1
Introduction

Many tasks in everyday life involve making decisions over a large number of options (Schwartz 2005): we must decide which clothes to wear, what to eat, where to go for fun. Both these frequent decisions and infrequent ones, such as shopping for expensive goods or planning vacations, demand an effort that can be reduced by delegating decision-making to intelligent agents. For agents to appropriately perform tasks on our behalf, however, they must be aware of individual user preferences and the available options.

Our vision is for agents to make decisions on behalf of users so that their choices match those of users themselves, given adequate time and knowledge. However, it is important to understand that humans do not make decisions in isolation, and agents acting on their behalf should not do so either. Where the option chosen for one user may affect that of another (e.g. in deciding at which hotel to stay, we both prefer to stay at the same hotel), agents need to coordinate their actions. Such coordination between users reflects just one among the many interacting preferences that agents may need to consider. We argue that, by reflecting how users themselves decide, there is a *rationale* for choices that is *convincing* to users. Nevertheless, before we can make decisions appropriate to multiple users, we must first have agent reasoning appropriate to a single user, which is addressed in this thesis, consisting of a first step towards the broader vision.

Choosing from a set of available options often requires resolution of trade-offs, but it can be unfeasible for humans to carefully evaluate each option of a large set due to the required amount of time and cognitive effort, so that they are often unsatisfied with their choices (Schwartz 2005). As understanding human decision making and how to support them in making choices is important from many perspectives, such as understanding consumer behaviour and aiding managers in making high-impact business decisions, decision making has been extensively studied in a wide range of areas, including *economics* (Keeney and Raiffa 1976), *marketing* (Wierenga 2008), and *psychology* (Tversky 1996), for many decades. Moreover, receiving support or delegating a decision to a software system that is aware of users' preferences and is able to make decisions like them without effort restrictions and with adequate time can be very helpful. Therefore, decision making

has also received much attention in *computer science*, including in the areas of artificial intelligence, databases, and semantic web. Our goal in particular is to automate the decision making process and, in order to do so, we deal with three widely investigated issues: (i) how to capture and represent user preferences; (ii) how to reason about preferences and make decisions; and (iii) how to justify to users the decisions made.

According to Lichtenstein and Slovic (Lichtenstein and Slovic 2006), humans have a set of preferences that they are aware of — referred to as *known* preferences — which guide the decision making process. These preferences are expressed by individuals in different forms, but existing work on preference reasoning is only able to handle a restricted set of preference types, thus constraining users in expressing their preferences. One way to deal with that is to capture preferences through elicitation processes, but these can be tedious, discouraging users from using such processes. Moreover, these processes not only capture known preferences but also those needed to resolve trade-offs because the choices available often present a conflict among known preferences, so that trade-offs must be made. However, these additional preferences are constructed (as opposed to revealed) during the decision making process, as "*people do not maximise a precomputed preference order but construct their choices in the light of available options*" (Tversky 1996), and trade-off resolution requires cognitive effort of users (Schwartz 2005), thus compromising the acceptance of decision support systems that use elicitation processes.

Many approaches to reasoning about preferences are based on Multi-Attribute Utility Theory (MAUT) (Keeney and Raiffa 1976), which is designed to handle trade-offs among multiple objectives assuming a set of axioms (von Neumann and Morgenstern 1944) for preferences and utilities. Three of these axioms are (i) *description invariance*: preference order is always the same no matter how options are presented; (ii) *procedure invariance*: preferences do not depend on the elicitation process; and (iii) *context independence*: the addition of options does not impact preferences. A decision-maker whose preferences satisfy these axioms is considered rational, from an economic perspective, and consequently has a utility function (UF) that quantitatively represents preferences. However, as Tversky (Tversky 1996) has observed, human preferences often do not satisfy these axioms and, considering human irrationality, can we say that human preferences over a set of options are wrong? Moreover, should they be changed, in order to be considered rational? We assume that human preferences are not wrong, and if a decision model is not consistent with them, the model has to change. Consequently, preferences represented in the form of UF are not only hard to elicit but may also be inadequately represented.

Furthermore, *explanations* play an essential role in decision making. Humans often make and accept decisions made by others when they are able to identify the reasons for accepting and rejecting choices (Shafir et al. 1998), so that there are plausible arguments that justify the decision. Therefore, providing users with explanations that justify automated decision making is as important as providing adequate preference representation and reasoning.

Given this context, we present the problem we address in this thesis and the limitations of existing work in Section 1.1. We next describe our proposed solution and provide an overview of the contributions in Section 1.2, and then detail the structure of the remainder of this thesis in Section 1.3.

1.1 Problem Statement and Limitations of Existing Work

As introduced earlier, our research aims to automate decision making by tackling problems in three directions: (i) how to represent user preferences at a high level of abstraction (end-user level); (ii) how to make a choice from a set of available options using such preferences as input; and (iii) how to provide users with acceptable explanations that justify the decision. Based on these three issues, we state our primary research question below.

Research Question.

How can an automatic mechanism aware of a user's preferences choose one option from a set of available options and explain that choice, such that the user would be convinced of the adequacy of the choice?

Limitations of existing work, which are associated with this research question, are listed as follows.

There is no in-depth investigation of how humans express preferences nor is there a model that represents them. Many preference representation models have been proposed, to capture user preferences for decision making processes. However, such models are able to *represent* only a restricted set of preferences, constraining users in expressing their preferences, and creating the need for tedious interactive elicitation methods. Moreover, existing preference models are not justified by research studies, but are built in an *ad-hoc* way, based on intuition.

Existing approaches to reasoning about preferences are able to handle a restricted set of preference types and are limited to the identification

of non-dominated options. Even if there were models to represent high-level preferences, existing approaches to preference *reasoning* cannot handle all the constructions typically adopted by humans to express preferences. Therefore, it is important not only to represent different types of preferences, but also to be able to use them to make decisions.

Furthermore, preferences that users are able to specify without the aid of elicitation processes, i.e. their known preferences, are not enough to resolve trade-offs that emerge during the decision making process and, as a consequence, a decision making technique must provide a way to resolve trade-offs. Existing techniques are limited to selecting options that can be considered better according to provided (known) preferences. But, as these preferences often conflict — for example, maximising quality and minimising price — they are not enough to choose *one* option, but allow only the selection of a *subset* of options that have both pros and cons with respect to each other. The difficult step of the decision making process, namely trade-off resolution, remains for the user to perform.

Finally, most existing approaches rely on classical decision theory, which does not match how humans make decisions. Therefore, in order to make decisions like humans do, there is a need to take into account *human* decision making.

There is no consensus on what constitutes a good explanation to justify choice. The main goal of research into decision support and recommender systems has been to improve their *accuracy* (typically measuring the mean squared error of predicted ratings), associating this measure with the quality of the choice or recommendation. However, as argued by McNee et al. (McNee et al. 2006), the most accurate systems (based on standard metrics) may not be those that provide the most useful choices to users. Other aspects, such as *trust* and *transparency*, have also been considered, and many of these can be improved by providing users with *explanations* (Tintarev and Masthoff 2007).

There are different existing approaches to generating explanations (Klein and Shortliffe 1994, Labreuche 2011), from exposing the rationale of the underlying recommendation technique to selecting the essential attributes on which the decision is based. However, there is no consensus on what constitutes a *good* explanation, and what kinds of information must be presented to users in such explanations. Even though existing work (Tintarev and Masthoff 2007) provides qualitative arguments that characterise good explanations, there is only limited research on the kinds of explanation that users expect and need to understand recommendations or decisions made on their behalf. Where work does exist in this context, it is particular to a specific system.

1.2 Proposed Solution and Contributions Overview

In order to provide the kind of support we aim to give users — i.e. making choices that are consistent with high-level user preferences but going beyond them to resolve trade-offs — we propose an approach that involves preference representation, decision making and explanation generation, and these are founded on work in psychology and studies we performed with humans.

With the goal of providing a deeper understanding of how users express their preferences, we performed a study that involves the investigation of seven research questions, including how knowledge about a domain influences the expression of preferences and how users change their preferences after being exposed to decision-making situations. This study allowed us to identify the kinds of support users need to better express their preferences so that a system can make choices on their behalf. Given this study, we propose a preference metamodel that captures different kinds of preferences that humans adopt.

In order to tackle the problems associated with decision making, we propose a novel technique for making choices based on preferences and available options, whose main contributions are the following. First, it is able to handle qualitative preferences expressed in a high-level language, allowing users to express their preferences in a similar way to natural language, thus requiring less user effort than using a restricted preference language. Second, it incorporates psychological principles concerning how humans resolve trade-offs, as the provided user preferences are often not enough for making a decision. Our technique thus chooses one option from a finite set available, based on user preferences that have natural-language-like expressions, such as expressive speech acts (e.g. *like*, *accept* or *need*) — which are part of our preference metamodel. The decision making process is inspired by research work on human decision making (Shafir et al. 1998, Tversky 1972).

With regard to the identification of explanations that users expect to receive to justify choices, we present a study from which we extract types of explanation that humans use to justify a choice from a set of available options. As, based on the design of the study, we can assume that the explanations provided by study participants are those that the users would expect to receive, we derive a set of *guidelines* and *patterns*, which are a basis for generating explanations for users as to why particular options are chosen by decision support systems. Considering these identified explanation patterns, we also propose an explanation generation technique in order to produce appropriate and convincing explanations. The input for generating explanations is decision models generated during the decision making process of our user-centric preference-based technique. We include

algorithms to choose the appropriate explanation pattern in a given instance, and derive the parameters required to complete the explanation.

In summary, the main contributions of this thesis are:

(i) a **study of how humans express preferences**, which identifies preference constructions adopted by humans in natural language;

(ii) a **preference metamodel**, which allows the modelling of preferences at a high level;

(iii) an **automated decision making technique**, which chooses one option from a set available, based on high-level preferences;

(iv) a **study of how explanations can justify choices**, which identifies guidelines and patterns to generate explanations for users;

(v) an **explanation generation technique**, which justifies a choice made based on models produced by our decision making technique; and

(vi) a **user study**, which evaluates different aspects of our approach (preference metamodel, decision making technique and explanations), and also compares existing explanation generation approaches.

1.3 Outline

The remainder of this thesis is organised in four parts. Part I is related to the representation of preferences. Chapter 2 presents a study of how humans express preferences, which allows us to identify expressions that humans adopt to state their preferences. Based on the results of this study, Chapter 3 describes a preference metamodel that provides the different forms that are used to express preferences. Chapter 4 then discusses existing preference representation models, and compares them with our metamodel.

After discussing preference representation, Part II focuses on preference reasoning. First, a systematic review of preference reasoning approaches is presented in Chapter 5, introducing the work proposed in different areas of computer science. Chapter 6 details our novel decision making technique, which is able to handle different types of preference and incorporates user-centric principles. This chapter also compares our technique with existing work and presents an empirical evaluation.

Part III is concerned with another important issue in automated decision making: user explanations. Chapter 7 first presents a literature review of explanations, followed by Chapter 8, which describes a study in which we

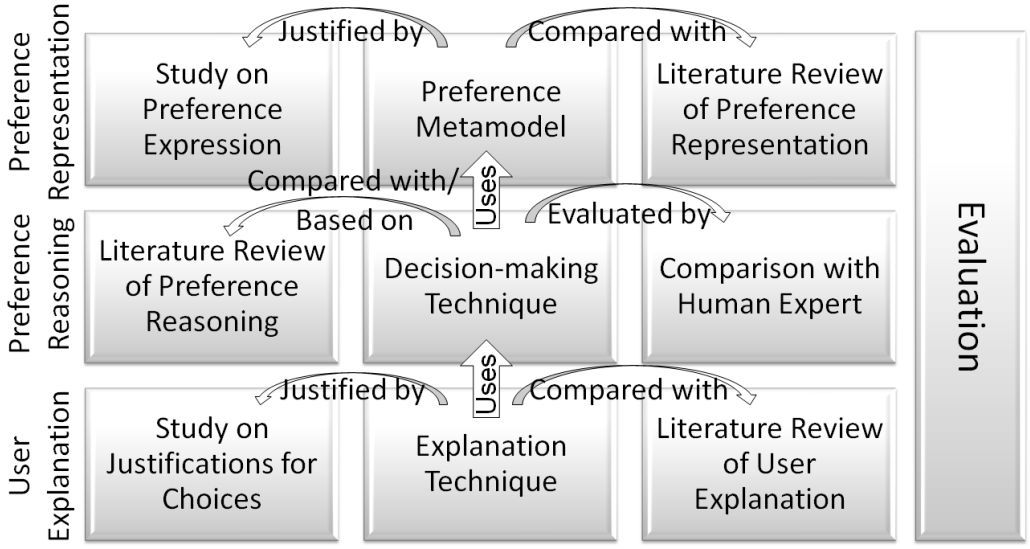

Figure 1.1: Thesis components and their relationship.

investigate explanations given by humans to justify a choice, from which we derive patterns and guidelines to be adopted by explanation approaches. Based on this result and our decision making technique, an explanation generation technique is detailed in Chapter 9.

Part IV connects all the previous parts and concludes this thesis. Chapter 10 describes a user study performed to evaluate all aspects of our approach: preference metamodel, decision making technique and explanation generation technique. Finally, conclusions and future work are presented in Chapter 11. We summarise in Figure 1.1 the different parts that will be presented in this thesis, and how they are related to each other.