the user is asked for an explanation. If s'' does not have higher priority than s' in S, the system updates the priority relation. And if s'' was not listed as a subcriterion of c, the system adds it with the right priority.

Equal preference The interaction diagram for updating a preference model with an equal preference between two outcomes α and β by a lexicographic criterion c is given in Figure 4. Such a preference is explained by the fact that, according to the user, all subcriteria are indifferent. There can only be one reason that the indifference does not follow from the current preference model S.

- There must be a subcriterion s in S that is not indifferent. In this case, the user is asked to clarify this issue. He can do so in two different ways. (i) If the user states that s is actually indifferent, he is asked to give an explanation. (ii) If the user states that s is not actually a subcriterion of c, then the system removes s from the set of subcriteria C_c.

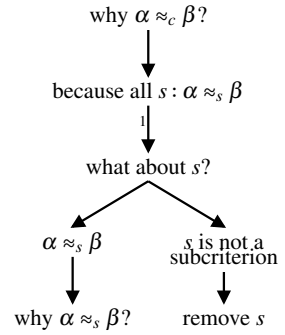
Incomparability The interaction diagram for updating a preference model with an incomparability between two outcomes α and β by a lexicographic criterion c is given in Figure 5. Since there are two kinds of explanation of such an incomparability, the interaction tree splits into two branches. If the incomparability is explained by a subcriterion that cannot compare between α and β according to the user, the possible responses are very similar to the case of strict preference. Therefore we do not discuss this case here but refer to the lefthand branch in Figure 5 for the details. If the incomparability is explained by two contradicting subcriteria s_1 and s_2, where $\alpha \succ_{s_1} \beta$ and $\beta \succ_{s_2} \alpha$ according to the user, there can be different reasons why these subcriteria do not decide c's preference in the current preference model S.

- First, it may be that $\alpha \not\succ_{s_1} \beta$ or $\beta \not\succ_{s_2} \alpha$ according to the current preference model S. In this case, the user is asked to explain that preference.
- Second, s_1 or s_2 may not be listed as a subcriterion of c in S. In this case, the system adds it to the set of subcriteria C_c.
- Third, according to S there may be another subcriterion s_1' that overrules s_1. In this case, the user can reply in different ways. (i) If the user states that s_1' is actually indifferent between α and β, he is asked for an explanation. (ii) If the user states that s_1' does not actually have higher priority than s_1, the system updates the priority relation ac-



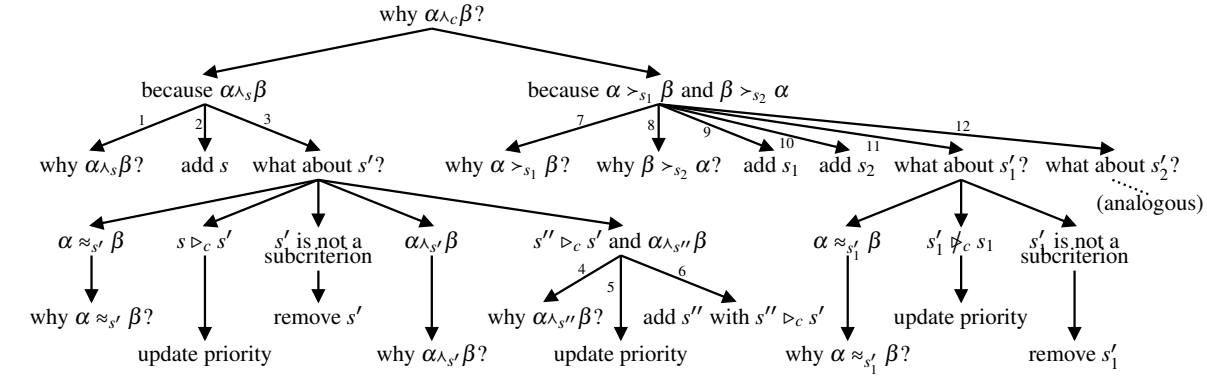
1: if $\alpha \not\succ_s \beta$
2: if $s \notin C_c$
3: for any $s' \in C_c$ such that $s' \rhd_c s$ and $\alpha \not\approx_{s'} \beta$
4: for any $s' \in C_c$ such that $s' \triangle s$ and $\alpha \not\succeq_{s'} \beta$ and s' is not overruled
5: if $\alpha \not\succ_{s''} \beta$
6: if $s'' \not\rhd_c s$
7: if $s'' \notin C_c$

Figure 3. Updating with a strict preference by a lexicographic criterion



1: for any $s \in C_c$ such that $\alpha \not\approx_s \beta$

Figure 4. Updating with an equal preference by a lexicographic criterion



1: if $\alpha \not\wedge_s \beta$
2: if $s \notin C_c$
3: for any $s' \in C_c$ such that $s' \rhd_c s$ and $\alpha \succ_{s'} \beta$ or $\beta \succ_{s'} \alpha$
4: if $\alpha \not\wedge_{s''} \beta$
5: if $s'' \not\rhd_c s'$
6: if $s'' \notin C_c$
7: if $\alpha \not\succ_{s_1} \beta$
8: if $\beta \not\succ_{s_2} \alpha$
9: if $s_1 \notin C_c$
10: if $s_2 \notin C_c$
11: for any $s'_1 \in C_c$ such that $s'_1 \rhd_c s_1$ and $\alpha \not\approx_{s'_1} \beta$
12: for any $s'_2 \in C_c$ such that $s'_2 \rhd_c s_2$ and $\alpha \not\approx_{s'_2} \beta$

Figure 5. Updating with an incomparability by a lexicographic criterion

cordingly. (iii) If the user states that s'_1 is not actually a subcriterion of c, then the system removes s'_1 from the set of subcriteria C_c.

• Fourth, according to S there may be another subcriterion s'_2 that overrules s_2. This case is handled analogously to the third case.

Goal-based cardinality criteria

Strict preference The interaction diagram for updating a preference model with a strict preference of an outcome α over an outcome β by a goal-based cardinality criterion c is given in Figure 6. The explanation of such a preference is given by a set of subgoals $g_1, \ldots, g_n$ that are all satisfied by α but not by β according to the user. There can be different reasons why this set of goals does not decide c's preference in the current preference model S.

• First, one of the goals may not be satisfied by α in S. In this case, the user is asked to explain this fact.

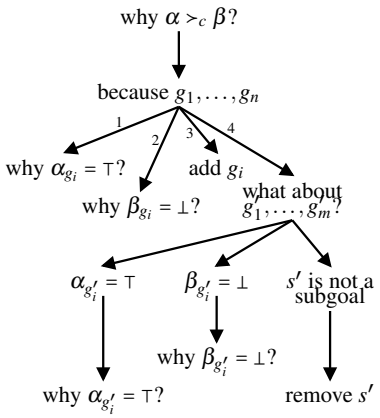
• Second, one of the goals may be satisfied by β in S. In this case, the user is also asked to give an explanation.

• Third, one of the goals may not be listed as a subgoal of c in S. In this case, the system adds it to the set of subgoals C_c.

• Fourth, there may be a set of goals $g'_1, \ldots, g'_m$ that are all satisfied by β but not by α according to S, which contains at least as many goals as $g_1, \ldots, g_n$. In this case, the user is asked to clarify this issue, and may respond in several ways. (i) If the user states that one of the goals is actually satisfied by α or (ii) not satisfied by β, he is asked to for an explanation. (iii) If the user states that one of the goals is actually not a subgoal of c, then the system removes this goal from the set of subgoals C_c.

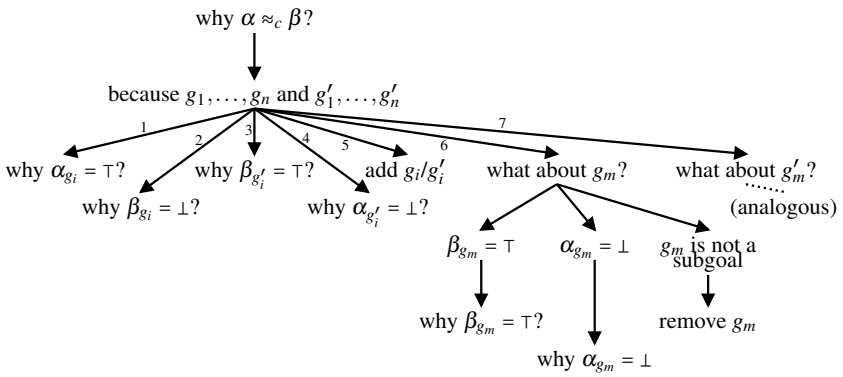
Equal preference The interaction diagram for updating a preference model with an equal preference between two outcomes α and β by a goal-based cardinality criterion c is given in Figure 7. The explanation of such a preference is given by two equally sized sets of subgoals: $g_1, \ldots, g_n$ that are all satisfied by α but not by β, and $g'_1, \ldots, g'_n$ that are all satisfied by β but not by α according to the user. Again, there can be different reasons why these sets of goals do not decide c's preference in the current preference model S.

• First, according to S, α may not satisfy some g_i, β may satisfy some g_i, β may not satisfy some g'_i, or α may satisfy some g'_i. In this case, the user is asked to give an explanation.



1: if $\alpha_{g_i} \neq \top$
2: if $\beta_{g_i} \neq \bot$
3: if $g_i \notin C_c$
4: for any $g'_1, \ldots, g'_m$ such that $m \geq n$ and for all $g'_i : \beta \succ_{g'_i} \alpha$

Figure 6. Updating with a strict preference by a goal-based cardinality criterion



1: if $\alpha_{g_i} \neq \top$
2: if $\beta_{g_i} \neq \bot$
3: if $\beta_{g'_i} \neq \top$
4: if $\alpha_{g'_i} \neq \bot$
5: if $g_i/g'_i \notin C_c$
6: for any $g_m \in C_c$ such that $g_m \notin g_1, \ldots, g_n$ and $\alpha \succ_{g_m} \beta$
7: for any $g'_m \in C_c$ such that $g'_m \notin g'_1, \ldots, g'_n$ and $\beta \succ_{g'_m} \alpha$

Figure 7. Updating with an equal preference by a goal-based cardinality criterion

• Second, any g_i or g'_i may not be listed as a subgoal of c in S. In this case, the system adds it to the set of subgoals C_c.
• Third, according to S there may be a goal g_m in C_c that is not in $g_1, \ldots, g_n$ and is satisfied by α but not by β. In this case, the user is asked to clarify this issue and may respond in several ways. (i) If the user states that β actually satisfies g_m, or (ii) α actually does not satisfy g_m, he is asked to explain this fact. (iii) If the user states that g_m is actually not a subgoal of c, then the system removes g_m from the set of subgoals C_c.
• Fourth, according to S there may be a goal g'_m in C_c that is not in $g'_1, \ldots, g'_n$ and is satisfied by β but not by α. This case is handled analogously to the third case.

5 CONCLUSION

Qualitative Preference Systems (QPS) [8, 9] provide a general framework for the representation of qualitative, multi-criteria preferences. We have shown that the composite tree structure of multiple criteria, combined with the non-compensatoriness of a qualitative approach provides a basis for the generation of explanations for the preferences that follow from a preference model represented in the QPS framework. The explanation strategy that we proposed is based on the intuition that preferences between outcomes can be explained by the criteria that are deciding in the overall preference. We identified the explanations that can be given for different preferences by different kinds of criteria. We then showed that the same explanations can also be useful when updating a preference model, because they provide information on how exactly the model should be updated.

Some interesting issues remain for future work. First, in some instances it may be necessary to explain facts about the outcomes involved in a preferential comparison, e.g. to explain why they do or do not satisfy a particular goal. Explanation of knowledge and reasoning is a separate field of study that may provide solutions to this issue. Second, when the system updates the priority relation between two subcriteria of a lexicographic criterion, this relation has to remain a partial order. Moreover, as the system iteratively engages in an interaction with the user as described here, it has to ensure that the previous preferences and explanations expressed by the user remain valid. It is important to investigate how such consistencies can be ensured. Third, the explanation of preferences may be part of a larger picture, for example in recommendation, decision making or planning. We would like to investigate how the explanation mechanism presented here can be embedded in other explanation mechanisms, such as the one presented in [1], where a tree structure of goals and beliefs is used to explain actions. Besides these theoretical considerations, we would like to take a more practical approach and implement the QPS framework together with the proposed explanation mechanism and update mechanism. We can then experimentally test the validity of our intuitions. This is related to the work of [3], who tested the predictive performance of the Take the Best (TTB) heuristic [4], which is a simplified instantiation of the lexicographic rule.

Acknowledgements This research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs. It is part of the Pocket Negotiator project with grant number VICI-project 08075.

REFERENCES

[1] J. Broekens, M. Harbers, K. Hindriks, K. van den Bosch, C. Jonker, and J.-J. Meyer, 'Do you get it? User-evaluated explainable BDI agents', in *Multiagent System Technologies*, *LNCS* 6251, 28–39, (2010).
[2] G. Carenini and J.D. Moore, 'Generating and evaluating evaluative arguments', *Artificial Intelligence*, **170**(11), 925–952, (2006).
[3] A. Dieckmann, K. Dippold, and H. Dietrich, 'Compensatory versus noncompensatory models for predicting consumer preferences', *Judgment and Decision Making*, **4**(3), 200–213, (2009).
[4] G. Gigerenzer, P.M. Todd, and ABC Group, *Simple heuristics that make us smart*, Oxford University Press, 1999.
[5] D.A. Klein and E.H. Shortliffe, 'A framework for explaining decision-theoretic advice', *Artificial Intelligence*, **67**(2), 201–243, (1994).
[6] C. Labreuche, 'A general framework for explaining the results of a multi-attribute preference model', *Artificial Intelligence*, **175**(7-8), 1410–1448, (2011).
[7] B. Moulin, H. Irandoust, M. Bélanger, and G. Desbordes, 'Explanation and argumentation capabilities:towards the creation of more persuasive agents', *Artificial Intelligence Review*, **17**(3), 169–222, (2002).
[8] W. Visser, R. Aydoğan, K.V. Hindriks, and C.M. Jonker, 'A framework for qualitative multi-criteria preferences', in *Proc. ICAART*, pp. 243–248, (2012).
[9] W. Visser, K.V. Hindriks, and C.M. Jonker, 'Goal-based qualitative preference systems', in *Proc. DALT*, (2012).